

**COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR THE
PREDICTION OF ANAEMIA AMONG WOMEN AT REPRODUCTIVE AGE.**

A

THESIS

SUBMITTED TO

SHRI JAGDISHPRASAD JHABARMAL TIBREWALA UNIVERSITY

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

In

Statistics



POOJA SHIVAJI ZANJURNE

(Registration No.: 27621046)



UNDER THE GUIDENCE OF

Dr. Farooqui M. Ali Zakirhussain

(Registration No.: JJT/2K9/SC/0172)

UNDER THE CO-GUIDENCE OF

Dr. Vaishali Vilas Patil

(Registration No.: JJT/2K9/SC/1708)

DEPARTMENT OF STATISTICS


SHRI JAGDISHPRASAD JHABARMAL TIBREWALA UNIVERSITY,

VIDYANAGAR, JHUNJHUNU, RAJASTHAN-333001

2024

DECLARATION BY CANDIDATE

I declare that thesis entitled "Comparative Study of Machine Learning Algorithms for the Prediction of Anaemia among Women at Reproductive Age." is my own work, conducted under supervision of Dr. Farooqui M. Ali Zakirhussain, & Co-Supervision of Dr. Vaishali V. Patil, Shri JJT University at approved by Research Degree Committee. I have worked more than 200 days/600 hours of attendance with supervisor. I further declare to best of my knowledge that thesis does not contain any part of any work which has been submitted for award of degree either in this university or any other university/deemed university without proper citation.


Ms. Pooja Shivaji Zanjurne
NAME OF CANDIDATE

CERTIFICATE OF SUPERVISOR

This is to certify that work entitled "Comparative Study of Machine Learning Algorithms for the Prediction of Anaemia among Women at Reproductive Age" is piece of research work done by Pooja Shivaji Zanjurne under my supervision for degree of Doctor of Philosophy in Statistics of Shri JJT University, Jhunjhunu, Rajasthan, India that candidate has worked more than 200 days/ 600 hours with me. To best of my knowledge & belief thesis

- I. Embodies work of candidate herself
- II. Has duly been completed
- III. Fulfils requirement of ordinance related to Ph.D. degree of University and
- IV. Is up to standard both in respect of content & language for being referred to examiner

Dr. Farooqui M. Ali Zakirhussain



(SIG. AND STAMP)

**RESEARCH SUPERVISOR
SHRI J.J.T. UNIVERSITY
JHUNJHUNU (Raj.)**



CERTIFICATE OF CO-SUPERVISOR

This is to certify that work entitled "Comparative Study of Machine Learning Algorithms for the Prediction of Anaemia among Women at Reproductive Age" is piece of research work done by Ms. Pooja Shivaji Zanjurne under my supervision for degree of Doctor of Philosophy in Statistics of Shri JJT University, Jhunjhunu, Rajasthan, India that candidate has worked more than 200 days/ 600 hours with me. To best of my knowledge & belief thesis

- I. Embodies work of candidate herself
- II. Has duly been completed
- III. Fulfils requirement of ordinance related to Ph.D. degree of University and
- IV. Is up to standard both in respect of content & language for being referred to examiner

Dr. Vaishali V. Patil



Dr. Vaishali V. Patil
Associate Professor(Statistics)

Reg.No.: JJT/2K9/SC/1708

(SIG. AND STAMP)



ACKNOWLEDGEMENT

It gives me great pleasure to offer my sincere gratitude and most sincere regards to everyone who has made my research work feasible and a very memorable experience. My guide, well-wishers, and all of my relatives, among many others, helped, cooperated, guided, and encouraged me to finish my research work, which was a collaborative effort. Without their unwavering support, this endeavour would not have been possible. The time is here for me to express my sincere gratitude to each and every one of them. To be honest, at this fortunate time, words simply cannot capture my feelings and emotions.

I extend my deepest appreciation to my PhD guide Dr. Farooqui M. Ali Zakirhussain for his fullest support, willing cooperation, motivation, and guidance during the entire progress of this work. I am immensely grateful to Dr. Vaishali Vilas Patil for her guidance, mentorship, and unwavering support throughout this research journey. Your expertise and insightful feedback have been invaluable in shaping this thesis. I am wordless to thank Prin. Dr. Avinash Jagtap and Prof. Dr. Vikas Kakade for their continuous support to complete my Ph.D. study and related research work. Their concern and support have been invaluable to me in the completion of my research work. I am indebted to Dr. Neeta Dhane for her guidance, advice, and inspiration that have shaped my academic pursuits. I acknowledge to Principal, Registrar, President and Secretary of my college for providing the necessary resources, facilities, and a conducive academic environment for carrying out this research.

I extend my thanks to the research participants, Ms. Sarita Wadkar, Mr. Chandrashekhar P. Swami, Ms. Priti Malusare and Ms. Nilambari Jagtap who generously contributed their time and insights to my research, helped me a lot in the research work, and many other things. Without their willingness to participate, this study would not have been possible.

To my parents Shivaji Sadashiv Zanjurne and Shobha Shivaji Zanjurne, your unwavering belief in my abilities and the sacrifices you made to provide me with an education have made all the difference. Your guidance and unwavering support have been my constant motivation. Thank you for instilling in me a love for learning and a strong work ethic.

To my uncle and aunty, Sunil Zanjurne and Pranita Zanjurne, your guidance and encouragement instilled in me a love for learning and a strong work ethic. I want to express my deepest gratitude to my siblings Neha Zanjurne, Aditya Zanjurne, for


your patience, camaraderie, and occasional distractions were much-needed breaks and moments of laughter during this academic journey.

To my dear mother-in-law, Rekha Laxman Gaikwad, your unwavering support, love, and assistance with household responsibilities were a lifeline during this journey. I am deeply grateful for the sacrifices you all made and the patience you displayed during the long hours I spent working on this thesis.

To my loving husband, Sujit Laxman Gaikwad, your belief in my dreams and your constant encouragement were the driving force behind the successful completion of this thesis. Your willingness to shoulder additional responsibilities at home and provide me with uninterrupted study time was invaluable. You are not just my partner in life, but also my partner in this academic endeavour.

I also have no words to express my sense of gratitude to my lovely daughter, Siya Sujit Gaikwad, for her love and support during this journey. To conclude, I humbly believe that I could achieve success in my research work as well as in my life due to the blessings of my late grandmother. Smt. Hirabai Sadashiv Zanjurne.

I want to dedicate this thesis to the loving memory of my late little brother, Pravin Shivaji Zanjurne. Though he is no longer with us, his spirit and presence continue to inspire me every day. I often find myself thinking about how proud he would have been of my achievements. Although he cannot be here to witness this milestone, I know he is with me in spirit.



Date: 11-2-2024

Place: Baramati

Ms. Pooja Shivaji Zanjurne

Research Scholar

Dedicated to the Pillars of my existence

Mr. Shivaji Zanjurne

&

Mr. Sujit Gaikwad

TABLE OF CONTENTS

Title page.....	(i)
Declaration by The Candidate.....	(ii)
Certificate of Supervisor.....	(iii)
Certificate of Co-Supervisor.....	(iv)
Acknowledgement.....	(v)
List of Contents.....	(viii)
List of Figures.....	(x)
List of Tables.....	(xi)
List of symbols, Notations and Abbreviations.....	(xii)
Abstract.....	(xiv)

CHAPTER 1	INTRODUCTION	Page no.	
	1.1	Research background	1
	1.2	Introduction to anaemia	2
	1.3	Causes of anaemia	4
	1.4	Consequences of Anaemia	6
	1.5	Government initiatives	9
	1.6	Data mining	12
	1.7	Machine learning algorithms	13
	1.8	Supervised learning algorithm	14
	1.9	Unsupervised learning algorithms	21
	1.10	Reinforcement learning	22
	1.11	Software	22
	1.12	Problem in hand	25
	1.13	Research objectives	25
	1.14	Scope of the research	25
	1.15	Outline of the thesis	26
	1.16	Terminology	26
CHAPTER 2	LITERATURE REVIEW		
	2.1	Introduction	28
	2.2	Literature Review	28
	2.3	Findings	64
CHAPTER 3	METHODOLOGY		
	3.1	Introduction	65
	3.2	Data pre-processing and its need	65
	3.3	Data visualisation and its importance	66
	3.4	Analysis of data	66
	3.5	Pilot study	66
	3.6	Statistical analysis	74
	3.7	Confusion Matrix	92
CHAPTER 4	PILOT STUDY		
	4.1	Prevalence of anaemia in India	94

	4.2	Prevalence of anaemia in Maharashtra	94
	4.3	Area wise distribution of anaemia	97
	4.4	Anaemia according to pregnancy status	98
	4.5	Relationship of anaemia and marital status of WRA	99
	4.6	Prediction of anaemia with Machine learning techniques	100
	4.7	Relationship of significant factors with anaemia	109
CHAPTER 5			
	5.1	Introduction	114
	5.2	Prevalence of anaemia among unmarried WRA	114
	5.3	Decision tree	115
	5.4	Ensemble techniques	140
	5.5	Overall comparison of machine learning algorithms	151
	5.6	Relationship of anaemia with significant contributors in prediction	153
CHAPTER 6			
	6.1	Introduction	162
	6.2	Prevalence of anaemia among non-pregnant WRA.	162
	6.3	Decision tree.	163
	6.4	Support vector machine with cost sensitive model	168
	6.5	K nearest neighbour algorithm	175
	6.6	Ensemble techniques to boost the model performance	177
	6.7	Comparison of machine learning algorithms by accuracy	187
	6.8	Stocking in Ensemble algorithm	189
	6.9	Relationship between significant factors and anaemia status	197
CHAPTER 7			
	7.1	Introduction	210
	7.2	Prevalence of anaemia among pregnant WRA.	210
	7.3	Decision tree.	211
	7.4	Support vector machine with various kernels.	215
	7.5	K nearest neighbour algorithm	226
	7.6	Ensemble techniques on pregnant women	228
	7.7	Comparative examination of developed machine learning algorithms	242
	7.8	Relationship of anaemia and influential factors	244
CHAPTER 8			
	SUMMARY AND CONCLUSION		
	8.1	Introduction	255

	8.2	Summary and conclusion	258
	8.3	Recommendations	267
	8.4	Future scope	269
	8.5	Limitations of research	269

LIST OF FIGURES

Fig No.	Name of Figures	Page No.
1.1	Machine learning algorithms	14
1.2	Polynomial curve	16
1.3	Non-linear regression	17
1.4	Logistic graph	18
1.5	K-NN classifier	19
1.6	SVM with linearly separable data	20
1.7	Confusion matrix	27
2.1	Flow of research	29
3.1	Types of data	65
3.2	Flow of Supervised learning algorithm	74
3.3	Types of supervised learning models	75
3.4	Classification tree	77
3.5	Regression tree	78
3.6	partition of feature space	79
3.7	Dividation of sample space	79
3.8	Decision tree	80
3.9	Linearly separable 2D data	82
3.10	Margin of Hyperplane	82
3.11	Maximum margin Hyperplane	84
3.12	SVM decision	86
3.13	Ensemble learning	88
4.1	Anaemia Prevalence in India	94
4.2	Anaemia prevalence in Maharashtra	96
4.3	State wise anaemia distribution	97
4.4	anaemia prevalence by Rural-Urban status	98
4.5	Anaemia prevalence according to pregnancy status	99
4.6	Decision tree plot for pilot study	103
4.7	Variable Importance plot of Random Forest	107
5.1	Prevalence of anaemia among Unmarried WRA	114
5.2	Decision tree plot 1	118

5.3	Decision tree plot 2	122
5.4	Decision tree plot 3	130
5.5	Comparison of SVM model with various kernels	138
5.6	Variable Importance plot by bagged decision tree	142
5.7	Important variable by RF classifier	146
5.8	Comparison of Machine learning Algorithms for Unmarried WRA	152
6.1	Prevalence of Anaemia among Married Non- pregnant WRA.	162
6.2	Decision tree 1 for Non-pregnant WRA	166
6.3	Accuracy with various kernels of SVM	170
6.4	Variable importance by RF classifier	181
6.5	Comparative accuracy plot of fitted machine learning algorithms	188
6.6	Meta model Variable importance plot	195
6.7	Plot of anaemic WRA with different levels of husband's education.	208
7.1	Anaemia Prevalence in Pregnant WRA.	210
7.2	Decision Tree plot 1 for Pregnant WRA	215
7.3	VarImpPlot by RF for pregnant WRA	236
7.4	Accuracy Comparison of Machine learning Algorithms for pregnant WRA	243
7.5	Anaemia distribution among eating habits	248

LIST OF TABLES

Table No.	Name of Table	Page No.
2.1	Anaemia Categories.	54
3.1	Confusion matrix for 4 classes.	92
4.1	Anaemia distribution in India.	94
4.2	State wise anaemia prevalence.	96
4.3	Area-wise anaemia percentage.	98
4.4	Anaemia distribution according to pregnancy status.	98
4.5	Marital status with anaemia.	100
4.6	Variable importance by CART.	101
4.7	Confusion matrix by CART.	104
4.8	Variable importance by Random Forest.	106
4.9	Confusion Matrix for Random Forest.	108
4.10	Anaemia with working status of WRA.	109
4.11	Anaemia and Husband's job status.	109
4.12	Anaemia status and average wealth index.	110
4.13	Anaemia status and average number of residential years.	110
4.14	Age at first birth and anaemia.	111
4.15	Biological Age of Household Head and anaemia severity.	111
4.16	Household Members Total with anaemia.	112
4.17	Education of women with anaemia.	112
4.18	Age of WRA and anaemia.	113
5.1	Variable importance table of CART.	117
5.2	Significant variables by updated CART.	121
5.3	SVM model's accuracy comparison.	137
5.4	Top Important variable by random forest.	145
5.5	Top significant variable by Ada Boot for unmarried WRA.	150
5.6	Comparison of machine learning algorithms for Unmarried women.	151
5.7	Average weight of unmarried WRA according to Anaemia status.	153

5.8	Average BMI of unmarried WRA according to Anaemia status.	154
5.9	Average age of unmarried WRA according to Anaemia status.	155
5.10	Frequency distribution of age of unmarried WRA according to Anaemia status.	155
5.11	Average of Number of years lives in residential area.	156
5.12	Average age at menstrual cycle begins with anaemia.	157
5.13	Average height of unmarried WRA according to Anaemia.	158
5.14	Income of family and anaemia.	158
5.15	HIV status and anaemia.	159
5.16	Number of days of blood flow with anaemia.	160
5.17	No of pads (per day) with anaemia.	160
5.18	No of pads (per day) according to anaemia categories.	160
6.1	Comparison of Machine learning algorithms by Accuracy.	187
6.2	Feeling weak or dizziness over anaemia severity.	197
6.3	Average BMI among Anaemia.	198
6.4	Average of Number of years lives in residential area across anaemia.	198
6.5	Anaemia and average of Age at the marriage.	199
6.6	Average of Age of last children (month) Vs Anaemia.	199
6.7	Anaemia with average age of WRA.	199
6.8	Anaemia with average height of WRA.	200
6.9	Anaemia with average weight of WRA.	200
6.10	Average of husband's age at marriage with anaemia status.	201
6.11	Average age of husband (years) of WRA with anaemia.	201
6.12	WRA's age at first birth of child versus anaemia.	202
6.13	Average of number of days of blood flow according to anaemia classes.	202
6.14	Average family income with anaemia.	203
6.15	Contingency table of anaemia and husband's occupation	204
6.16	Contingency table of Region and anaemia.	205
6.17	Average of Age at menstrual cycle begins with anaemia.	206
6.18	Average of family size with anaemia.	207

6.19	Frequency table of husband's education and anaemia status.	207
6.20	Percentage of anaemic WRA with different levels of husband's education.	208
6.21	Alcohol consumption status of WRA with Anaemia severity.	209
7.1	Anaemia distribution of pregnant WRA.	210
7.2	Most influential factors related to anaemia by CART.	213
7.3	Table of variable importance by RF.	236
7.4	Sorted important variables by Ada Boost.	241
7.5	Table of accuracy for developed models.	242
7.6	Average of Age at the marriage with anaemia in pregnant WRA	246
7.7	Average of weight(kg) of pregnant women.	246
7.8	Table of Average BMI of pregnant WRA among anaemia categories.	247
7.9	Distribution table of taste preferences.	247
7.10	Alcohol consumption in Pregnant women	249
7.11	Distribution of average age of pregnant WRA.	249
7.12	Anaemia status and age at menstrual cycle begins with anaemia.	250
7.13	Table of Anaemia status and average number of days of blood flow during menstrual.	250
7.14	Average husband's age of pregnant WRA.	250
7.15	Average family size Vs Anaemia.	251
7.16	Table of Average of husband's age at marriage.	251
7.17	Average of number of years lives in residential area versus anaemia category in pregnant WRA.	251
7.18	Relationship between Anaemia status and Income of the family.	252
7.19	Relationship between Anaemia status and Age of last children in pregnant WRA.	252
7.20	Relationship between Anaemia status and Height of WRA.	253
7.21	Relationship between Anaemia status and Age at first birth of child.	253
7.22	Relationship between Anaemia status and HIV status of WRA.	254
7.23	Relationship between Anaemia status and No of pads (per day).	254

ABBREVIATIONS

WRA	Women at reproductive age.
PW	Pregnant women
NPW	Non-pregnant WRA
LR	Linear Regression
MLR	Multiple linear Regression
SLR	Simple linear Regression
BNR	Binary Logistic regression
RL	Reinforcement Learning
ML	Machine Learning
DT	Decision Tree
CART	Classification And Regression Tree
DM	Data Mining
β	Constant terms
ϵ	Error term
R.M.S.E	Root Mean Squared Error
S.S.E.	Sum of Square of Error
S.A.E.	Sum of Absolute Error
S.E.	Sum of Error
R.F.	Random Forest
SVM	Support Vector Machine
KNN	K-Nearest Neighbour
ADA Boost	Adaptive Boosting
AI	Artificial Intelligence
D.F.	Degrees of Freedom
C.P.	Complexity Parameter
Signif.codes	Significance Code
AUC	Area Under the Curve
ROC Curve	Receiver Operating Characteristic Curve
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
n	Number of Observation

RBC	Red Blood Cell
WBC	White Blood Cell
Hb	Haemoglobin
MCH	Mean Cell Haemoglobin
PLT	Platelet
MCHC	Mean Cell Haemoglobin Concentration
RDW	Red Cell Distribution Width
MCV	Mean Corpuscular Volume
NEUT	Neutrophils
TIBC	Transferrin and Iron Binding Capacity
BDHS	Bangladesh Demographic and Health Survey
PR	Poisson regression
GLM	general linear model
CNN	convolutional neural network
NDHS	Nepal Demographic and Health Survey
NHANES	National Health and Nutrition Examination Survey
OLR	Ordinal logistic regression

ABSTRACT

Machine learning has emerged as a transformational technology in a variety of fields, and its applications are expanding at an exponential rate. With an ever-expanding array of machine learning algorithms, choosing the right one for a specific task be a big challenge. Machine learning, a subset of artificial intelligence, has emerged as a transformational force in the medical field. The health and well-being of reproductive age women plays crucial role in the overall health of communities and future generations. Anaemia remains a critical global public health issue, particularly affecting reproductive-age women. According to WHO, ‘Anaemia is a condition in which the number of red blood cells or the haemoglobin concentration within them is lower than normal. It mainly affects women and children.’ Therefore, the reproductive aged women take care into account for this research.

This study aims to provide a comprehensive comparison of various popular machine learning algorithms to predict anaemia among women at reproductive age (WRA) and identifies the factors associated with the status of anaemia. To achieve these objectives the primary data were collected using well designed questionnaire. Total 664 reproductive aged women were selected in this research. According to WHO guidelines anaemia categorisation was done. Anaemia was categorised into four categories ‘no Anaemia’, ‘mild’, ‘moderate’ and ‘severe’.

Popular machine learning algorithms such as, decision tree, support vector machine (SVM) with linear, polynomial, radial, sigmoid kernel, and k-nearest neighbour (KNN). To enhance the accuracy of prediction ensemble methods like Random Forest, Bagged Decision tree and Ada Boost. The results says that the ensemble methods gives comparative high accuracy than that of single algorithms. At the last it was observed that the individual level factors like age, education, height, weight, etc, household level factors like number of family members and community level factors like number of years lives in residential area affects the status of anaemia.



Pooja Shivaji Zanjurne

Registration Number 27621046

CHAPTER 1

INTRODUCTION

1.1 Research background:

Without women, we are unable to picture how successful life would be in general. They have a major share of the blame for the continued success of life on this planet. In the past, they were only considered to be wife and mother who were responsible for the sole responsibility of preparing meals, cleaning, and taking care of the entire family. But now that their condition has slightly improved, they have begun engaging in activities other than those with their family and children. The nation's pioneers are women. Given that women make up half of the world's population, Indian culture values them highly. Women make up 50% of the human resource pool, making them the second-largest behind men in terms of potential, according to a report from the UN secretary general. The success of sustainable development and family life depends on women.

While caring for responsibilities, women should take care of their own health as well, but more often than not, women tend to neglect their health. Due to both biological and gender-related distinctions, being a man or a woman has a substantial effect on one's health. Because they are often discriminated against due to socio-cultural issues in many nations, the health of women and girls is a topic of particular concern. Poverty is a significant obstacle to good health outcomes for both men and women, it tends to have a greater inverse impact on women's and girls' health. Women face particular healthcare issues and are more likely than men to receive a diagnosis for some disorders.

A woman's ability for reproduction is a biological wonder. The ability of women for reproduction is an intrinsic component which is important in existence and evolution of human. It includes the intricate interactions between hormones, organs, and physiological functions involved in conception, gestation, delivery, and raising children. The hallmark of womanhood is her ability to become pregnant, bring a pregnancy to term, and give birth. Beyond its biological limits, women's ability to reproduce comprises complex elements that interact with individual, social, and cultural domains, profoundly influencing lives and communities. Women who are of reproductive age represent a crucial demographic whose health and welfare have a significant influence on communities and societies as a whole, not just on individuals.

Numerous physical, mental, and social changes occur throughout this stage of a woman's life.

In India it was found that 19% of the maternal deaths were related to anaemia [65]. It is critical to address anaemia in pregnancy since it has been linked to adverse pregnancy outcomes like preterm delivery, low-birth-weight new-borns, fetal mortality, and, in certain circumstances, maternal death [66]. Therefore, in this research the women at reproductive age were taken into consideration.

In the next section we focus on anaemia of reproductive women since it is severe cause of death in WRA.

1.2 Introduction to Anaemia

Anaemia is a condition where there are either too few RBC or too little Hb in them. Tiredness, weakness, fainting, and shortness of breath are just a few of the symptoms of anaemia. Dietary deficiencies especially iron deficiency, as well as haemoglobinopathies, infectious disorders such malaria, tuberculosis, HIV, and parasite infections, and nutritional deficits in vitamin B-12 and folate are primary reasons of anaemia.

Anaemia affects many different demographic groups, particularly women at reproductive age (WRA), and is an important worldwide health concern. Reduced haemoglobin levels or a reduction in the number of RBCs in the blood cause anaemia, which lowers the blood's ability to carry oxygen. This disorder may have significant negative effects on WRA's health and wellbeing, with broad ramifications for mental development, general quality of life, and the health of mothers and children. Reduced oxygen-carrying capacity due to anaemia can cause a range of symptoms and health issues.

1.2.1 Prevalence and trends of anaemia:

According to WHO statistics, anaemia affects 40% of pregnant/expecting women and 42% of kids up to the age 5 years. According to DHS report 2011 anaemia impacts over 500 million women in developing countries, resulting in an intolerable amount of avoidable illness and death, reduced economic output, and missed chances for social, economic, and personal growth. According to the most recent estimates from the WHO, 3 out of every 10 NPW and 4 out of 10 PW are both anaemic. According to data, when anaemia prevalence is 20%, 50% of the population may be iron deficient. Everybody in the population experiences some level of iron deficiency when anaemia prevalence exceeds 40%.

Anaemia doubles the chance of pregnancy death and stunts children's mental development. Recent research indicates the distribution of anaemia among Indian women of childbearing age has typically been 20% higher than the global average. In India, one in two women are anaemic, compared to one in three worldwide. From the NFHS-5 report of India Anaemia prevalence among kids aged 6 months to 59 months grew from 59% to 67% between 2015–16 and 2019–21, and it remained higher among children in rural areas. Also, it was observed that 57 percent of women between the ages of 15 and 49 are anaemic which is main reason behind increment in above mentioned child anaemia statistic. Women are mildly anaemic in 26% of cases, moderately anaemic in 29% of cases, and severely anaemic in 3% of cases. In almost all of the subgroups of women, anaemia is continuously high, with a prevalence of more than 50%. Anaemia varies depending on the pregnancy status; 61 percent of breastfeeding mothers have anaemia, compared to 52 percent of pregnant mothers and 57% of women who were neither pregnant nor breastfeeding, with education, the prevalence of anaemia typically decreases, as household wealth rises, the percentage of anaemic men and women continuously decreases. The percentage of women who were anaemic was slightly lower in urban regions (54%), compared to the percentage of women who were anaemic in rural areas (59%). In recent research it was found that women's anaemia is influenced by their living situation (rural vs. urban), education level, financial standing, and, most importantly, whether or not they are pregnant.

Typically, mild iron deficiency anaemia doesn't result in consequences. Iron deficiency anaemia, however, can grow severe if left untreated and result in health issues like heart issues, early births, and low birth weight infants. Severe iron shortage in children and new-borns can result in anaemia, as well as slowed growth and development. Anaemia due to iron deficiency is also linked to a higher risk of contracting infections. So, it is a need of early diagnosis of anaemia by which we can avoid its dangerous effects. Therefore, the importance of addressing anaemia was discussed in next section.

In addition, maternal and infant health might be significantly impacted by anaemia in WRA. Anaemia increases a woman's chance of difficulties during pregnancy, including low birth weight, premature birth, and maternal death. Additionally, their new-borns might be born with iron-deficiency anaemia, which would continue the anaemia cycle into future generations. Public health initiatives frequently concentrate on tactics like nutritional education, iron supplementation, and

access to high-quality prenatal care in order to prevent and manage anaemia in this crucial population because they understand how important it is to treat anaemia in women of reproductive age. This will not only benefit women's health and wellbeing, but it will also increase the likelihood of better pregnancies and offspring. Therefore, in the next section importance of addressing anaemia was explained.

1.2.2 Importance of Addressing Anaemia:

Because anaemia has a significant influence on both individual and public health, it must be addressed immediately. Anaemia can cause a number of severe symptoms that can lower quality of life, including tiredness and exhaustion as well as cognitive impairment. Furthermore, because anaemia reduces labour capability and productivity—especially in underprivileged communities with limited access to healthcare and nutrient-dense food it has an impact on larger societal and economic domains. Notably, treating anaemia during pregnancy lowers the likelihood of problems like low birth weight and premature birth, improving mother and child health outcomes. Treatment for anaemia can also strengthen the immune system and lower the risk of cardiovascular problems. Addressing anaemia is a critical component of public health activities worldwide since it not only reduces healthcare costs but also promotes healthier communities and societies by minimising difficulties associated with underlying disorders and boosting cognitive development in children.

The common causes of anaemia are briefly discussed in the section that follows.

1.3 Causes of anaemia:

There are various types of anaemia based on different causes, which can be broadly categorized into the following:

1. Iron-Deficiency Anaemia:

In the world, this is the most prevalent kind of anaemia. It happens when the body doesn't have enough iron in it to make enough haemoglobin. Iron-deficiency Anaemia is a common type of anaemia that can be treated. It is defined by low iron levels in the body, which lowers the ability of body to make RBC and haemoglobin. Increasingly, iron requirements during pregnancy, poor dietary iron intake, impaired iron absorption from illnesses like celiac disease or inflammatory bowel disorders, and chronic blood loss from heavy menstruation, gastrointestinal bleeding, or other medical conditions are common causes of this type. Weakness, exhaustion, pallor, and dyspnoea are possible symptoms. In order to treat chronic bleeding or absorption problems, treatment usually consists of treating the underlying cause, nutritional adjustments, iron supplements,

and, in certain situations, medical intervention. In order to avoid problems and return normal red blood cell production, early diagnosis and therapy are essential.

2. Vitamin Deficiency Anaemia:

Vitamin deficiency anaemia encompasses several types of anaemia caused by insufficient levels of specific vitamins essential for red blood cell production. Folate deficiency anaemia and Vitamin B-12 deficiency anaemia are the most common forms of this type. A vitamin b-12 deficiency anaemia can be caused by disorders of the stomach or small intestine, malabsorption problems, or dietary deficiencies. This is because vitamin B12 is essential to produce of RBCs. Folate deficiency anaemia is caused by a diet low in leafy green vegetables, legumes, or fortified cereals, or by problems with the absorption of folate (vitamin B9). Fatigue, weakness, and other anaemia symptoms can be caused by either form of vitamin deficiency anaemia. Treatment usually entails vitamin supplementation, dietary changes, and addressing underlying causes. Restoring healthy red blood cell production and minimising consequences require early detection and management.

3. Haemolytic Anaemias:

A class of diseases known as haemolytic anaemias are defined by the rapid breakdown of red blood cells, which frequently leaves the bloodstream lacking in these cells. Haemolytic anaemias have a variety of causes, such as autoimmune illnesses, in which body erroneously targets RBCs, exposure to chemicals or certain drugs, and hereditary diseases such sickle cell anaemia and thalassemia. Fatigue, jaundice, and an enlarged spleen are among the symptoms associated with homolyses, or the breaking down of red blood cells. Depending on the underlying reason, treatment options vary and may include controlling symptoms, treating the autoimmune component, or, in extreme situations, removing the spleen surgically or giving blood transfusions.

4. Chronic Diseases:

Anaemia can result from persistent diseases such rheumatoid arthritis, chronic kidney disease, and some types of cancer. The body's reaction to a chronic sickness and inflammation can both affect how many red blood cells are produced.

5. Aplastic Anaemia:

Aplastic-anaemia is caused by severe reduction in capacity of bone marrow to generate all types of blood cells; the condition is extremely rare and can be deadly. It is frequently identified as a condition of bone marrow failure. This illness can be related to factors such as exposure to radiation, poisons, certain drugs, infections, or it can be

idiopathic, with no known cause (idiopathic aplastic anaemia). Low levels of RBCs, WBCs, and platelets cause aplastic anaemia, which manifests as weakness, exhaustion, recurrent infections, and easily bruised or bleeding. Blood transfusions, drugs that increases the generation of RBC, bone marrow transplants, and, when practical, treating the underlying cause are among forms of treatment. Effective management of this illness requires early diagnosis and intervention.

6. Sickle Cell Anaemia:

The genetic condition known as sickle cell anaemia causes red blood cells to have an irregular shape, causing them to break down more easily and leading to anaemia. It is common among individuals of African, Mediterranean, and Middle Eastern descent. It is frequently identified as a condition of bone marrow failure. This illness can be related to factors such as exposure to radiation, poisons, certain drugs, infections, or it can be idiopathic, with no known cause (idiopathic aplastic anaemia).

7. Thalassaemia:

A class of hereditary blood illnesses known as thalassaemia which impacts the synthesis of haemoglobin and the protein that carries oxygen in red blood cells. Individuals which are suffering from thalassaemia have abnormal haemoglobin production, which leads to a deficiency of healthy red blood cells and causes anaemia. Thalassaemia can range in severity from mild to severe and it often requires lifelong medical management, including blood transfusions and iron chelation therapy. Thalassaemia is most commonly found in people of Mediterranean, Middle Eastern, and Southeast Asian descent, and its symptoms can range from mild fatigue to severe complications, making early diagnosis and appropriate medical care crucial for those affected by this condition.

The severity of the consequences can vary depending on the underlying cause of anaemia, its duration, and its extent. In the next section the consequences of anaemia were discussed.

1.4 Consequences of Anaemia:

Both short and long term effects of anaemia can be extensive, impacting numerous aspects of health and wellbeing.

1. Weakness and Fatigue:

Fatigue and weakness result from anaemia. With little physical or mental effort, people with anaemia frequently feel exhausted.

2. Paleness:

Anaemia can lead to pale skin, especially in the mucous membranes and nail beds, as it results from a deficiency of healthy RBCs the pigment haemoglobin, which gives red colour to blood and gives our skin its rosy hue when oxygenated. When anaemia reduces the count of functional RBCs the skin and mucous membranes can appear noticeably paler. This pallor is sign of anaemia and serves as a visible reminder of the condition's impact on the body's oxygen supply, often prompting individuals to seek medical attention to address the underlying causes and manage the condition.

3. Shortness of Breath:

People with severe anaemia may find it difficult to breathe, especially while exerting themselves. The body fails to provide tissues enough oxygen, which results in this.

4. Dizziness and Light-headedness:

Anaemia is frequently accompanied by symptoms such as light-headedness and dizziness. People who have anaemia frequently feel lightheaded and dizzy, especially when they get up abruptly or perform abrupt, demanding tasks. These symptoms suggest that anaemia could be negatively impacting a person's overall health, since they are caused by the brain not getting enough oxygen, which can impair balance and coordination. To relieve these symptoms and treat the underlying anaemic disease, a proper diagnosis and course of therapy are important.

5. Cold Hands and Feet:

Anaemia frequently manifests as cold hands and feet, which are mostly the result of the disease's impaired circulation. Blood's ability to carry oxygen is diminished in anaemia, and blood flow to the extremities may be jeopardised as the body gives priority to oxygen delivery to essential organs. Consequently, people suffering from anaemia frequently have a continuous feeling of being cold in their hands and feet. This symptom acts as a clear reminder of the difficulties with circulation that anaemia causes; treating the underlying cause of anaemia can assist to enhance blood flow and lessen these discomforts.

6. Reduced Work Capacity:

Anaemia can limit an individual's ability to perform physical tasks or engage in regular activities, affecting work capacity.

7. Headaches:

In circumstances where the brain does not receive enough oxygen, anaemia can cause headaches.

8. Chest Pain:

Severe anaemia can potentially lead to chest pain, it is also known as angina. In some cases, it may even result in heart-related complications. The heart may have to pump more blood to make up for the body's tissues and organs receiving less oxygen when anaemia is severe and the blood's capacity to carry oxygen is severely impaired. This increased cardiac workload can lead to chest pain or discomfort, especially during physical activity or periods of increased demand on the heart. Individuals experiencing chest pain in the context of severe anaemia should seek immediate medical attention, as untreated anaemia could contribute to further heart-related issues. Proper diagnosis and management are essential to address the underlying anaemic condition and alleviate associated symptoms.

9. Cognitive Impairment:

Chronic anaemia has considerable cognitive consequences, especially in vulnerable populations like children and the elderly. Because anaemia reduces the amount of oxygen that reaches the brain, it can lead to a reduction in cognitive function, which can cause issues with focus, memory, and mental clarity. While older adults may perceive a loss in their cognitive capacities, children with anaemia may face delays in their learning and development. To lessen cognitive damage and improve the overall quality of life for affected people, treating the underlying anaemic illness and optimising blood parameters are essential.

10. Impaired Immune Function:

Anaemia can compromise the body's immune function, rendering individuals more vulnerable to infections. Haemoglobin and RBCs play an important role in transferring oxygen to organs and tissues, including the immune system. When anaemia reduces the blood's oxygen-carrying capacity, the immune system's ability to fight off infections is diminished. Consequently, anaemic individuals may be at a higher risk of contracting various illnesses and experiencing more severe symptoms. Maintaining proper blood parameters and addressing the underlying causes of anaemia are essential to bolster immune function and enhance the body's defences against infections.

11. Growth and Development Issues (in Children):

Anaemia in children can impede growth and cognitive development.

12. Complications of Underlying Causes:

Anaemia frequently indicates a more serious health issue. Anaemia that is poorly controlled or untreated can worsen the consequences of these underlying conditions, which include cancer, inflammatory illnesses, and chronic kidney disease.

13. Complications during Pregnancy:

Anaemia in pregnant women can lead to a range of complications that pose risks to both the mother and the developing baby. Maternal anaemia can result in various inverse effects, including an increased risk of preterm birth, where the baby is born before reaching full term, low birth weight which can lead to health problems for the infant. Additionally, anaemia during pregnancy can contribute to maternal mortality, as it places additional strain on the mother's body and can lead to complications during labour and delivery. To mitigate these risks, proper prenatal care, including regular monitoring and management of anaemia, is essential to ensure the well-being of both the mother and the baby.

Mostly anaemia affects severally to women because of ability of reproduction. It also carries anaemia in child. So, there is need to treat anaemia in reproductive age women. The government implementing various schemes to reduce anaemia, from this research we can easily identify the regions which are mostly affected also the cases of anaemia. This will helpful to implement government schemes more efficiently. In the next section the various government schemes in India were briefly explained.

1.5 Government Initiatives:

According to the recent report of NFHS-5(2019-21) the burden of anaemia increases. It was found that Anaemia significantly increased in India with 57% of WRA and 67% of children (6 months to 59 months) being anaemic. The government implements various schemes for dealing with anaemia. Below that schemes were discussed:

1. Anaemia Mukht Bharat (AMB):

The Indian government launched the AMB plan in 2018, with the aim of reducing anaemia in age groups that are particularly susceptible, including as women, children, and teenagers. With six target beneficiaries, six interventions, and six institutional mechanisms, AMB is based on a life cycle approach. The government of India started the Anaemia Mukht Bharat (AMB) programme under the 'Poshan Abhiyaan,' the country's main drive to combat malnutrition and enhance nutritional outcomes. Anaemia is a prevalent and dangerous health concern in India, particularly among women and children. AMB focuses exclusively on combating this issue.

- **Key features and components of the Anaemia Mukh Bharat scheme:**

Objectives of the Anaemia Mukh Bharat: AMB's primary goal is to eradicate anaemia from India by 2030. Its goal is to lower the national incidence of anaemia, especially in susceptible populations including children, pregnant women, and lactating mothers.

2. National Nutritional Anaemia Control Program (NNACP):

The Primary Health Centres and their subcenters aim to reduce anaemia in reproductive-age women. The three main strategies are promoting iron-rich food consumption, giving high-risk groups iron and folate tablets, and Recognition and to treatment severely anaemic cases.

3. Weekly Iron and Folic Acid Supplementation (WIFS):

Under the Government of India, the Ministry of Health and Family Welfare, developed a special program for WIFS of adolescents based on these scientific studies because adolescent anaemia is a major public health issue.

Objective: To reduce nutritional anaemia in adolescents (10-19 years old)

4. Janani Suraksha Yojana:

Janani Suraksha Yojana is implemented as per 22 December 2006. In this scheme, benefits are given to Scheduled Castes, Scheduled Tribes and Below Poverty Line beneficiaries in rural and urban areas. As per Central Government circular dated 8th May 2013, the conditions related to age of beneficiaries and children have been relaxed. The initiative provides additional benefits to poor pregnant women in various states in India which have low institutional delivery rates.

Objective: The objective of this scheme is to reduce maternal mortality and infant mortality in below poverty line, scheduled caste and scheduled tribe families.

5. Pradhan Mantri Surakshit Matritva Abhiyan (PMSMA):

Women who are expecting are eligible to get free prenatal care through the PMSMA government initiative. It is possible to diagnose and cure anaemia through routine prenatal check-ups.

6. Pradhan Mantri Surakshit Matritva Abhiyan (PMSMA):

PMSMA is a government program that offers free antenatal care check-ups for pregnant women. Regular check-ups during pregnancy can help identify and manage anaemia. This program was built on the essential components of the program as well as the lessons that were learnt from it.

Government schemes for anaemia control in India reflect a commitment to address the widespread health issue of anaemia among women and children. These programs

not only focus on treatment but also on prevention through nutrition, healthcare, and awareness. By adopting a multi-sectoral approach, strengthening health systems, and raising public awareness, the government is striving to make India anaemia-free and improve the general health and wellness of its population. These schemes represent a significant step towards achieving the goal of reducing anaemia and its associated health risks in the country. However, continued efforts, monitoring, and evaluation will be crucial to ensure the success of these initiatives in the years to come.

Analysing anaemia in women is imperative as it provides crucial insights into the public health impact of this condition, particularly concerning maternal and child health, economic repercussions, and gender inequalities. For the purpose of enhancing the general health and well-being of women and their communities, it aids in the identification of high-risk groups, evaluation of the efficacy of interventions, and direction of research and creative solutions.

Previous research on anaemia in women has spanned several decades, with a focus on understanding its prevalence, causes, and consequences. Early investigations primarily highlighted the high prevalence of anaemia among women, particularly during pregnancy. Subsequent studies delved into the multifactorial causes, including nutritional deficiencies, infections, and socioeconomic factors, shedding light on the complex aetiology of the condition. Research efforts also extensively explored the adverse health outcomes associated with anaemia in women, emphasizing its role in maternal and child morbidity and mortality. As research evolved, it increasingly emphasized the importance of effective interventions, such as iron supplementation and dietary improvements, to mitigate the impact of anaemia in this vulnerable population. This historical research continuum has laid the foundation for current efforts to combat anaemia in women and underscores the need for ongoing investigations to further refine preventive and treatment strategies.

A variety of statistical techniques have been employed in earlier studies on the prevalence of anaemia in females to offer a thorough grasp of the condition's consequences. Descriptive statistics, such as mean, mode, median and sd, have been employed to quantify the central tendency and variability of haemoglobin levels in female populations. Prevalence rates, often expressed as percentages, have been calculated to estimate the proportion of women affected by anaemia in various age groups and regions. Regression analysis has been applied to assess the connection between anaemia and factors like age, socio-economic status, and dietary patterns.

These statistical approaches have not only quantified the extent of the problem but have also allowed for the identification of vulnerable subgroups and the assessment of risk factors, thereby informing targeted interventions and policy decisions to address anaemia in women.

The primary objective of this study is to determine whether or not women who are at the reproductive age are at risk for developing anaemia. The study will reveal the significant variables related with anaemia among the WRA. This study will further helpful for taking precautions for most affected region. After identifying the stage of anaemia, we will able to avoid the cases of severe anaemia by taking care of respective WRA.

In recent researches mostly, various statistical techniques have been used to analyse the anaemia in WRA. Advanced machine learning methods were used to predict the anaemia. In the next section various machine learning techniques were explained which are helpful to develop best model for exact prediction of anaemia.

1.6 Data Mining:

Data mining is a multidisciplinary field that lies at the intersection of statistics and machine learning. Nowadays, with data being generated at an unprecedented rate, it is more important than ever to be able to extract meaningful insights from large datasets. People and organisations can use data mining, a multidisciplinary discipline at the confluence of statistics and machine learning, to find hidden patterns, correlations, and trends in massive amounts of data.

1.6.1 Introduction of data mining

In depth discussion of data mining's concepts, methods, applications, and ethical issues was provided below, which also emphasises how revolutionary data mining has become for our data-driven culture. It includes a collection of techniques and algorithms intended to sort through large datasets, find significant patterns, and convert unprocessed data into insights that can be put to use. Fundamentally, data mining looks for patterns in current data to answer queries, identify previously unidentified correlations, and forecast future events. The KDD process typically involves several key steps:

Data Collection: The process of gathering or collecting data from a variety of sources, including databases, spreadsheets, or internet platforms, is referred to as data collection.

Data Pre-processing: The term "data pre-processing" refers to the process of cleaning and preparing the data by addressing any missing values, deleting duplicates, and modifying variables in order to guarantee those variables' quality.

Data Transformation: In this step data is converted to a format such that it is easy to analyse, which may include aggregating, encoding, or scaling.

Pattern Discovery: Utilizing algorithms to identify patterns, associations, or trends within the data. Common techniques are included here such as clustering, classification, and association rule mining.

Evaluation: Assessing the discovered patterns for their relevance, validity, and significance.

The use of data mining has applications in various fields, transforming raw data into actionable insights and driving informed decision-making. In business and marketing, it enables customer segmentation and market trend analysis. In healthcare, data mining aids in disease prediction, personalized treatment, and hospital operations optimization. The financial sector relies on data mining for fraud detection, risk assessment, and investment analysis. In education, it helps tailor learning experiences, predict student performance, and reduce dropout rates. Additionally, data mining supports scientific research by facilitating knowledge discovery from large datasets, such as genomics, climate science, and social sciences, while contributing to advancements across various domains. Data mining broadly build upon the various machine learning tools, so approaches of machine learning were explored in the following section.

1.7 Machine learning algorithms:

Different machine learning tasks can be addressed by different algorithms. For example, simple linear regression can be applied to prediction issues, such as stock market prediction, while the KNN method can be used to categorization problems. Artificial intelligence (AI) and data science depend significantly on machine learning techniques. From given data computer are learn and make predictions or judgements without explicit programming. These algorithms are made to recognise trends, categorise data, forecast outcomes, and enhance their functionality over time. Machine learning algorithms have several forms and can be divided into three primary groups:

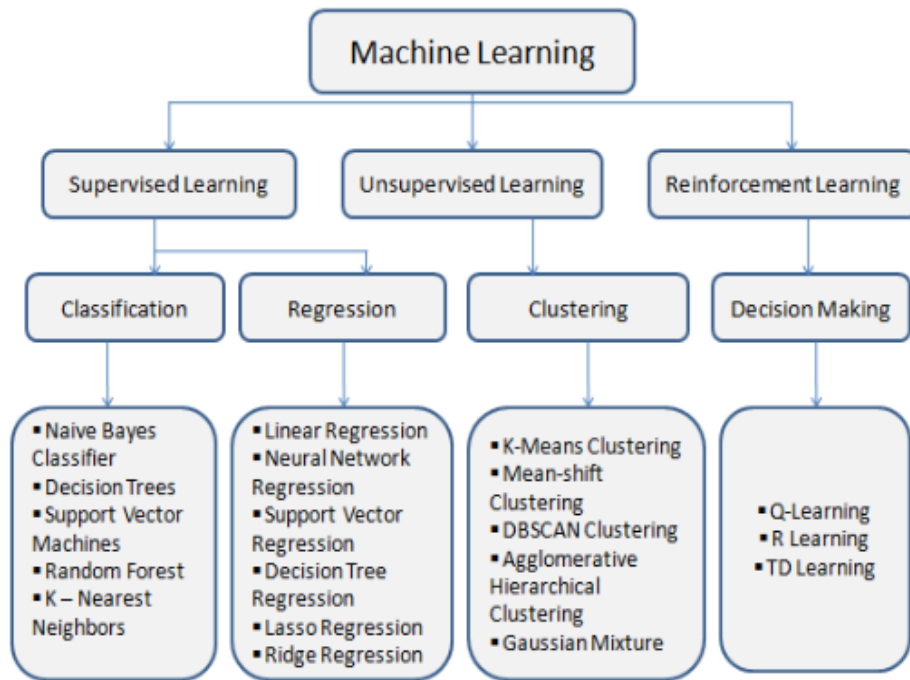


Fig 1.1 Machine learning algorithms ((Courtesy: Google)

1.8 Supervised Learning:

In the process of supervised learning, algorithms are trained/developed with the assistance of data that has been labelled. In this type of learning, each data point is associated to a certain target or outcome that has been specified initially. One kind of machine learning is supervised learning, in which the machine must have outside supervision in order to learn. To train the supervised learning models labelled dataset is used. After training and processing, a sample test set of data is provided to the model, which is then tested to see if it can correctly predict the output. The supervised learning is the same as when a student learns under a teacher’s guidance. Spam filtering is a prime example of supervised learning. Regression and classification are the two further issue groups into which supervised learning falls. Popular supervised learning methods include the KNN algorithm, Decision Trees, Logistic Regression, Simple Linear Regression, and others.

1.8.1 Regression analysis:

1. Linear Regression:

Regression analysis is a well-known method of supervised learning that is helpful in determining the relationship between variables. It also gives us the ability to predict the continuous response variable, which is sometimes called as the dependent variable, by using one or more independent variables, or simply predictors. There is a stepwise

procedure for developing regression model. If there is one response variable and one independent (predictor) variable then in the first step we have to examine the relationship between both variables by using scatter plot. If the scatterplot shows the relationship between them, then we can move towards the regression model building step. In this step a simple straight-line equation is derived to estimate the dependent variable.

$$Y = \beta_0 + \beta_1 * X \text{ -----(1.1)}$$

Where,

Y shows the response variable or dependent variable which we want to predict.

β_0 is an intercept

β_1 is slope parameter

X indicates the predictor or independent variable which is used to predict/estimate the response variable.

On the basis of number of predictors, the simple linear regression has two types i.e. Simple linear regression and multiple linear regression. Multiple predictor variables are used in the MLR to estimate or predict the response variable. This allows for more accurate results. Here we can say that the response variable is depends on more than one independent variables. Suppose we have k independent variables then the regression line of MLR is as follows:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \dots + \beta_k * X_k \text{ -----(1.2)}$$

While studying linear regression there are some assumptions,

1. Linearity: There should be linear relationship between the dependent/response variable and independent variables.
2. Independence: Errors are independent.
3. Homoscedasticity: The error variance is constant.
4. Normality: The residuals (differences between actual and predicted values) are normally distributed.

2. Polynomial regression:

In polynomial regression analysis, the regression model is developed by using as nth-degree polynomial equations by extending the concepts of SLR. Linear regression is expanded upon by this method. It can capture more complex patterns in the data by introducing polynomial terms. Suppose we have a dataset contains datapoints that exhibit non-linearity. In this scenario, linear regression is not the most appropriate fit

for the datapoints in the dataset. Polynomial regression is required to cover such datapoints.

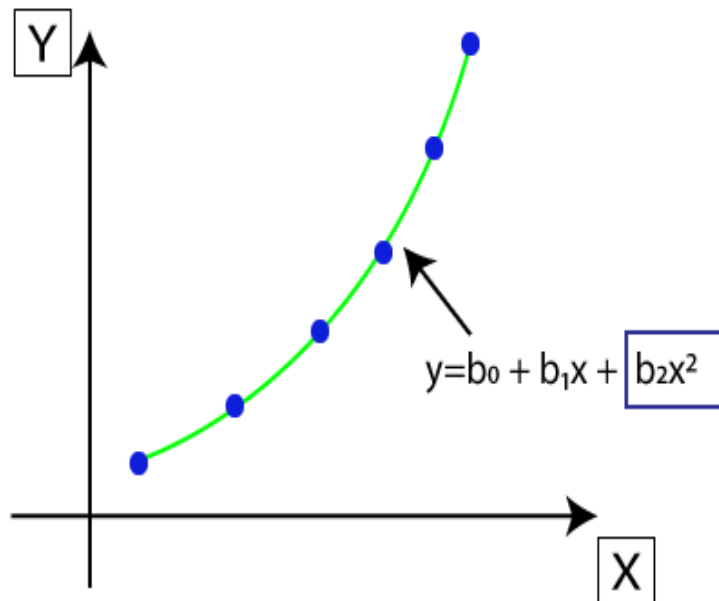


Fig 1.2 Polynomial curve (Courtesy: Google)

In polynomial regression, a linear model is used to model the original characteristics after they have been converted into polynomial features of a specified degree. This indicates that a polynomial line fits the datapoints the best. The model equation of polynomial regression will be defined by using LR equation that means if we have linear regression equation is, $Y = \beta_0 + \beta_1 * X$ then the Polynomial regression equation can be written as follows,

$$Y = \beta_0 + \beta_1 * X + \beta_2 * X^2 + \dots + \beta_n * X^n \text{ ----- (1.3)}$$

The above polynomial regression model is linear model since the model coefficients are linear in nature.

3. Non- linear regression:

Nonlinear regression has several uses since a large number of real-world data sets fail to establish a linear relationship. Similar to linear regression modelling, nonlinear regression modelling aims to visually trace a certain response variable from a group of predictor variables. Nonlinear regression models are more complicated than that of linear regression models because the nonlinear equation is created by iterations that may stem from trial-and-error. Several methods are established to develop the non-linear model equation, such as the Gauss-Newton method and the Levenberg-Marquardt method.

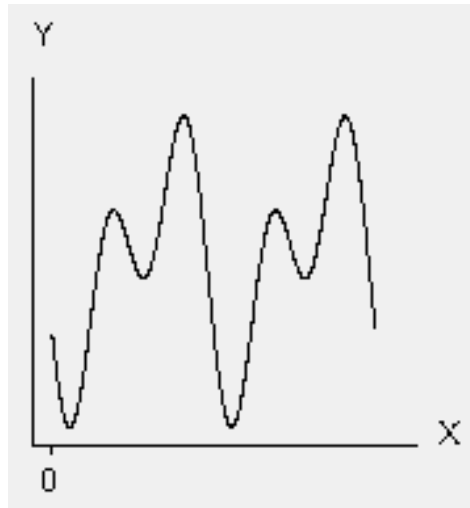


Fig 1.3 Non-linear regression (Courtesy: Google)

1.8.2 Classification

Assigning data points to predetermined groups or classes according to their properties or attributes is the method of classification, which is a fundamental concept of data mining and machine learning. This method, learns patterns from labelled data, is essential to data analysis because it allows automated decision-making and prediction. There are a wide variety of applications for classification algorithms, including the identification of pictures, the diagnosis of diseases, and the detection of spam emails.

In modern word we can say classification as supervised machine learning. The applications of classification are far-reaching. In healthcare, it aids in the diagnosis of diseases by analysing patient data and medical images. In finance, it is crucial for fraud detection and risk assessment. In natural language processing, text classification categorizes documents for sentiment analysis and information retrieval. E-commerce platforms use it to recommend products based on user preferences and behaviours.

Classification will continue to play a crucial part in data-driven decision-making and automation in the future, and it will have an impact on a variety of disciplines, including recommendation systems, personalized medicine, and autonomous cars. As machine learning techniques evolve, so too is the capabilities and applications of classification in our increasingly data-driven world. Following are some machine learning techniques and their types.

1. Logistic Regression:

A well-known example of a supervised machine learning method is known as the Binary-Logistic Regression (BLR), which is generally used for the classification

task using a one or more independent variables. It's specifically designed to handle problems where you need to classify data into two distinct groups, such as 'Yes' or 'No,' '0' or '1,' 'true' or 'false.' However, instead of delivering strict 0 or 1 results, logistic regression provides probabilistic values between 0 and 1.

Logistic regression is a statistical technique that involves fitting a curved 'S'-shaped logistic function rather than a straight line. This function predicts two extreme values, namely 0 and 1. The shape of this curve provides an indication of the probability that an event will take place. When it comes to the topic of machine learning, logistic regression is an extremely useful technique because it can provide probabilities and make classifications using both continuous and discrete data. This makes it a versatile tool that can be used in a variety of problems.

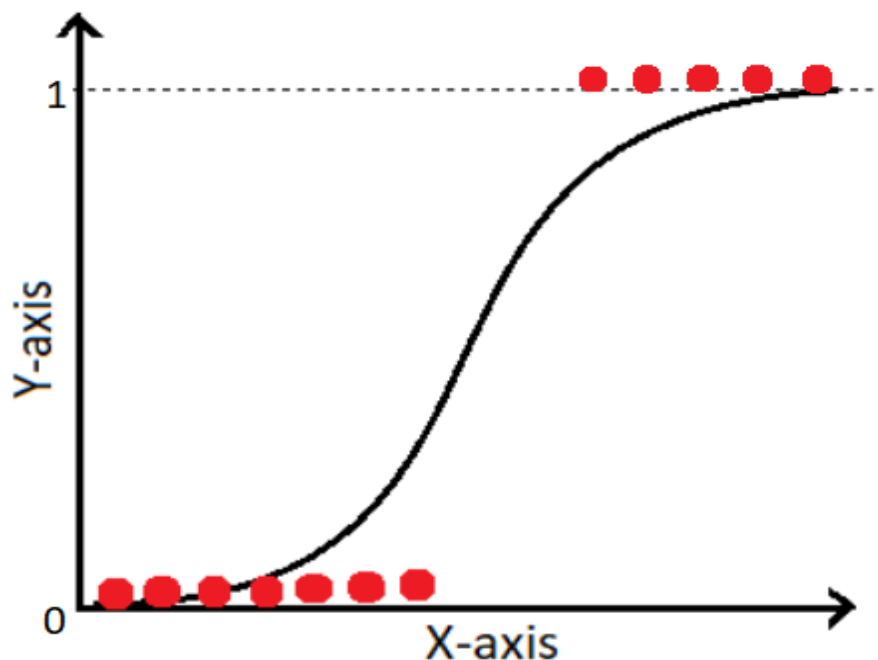


Fig 1.4 Logistic graph (Courtesy: Google)

Types of Logistic Regression:

Considering the nature of the classes the LR classified into three types such as Binary or Binomial logistic regression, Multinomial logistic regression and ordinal logistic regression.

Binomial: In binomial Logistic regression dependent variable has only two classes.

Multinomial: The dependent variable in multinomial logistic regression may have more than three different unordered classes such as 'cat', 'dogs', or 'sheep'

Ordinal: In the Ordinal-Logistic Regression dependent variable has more than 3 classes which are ordered such as 'low', 'Medium', or 'High'.

2 Decision Tree:

The DT is a well-known ML technique utilized in classification and regression applications. With each node represents a features or attributes and branches represents a outcomes or decisions, and leaf nodes indicating final predictions or classifications, it resembles an inverted tree structure. The goal of decision trees is to produce straightforward yet powerful guidelines for decision-making by recursively dividing the dataset according to the most informative attributes. They can handle both numerical and categorical data. Decision trees are simple to understand and apply. Despite their tendency for overfitting, decision trees are frequently included as a component of ensemble techniques such as random forests in order to enhance their robustness and predictive ability.

3 K- Nearest Neighbour:

A straightforward but efficient machine learning approach called K-Nearest Neighbours (KNN) is utilised for classification as well as regression type problems. It functions according to the idea that target values are typically similar across data points with comparable attributes. Since KNN is non-parametric and instance-based, it may be used to a variety of complicated and diverse datasets without requiring any assumption about the nature of the data. It may, however, be affected by the selection of the 'k' parameter and the distance metric that is employed to measure the similarity of the data points. Minkowski, Manhattan, and Euclidean distances are examples of common distance measures in KNN algorithm. The type of data determines which distance measure is suitable.

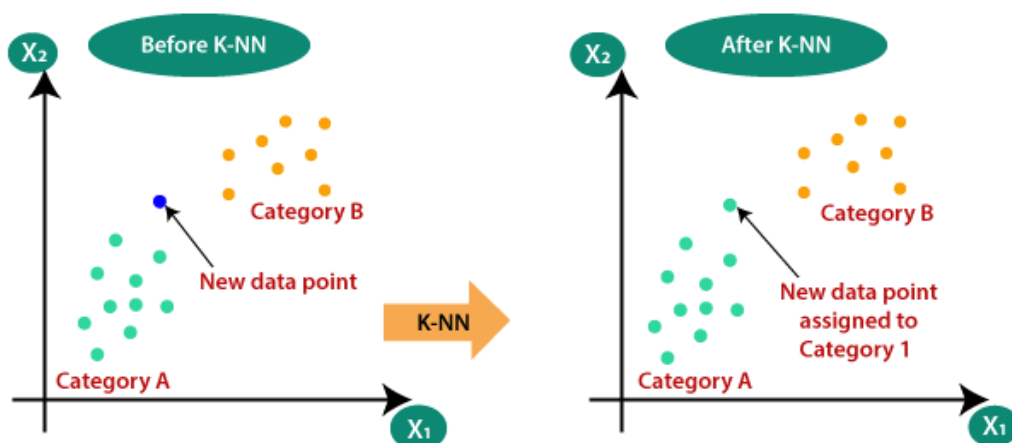


Fig 1.5 K-NN classifier (Courtesy: Google)

Regression and classification tasks are two applications for KNN. The anticipated class in classification is determined by the majority class among the ‘k’ nearest neighbours. The average or weighted average of the target values of the k nearest neighbours is used in regression type problem.

4 Naïve Baye’s Algorithm:

Naïve Bayes is a supervised machine learning algorithm which is used for classification task. It’s based on Bayes’ theorem and the assumption of conditional independence. Naïve Bayes algorithm procedure works by calculating the probability of a sample point belonging to each class and then assigns it to the class which has highest probability. Text classification tasks like sentiment analysis and spam detection are especially well-suited for it. Despite its simplicity and the simplifying assumption of independence, Naïve Bayes can perform effectively and is especially useful when you have limited data or need a quick and reliable classification solution. When the independence assumption is significantly violated then Naive Baye’s may not work properly.

5 Support Vector Machine (SVM):

The SVM is popular machine learning method for both classification and regression applications. Its capacity to identify the ideal hyperplane which divides data/sample points into distinct classes/categories and maximises the margin between the that classes. In order to determine the optimise position of the separating hyperplane, support vectors which are subset of data points are identified by SVM.

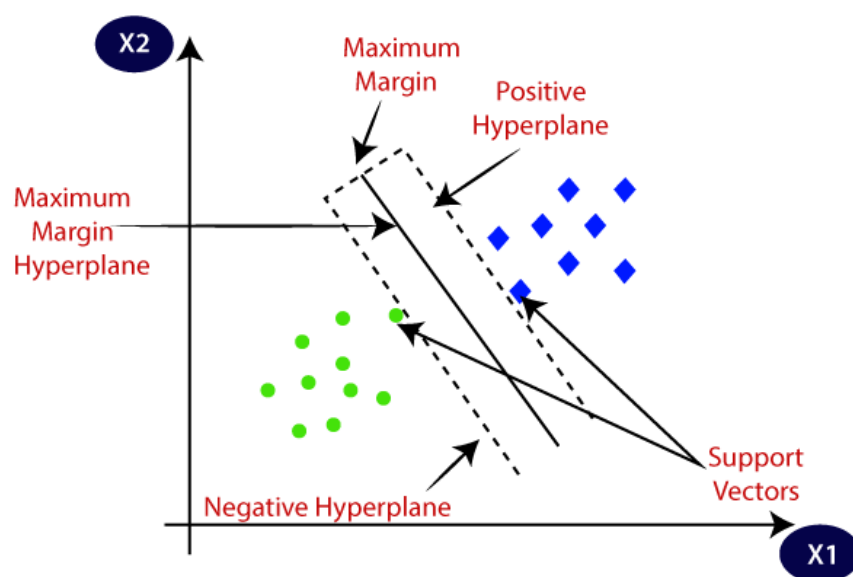


Fig 1.6 SVM with linearly separable data (Courtesy: Google)

SVM uses several kernel functions to translate the original data into a higher dimensional space, it performs especially well when the data is high-dimensional and have non-linear decision boundaries. SVM is renowned for its durability and capacity to manage noisy data. It can offer a powerful, broadly applicable solution, but its effectiveness depends on selecting the right kernel function and regularisation settings. SVMs are also frequently utilised in situations where precisely predicting outcomes requires the identification of a distinct boundary between classes, such as picture and text categorization.

1.9 Unsupervised learning algorithms:

Unsupervised learning is also the machine learning techniques where the algorithm is given unlabelled data and tasked with discovering patterns, structures, or inherent relationships within the data. Unlike supervised learning, which relies on labelled examples to make predictions or classifications, unsupervised learning operates on its own, seeking to uncover hidden insights or groupings in the absence of guidance. Common techniques in unsupervised learning include clustering, in which the data points as well as variables are grouped in the form of clusters based on their similarity. Unsupervised learning algorithms are valuable for tasks like data exploration, anomaly detection, and feature extraction, as it can reveal underlying structures and patterns that might not be apparent through manual inspection, making it a fundamental component of data analysis and AI.

1.9.1 Clustering:

In data mining and machine learning, clustering is a basic technique that includes putting similar data points in groups according to shared traits or patterns. It is extensively employed in many different domains, including as natural language processing, image processing, and data analysis. Algorithms for clustering data are essential for revealing hidden patterns in datasets and supporting data exploration. Finding naturally occurring groups or clusters within a dataset is the main goal of clustering. The similarity or dissimilarity of data points based on selected criteria or attributes defines these groups. There are various clustering algorithms, each have their own strengths and weaknesses. Some of the most commonly used ones include:

- 1. K-Means Clustering:** Data points are grouped into K clusters and the centroid of each cluster is used as a representation of cluster. The main goal of the computation is to minimize the total squared distances between the centroids that are associated with every data point.

- 2. Hierarchical Clustering:** This technique builds a hierarchical representation of clusters, forming a tree-like structure known as a dendrogram. It allows for a flexible approach to cluster extraction. In this method there are three techniques such as single linkage method, average linkage method and complete linkage method each have their own advantages and disadvantages.

1.10 Reinforcement Learning:

Machine learning includes the discipline of reinforcement learning. It involves performing the appropriate actions to maximize reward in an instance. It is utilized by a variety of machines and software to determine the optimal course of action or conduct in an instance. Supervised learning differs from reinforcement learning in that the answer key is included in the training data for supervised learning, ensuring that the model is trained using the correct answer. In contrast, RL operates without an answer, with the reinforcement agent determining the appropriate course of action to accomplish the given task. It is inevitable that without a training dataset, it will acquire knowledge through experience.

The choice of ML algorithm is determined by the particular task, the type of data, and the project objectives. To find the optimum method for a particular problem and dataset, data scientists and machine learning engineers in practise frequently test out many versions of the algorithms.

For the statistical analysis purpose, the various softwares plays crucial role. In the next section those softwares were explained which were used in the study.

1.11 Software:

There is many software used in predictive analysis such as SPSS, MINITAB and R.

1.11.1 Minitab:

Minitab is a statistical software which is used for statistical analysis. It provides various tools and features that are especially useful for professionals in various fields, including business, engineering, healthcare, and academia. Here are some essential details regarding Minitab:

- **Statistical Analysis:** Data analysis, hypothesis testing, regression analysis, analysis of variance (ANOVA), multivariate analysis and many other statistical tasks can be performed using Minitab. It covers both fundamental and advanced statistical methods.

- **Data Visualization:** Minitab provides a variety of data visualization techniques, such as scatterplots, histograms, box plots, and control charts, making it easier to understand data and present results.
- **Data Manipulation:** By using Minitab we can easily import, clean, and manipulate data for efficient data preparation.
- **Simple to Learn:** Minitab is well known for its simple user interface, and it provides a number of resources, including tutorials and online assistance, to help users learn how to use the software properly.
- **Integration:** Minitab is an adaptable data analysis tool that may be combined with other popular software applications, such as Microsoft Excel.
- **Industries and Applications:** A variety of industries, including manufacturing, healthcare, finance, and education uses Minitab.
- **Versions and pricing:** Minitab have several software versions, including Minitab, Minitab Express, and Minitab Workspace, each suited to distinct requirements and price ranges. The version and licencing choices affect price.

1.11.2 R software:

R- software is a free and open-source programming language which is primarily designed for statistical computing and data analysis. Here are some important details regarding R software:

- **Statistical Analysis:** R is well-known for its vast library of statistical and graphical tools. It gives a comprehensive range of tools for time-series analysis, linear and nonlinear modelling, hypothesis testing, and data analysis.
- **Data Visualization:** Powerful data visualisation capabilities are provided by R. Users can design a wide range of graphs and charts, including scatterplots, histograms, bar charts, box plots, and customised visuals, to effectively explain their findings.
- **Data Manipulation:** R supports cleaning, transforming, and manipulating data. Tasks including data reshaping, combining, filtering, and aggregating are supported by the language and are essential for getting data ready for analysis.
- **Machine Learning:** R is widely used in the fields of predictive modelling and machine learning. There are numerous packages like ‘caret’ and ‘randomForest’ that can be used to create machine learning algorithms.

- **Industries and applications:** R is extensively utilised in a wide range of sectors, including academics, healthcare, finance, marketing, and more. Statisticians, data scientists, researchers, and analysts favour it as a tool. It is relevant in a variety of sectors.
- **Cost and licensing:** You are free to download, use, change, and distribute it without paying any licencing fees.

1.11.3 SPSS software:

A programme that is used for statistical analysis, data mining, and decision support is called SPSS, which stands for Statistical Package for the Social Sciences. It was created by IBM (International Business Machines Corporation) and is frequently used for data analysis and research across many industries. Here are some essential details regarding SPSS:

- **Statistical Analysis:** SPSS can be used for a variety of statistical analyses since it offers a complete collection of statistical tools and methodologies. Users can perform descriptive statistics, factor analyses, correlation analyses, regression analyses, and many more.
- **Data management:** SPSS enables effective data management, cleaning, and preparation. Working with datasets is made easier by its support for data entry, transformation, and manipulation operations.
- **Data Visualisation:** To aid users in understanding data and presenting findings in a visually appealing manner, the software provides a variety of data visualisation choices, such as bar charts, histograms, scatterplots, and more.
- **Advanced predictive analytics features:** Advanced predictive analytics features such as decision trees, logistic regression, and cluster analysis, are available in SPSS and are useful for data mining and predictive modelling.
- **Customization and Automation:** For individuals who would rather not code, SPSS also provides a convenient point-and-click interface. Users can write custom functions and scripts to automate repetitive activities.
- **Survey Research:** It has elements for survey research, including survey data analysis and questionnaire design.
- **Reporting and Output:** SPSS produces thorough output reports with tables and charts that are useful for describing and presenting analytic results.

1.12 Problem in Hand:

In Maharashtra, anaemia is still a widespread and serious problem for WRA, affecting the health of both mothers and the unborn child. The anaemia prevalence remains seriously high in spite of several interventions, having a significant negative influence on mother and child health outcomes. It is crucial to address this issue if we are to enhance the general wellbeing of women and children in the area.

Developing accurate and accessible predictive algorithms to predict anaemia among reproductive-age women is a pressing need. Early identification of individuals at risk of severe anaemia can facilitate targeted interventions and preventive measures, ultimately improving the health outcomes of women in this demographic. The challenge lies in creating predictive models that are reliable, cost-effective, and accessible to healthcare providers and women themselves.

This problem statement highlights the need for research to create predictive algorithms for anaemia risk, which could involve factors such as individual, household and demographic. Addressing this problem can lead to the development of campaigns and strategies that empower healthcare professionals and women to proactively manage anaemia risk.

1.13 Research objective:

- To find the anaemia prevalence among the WRA in Maharashtra.
- To identify the key factors associated with anaemia among the WRA.
- Develop models to predict Anaemia
- Develop a best machine learning model to predict anaemia among women at reproductive age.
- To predict the anaemia status of the reproductive age women (WRA).

1.14 Scope of the research:

The research aims to predict anaemia risk in women at reproductive age (WRA), with a focus on a Baramati region. This study seeks to develop a predictive algorithm for anaemia by analysing various factors including personal, socioeconomic, marital, lifestyle, dietary, demographic level factors. There is necessity of developing a highly accurate predictive algorithm to predict the status of anaemia by using these above factors. Therefore, the present study has scope to find accurate stage of anaemia of that respective WRA. By identifying the factors that influences the status of anaemia, this research has the potential to inform targeted public health interventions, policies, and

programs aimed at preventing and managing anaemia in reproductive-age women in the chosen region. The findings will contribute to the field of public health and could have practical applications for healthcare providers and policymakers.

1.15 Outline of thesis:

There are total 6 chapters in this thesis. Chapter-1 is introduction of the thesis. It contains research background, introduction of anaemia, causes of anaemia, consequences of anaemia, prevalence and trends of anaemia, importance of addressing anaemia, various ML algorithms and their types.

Chapter-2 comprises a complete review of the previous work on prevalence of anaemia, prediction of anaemia using machine learning algorithms, and the ensemble methods. This chapter includes literature before 10 years till date.

Chapter-3 is devoted to methodology of proposed thesis. This chapter contains the concepts of data, data types, data pre-processing and its need, data visualisation and its importance, data analysis, pilot study, DHS data description, supervised machine learning algorithm and its various algorithms, and ensemble methods.

Chapter-4 includes the pilot study on DHS data. Various machine learning algorithms with ensemble techniques is developed on this DHS data.

In Chapter- 5 various machine learning algorithms on the primary data which was collected by pre-designed questionnaire are developed.

The full conclusion of the suggested work, which is covered throughout the doctoral work and future study discussion, is the focus of Chapter 6.

A reference section is designated for Chapter 6.

1.16 Terminology:

In this research study various terms are used in classification algorithms.

- **Sample:**

The term ‘sample’ is frequently used for subset of data and also commonly known as training data sample.

- **Dependent variable:**

The dependent variable, response variable, outcome variable, class and target variable are referred as outcome of the values of target variables in the data.

- **Independent variable:**

The independent variables, predictor variables, regressors, features or attributes are referred as independent variables in the data which are used to predict target variable.

- **Continuous variable:**

The continuous variables are used numerical scale. For example, BMI of WRA is continuous variable as it measures on numerical scale.

- **Categorical variable:**

Nominal, attribute, or discrete variables are terms used to describe the category variables. It cannot measure on any numerical scale. For example, pregnancy status like Yes, No.

- **Classifier:**

A classifier is a model or algorithm that assigns labels or categories to input data based on patterns and features it has learned from a training dataset.

- **Confusion matrix:**

By summarizing performance of the model and displaying the counts of true-positive, true-negative, false-positive, and false-negative predictions, a confusion matrix provides a comprehensive overview of the model's accuracy and error rates in a classification model.

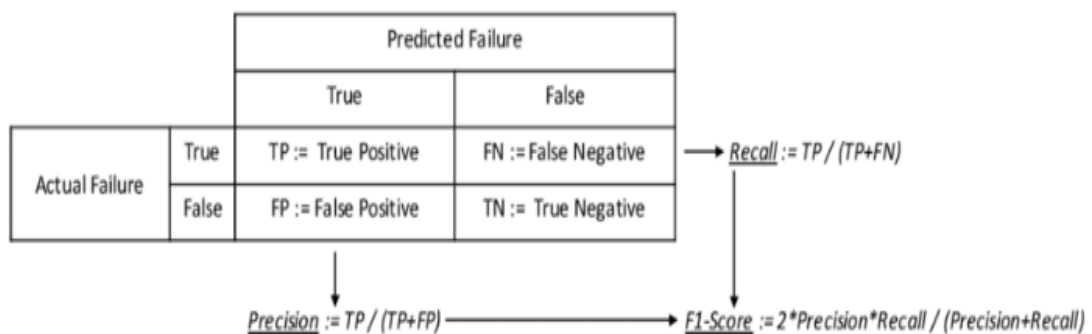


Fig. 1.7 Confusion matrix (Google courtesy)

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction:

This chapter is on literature reviews a critical synthesis of prior research and scientific work that gives the reader a thorough picture of the state of knowledge in the subject area. It acts as the framework for the research, assisting in the gap analysis, establishing the study's setting, and highlighting the importance of the research. The prediction of anaemia in WRA will be discussed in this chapter along with relevant research, hypotheses, and approaches, with an emphasis on dietary practises, socioeconomic impacts, and health outcomes. I intend to identify the current knowledge landscape, highlight areas that require additional research, and establish the framework for the research process by critically assessing the available literature.

The next section contains the literature reviews and findings are discussed in the following section.

2.2 Literature Review:

Nasim Uddin (2023) et. al. have developed innovative loan prediction system by using machine learning (ML) and intended to identify qualifying loan applicants automatically. In order to predict loan recipients, the best model must be found and integrated with the user interface. A web application created by the author determined whether or not a consumer is qualified for a loan. The data was pre-processed, including data partitioning into training and testing sets, as shown in the following image. After nine Machine Learning algorithms were taught, the best three models were utilized to construct an ensemble model. The performance of the model was assessed using the F1 score, recall, accuracy, and precision. Some of the most important criteria for determining loan acceptance include the following: gender of the applicant, marital status, dependents, education, self-employment, income of the applicant and any co-applicants, loan amount, loan term, loan requirements regarding credit history, and property valuation.

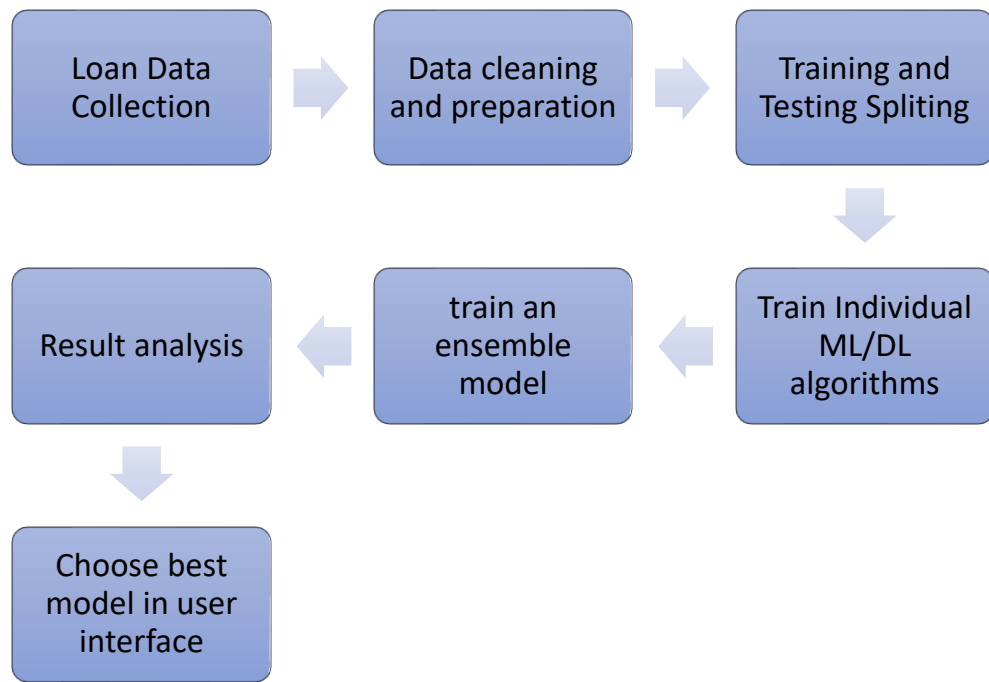


Fig 2.1 Flow of research

A Kaggle dataset on the loan prediction was used for this research. The initial dataset was found to be unbalanced. Two methods were employed to balance the data after it was discovered that it was out of balance. SMOTE is one method for balancing the dataset. The authors used a different strategy that involved using the existing data to train a straightforward machine-learning model in order to attain dataset balance. as soon as the dataset was almost in balance. The dataset, which was used to forecast loan recipients, had 806 rows and 13 columns. To create the various machine learning models, data was pre-processed. Data was first split into two sets train and test sets, with 75-25 pattern.

Author developed a technique that can predict whether or not a loan would be approved. The system architecture and method for choosing the best model that was connected to a user interface are shown in the following figure. In First, the author trained nine various machine learning models and evaluated their performance. The author developed two ensemble models in the second section: one with the nine classifiers and the other with the top three models in terms of performance. The performance of developed ensemble approaches was evaluated. The model with the highest accuracy was chosen in the end to be used in a desktop application. In order to predict loan defaulters, a user application was eventually created and the best ML model was used. The User Interface (UI) was created by the author using the Python Tkinter

package. A visual representation of bank loan availability checks is made possible by that interface. The following procedure was used to connect the best model to the UI.

Jeetendra Yadav, et. al. (2021) explored the significant determinants of anaemia, as well as the places where under-fives have low anaemia status. Anaemia can have an adverse influence on a child's mental development and social performance. During their first two years of life, children with iron deficiency anaemia have weaker brain development; poor school performance. Childhood anaemia is very serious problem. In this study the author identified the factors associated with childhood anaemia for that purpose the secondary data was collected from fourth round of the National Family Health Survey (NFHS-4). This survey included a total of 145,924 youngsters from 640 districts across India. The four types of anaemia were determined by the standards given by the National Health Mission. Individuals were categorized into two groups for the purpose of conducting multi-level analysis: those with anaemia and those without. The multilevel LR model was fitted to study the individual, community and district level factors on the Anaemia. In the first level null model was examined which consists of no predictor at all in the second level individual and village factors were taken and in the third level individual, village and district level factors were included. VIF was used to check the multicollinearity between the predictors. From this study author concluded that 60.13 % children were anaemic in which 30.69% mild anaemic, 27.82% moderate anaemic, and 1.62% severe anaemic. Also, it was found that majority of the childhood anaemia in rural areas. From the multilevel LR results author concluded that child's current age, size of child at birth, birth interval, mother's age at the time of birth, education of mother and father, religion, social group, wealth index/quintile, and region of residence are significant factors for the childhood anaemia.

Subba Rao Polamuri (2021) described an approach for predicting heart disease using AI techniques. The data was taken form UCI storehouse. Data contains 303 record after data pre-processing total 297 patients were taken for analysis. After the data pre-processing DT, BLR, RF, SVM, GBT, etc. were built and the model performance was evaluated by several measures like exactness, accuracy and mistakes. Among all the developed models the GBT(Gradient Boosted Decision Tree) and Naïve Baye's shows maximum accuracy.

B. Meena Preethi. et. al. (2021) developed a system to categorise medical records according to whether they are diseased or not and determine which diseases are becoming more prevalent. The Radiology reports that have been collected from

hospitals and scan centres are the text documents that make up the data set for this study. To generate a training set, the radiology reports in the (.doc/.txt) file are categorised based on classes and saved in the training data base. After the data pre-processing the ensembling models like mean feature voting ensemble classifier, Nearest mean classifier, KD Tree KNN, Random Forest were developed. Accuracy was measured for each fitted model for performance evaluation. It was discovered and mean feature voting classifier (70.1818%) shows maximum accuracy than other ensemble algorithms. According to the author, the machine learning algorithm did an adequate job of classifying the disease condition in the medical records.

Mirza Muntasir Nishat et al. (2021) Predict the Diabetes Mellitus by using machine learning algorithms. For this purpose, kaggle data was used. There are 2000 instances in the data, each having 9 features. 1600 of the 2000 samples were used for training the model, while the remaining 400 were used for testing the model. The data pre-processing was done by using Pandas. Missing values are replaced with respective means. After the data pre-processing the ten ML models with cross validation method that is 5- fold was built on the data. Also, hyper parameter tuning technique was used for better performance of machine learning models. After building models, model performance was evaluated by using confusion matrix. From the confusion matrix accuracy, sensitivity, specificity, precision, F1 score and ROC_AUC are calculated. After finding these parameters it was observed that the machine learning algorithm Gaussian Process (GP) provides more accurate predictions with accuracy (98.25%) and higher performance than the other methods. Also, Random Forest and Artificial Neural Networks had 97.25% and 96.5% accuracy respectively, which was also considerable rather than other algorithms.

Mithun Mog et.al (2021) established how common anaemia was in Maharashtra's districts. Secondary data for this study were obtained from NFHS-4 between 2015 and 2016. This study evaluated earlier NFHS-4 data with the recently released NFHS-5 factsheet and the body of existing literature. Bivariate analysis was done to examine the prevalence of anaemia, and ArcGIS 10.8's spatial analysis software was utilized by the researcher to determine the prevalence in area. STATA software was used for data analysis. Among WRA, the district of Nandurbar had the highest rate of anaemia (60.22%) in 2015–16, while the district of Washim had the lowest (35.46%). In contrast, Sindhudurg district founds the lowest prevalence of anaemia (41.20%) and Gadchiroli had the highest (66.20%) in 2019–20. The districts of Washim, Wardha, and

Buldhana exhibit the greatest improvements. Thus, there were significant health problems. In contrast, there have been negative changes in Sindhudurg, Mumbai City, and Ratnagiri, which indicates that the health status of women has improved. In Maharashtra, the spatial prevalence of anaemia among married women was as follows in 2015–16: out of 36 districts, 16 had a greater prevalence of anaemia (more than 55%), while 12 districts had a 50% prevalence and 5 districts had a prevalence of less than 50%. Of the 36 districts in 2020, only two had the highest prevalence of anaemia; seven had a 50% prevalence, and 27 had a prevalence of less than 50%. According to data on anaemia from 2015 to 2020, 16 districts out of 36 showed a greater prevalence of anaemia. This indicates that health problems were a very major concern in sixteen districts.

Lire LemmaTirore et.al. (2021) employed Multilevel ordinal logistic regression to identify the characteristics at the personal, domestic, and community/social levels that are linked with anaemia in WRA. The Ethiopia Demographic and Health Survey (EDHS) data set, which was gathered from Ethiopia's two administrative cities and nine regions, provided secondary data. A two-phase sampling technique was utilized to choose residences and enumeration zones. A total of 645 EAs were chosen in the initial phase, with 202 located in the urban areas and 443 in the rural areas. A total of 18,008 houses were chosen in the second stage of the study through the utilization of systematic sampling. The women of reproductive age were then selected from these households. There were total 14489 WRA were selected for this study. The individual household and community level variables were obtained from the EDHS data. Individual-level factors were extracted, including age, women's educational attainment, religion, marital status, media exposure, alcohol intake, khat use, maternity status, history of abortion, method of contraception, the number of births in the previous 5 years, use of deworming drugs, iron supplementation, nutritional counselling, and HIV status. Factors at the household level include household size, the wealth index, the type of fuel used for cooking, the availability of clean drinking water, and the state of the toilet. Place of residence, area, unemployment, poverty, exposure to the media in the community, and women's education are among the community-level determinants. The dependent/response variable in this study was the presence of anaemia. Anaemia status was used as a dependent variable. According to WHO anaemia categorised as no-anaemia, mild-anaemia, moderate-anaemia and severe-anaemia. if Hb level is greater than 12g/dl then it is categorised as non-anaemic, if Hb level is 11.0 to 11.6 g/dl then it

is categorised as mild anaemia, if Hb level is 8.0 to 10.9 g/dl then it is categorised as moderate anaemia and if Hb level is below 7 g/dl then it is categorised as severe anaemia but in pregnant women categorisation is different. In pregnant women if Hb level is greater than 11g/dl then it is categorised as non-anaemic, if Hb level is 10 to 10.9g/dl then it is categorised as mild anaemia, if Hb level is 7 to 9.9g/dl then it is categorised as moderate anaemia and if Hb level is 4.0 to 6.9g/dl then it is categorised as severe anaemia. For statistical analysis STATA software was used. The multilevel logistic regression model was developed. Bivariate analysis was done firstly to identify significant variables. The variables which are significantly associated with anaemia were further used in multivariate logistic regression. Estimates were made for adjusted odds, along with a 95% confidence interval. Variance partition coefficients (VPC) were utilized to determine the fraction of variability in the likelihood of anaemia that may be attributed to differences between households and communities. The researcher employed the AIC to determine the optimal model for the given data. All models' AICs were evaluated, and the best model was selected according to best AIC. The appropriateness of the model was verified by examining the random effects' normality. Since the random effects of the cluster and the household were roughly normally distributed, the fitted model was sufficient. There were a variety of factors that have been linked to anaemia, including HIV infection, pregnancy, a higher number of births, and living with a husband. Nevertheless, there was an adverse correlation between anaemia and secondary and tertiary education, as well as the utilization of contraceptive pills, implants, or injectables. Anaemia shown a positive correlation with residing in households characterized by a high number of family members and lower levels of wealth, including the poorest, poorer, and intermediate wealth index. An association was found between anaemia and residing in rural areas within the Afar, Harari and Somali regions, as well as in Dire Dawa city.

Morshedul Bari Antor et.al.(2021) have used some popular data mining algorithms like SVM, LR, DT, and RF to predict Alzheimer or Dementia in different patients from the MRI data. There are eight characteristics that make up the data, and they are as follows: gender (male or female), age, year of schooling, socio-economic status of patient, mini-mental state examination (MMSE), estimated total intracranial volume (ETIV), normalized whole brain volume (WBV), and Atlas scaling factor (ASF). Correlation matrix was used to find interrelationship between attributes. From the correlation matrix it was observed that higher the ASF and SES values, the more likely

you are to get dementia. After fitting the data mining models, model evaluation and comparison was done by using ROC, AUC and confusion matrix. It was observed that Support Vector Machine (SVM) gave best results rather than other models.

Dev Ram Sunuwar et.al. (2021) analyse the factors which significantly contribute to anaemia in WRA by using multilevel mixed effect LR model. An estimation of the rate of anaemia was also conducted through the utilization of spatial analysis. The data was taken from NDHS 2016. The reproductive women (WRA) from the NDHS 2016 comprised the research population. The outcome variable that is anaemia status is decided according to the WHO guidelines. The terms "anaemic" and "not anaemic" were used to further categorize the anaemia groups. As predictors, the sociodemographic, individual, household, and community characteristics were chosen. Standard deviation, mean, weighted frequencies, and weighted percentages were all used in descriptive analysis in statistical study. For categorical variables, the Pearson χ^2 test was employed, and for continuous variables t-test was used. All statistical analysis done by using STATA software. The author found that among WRA in Nepal, anaemia was quite common. In this study, anaemia affected more than 40% of the WRA. Anaemia was less common in middle-class and older women, although it was more common in those with less formal education and those who used chemical contraception. Women residing in Province 2, low community levels, and female education levels have all been linked to higher rates of anaemia.

Jahidur rahman khan et.al. (2021) created a number of ML algorithms to predict childhood anaemia in Bangladesh. This information has been obtained from the 2011 BDHS. This study included 13 children aged between 6 months to 59 months. Then, to predict the status of childhood anaemia, machine learning methods of classification such as DT, LDA, K-NN, SVM, RF, and binary LR were created. All machine learning models' performance was evaluated by computing several performance metrics such as accuracy, sensitivity, specificity, and AUC. Following the investigation, the author discovered that the RF model had higher accuracy (68.53%), 70.73% sensitivity, 66.41% specificity, and 0.6857 AUC. Also, BLR models demonstrated significant accuracy (62.41% with 63.41% sensitivity, 62.11% specificity, and 0.6276 AUC). Among all algorithms, the k-nn has the lowest accuracy.

Sagar Yeruva et.al. (2021) Have developed various machine learning models for prediction of sickle cell anaemia. SCA is one kind of blood disorder that affects the haemoglobin in red blood cells (RBCs). The sickle cell has a disc form that is sticky

and hard, causing blood flow to be blocked in the human body it may causes heart strokes, acute chest syndrome, Pulmonary Hypertension, Gallstones etc. To avoid major complications from this disease, early care is necessary as soon as possible. So the early diagnosis is important. In this Paper author classifies the SCA in three categories such as N-0, S-1, & T-2 where N tends for normal sickle cell, S for Sickle cell and T tends for Thalassemia cells. For this data was collected from Thalassemia and SickleCell Society in which there were 1387 patients with 13 parameters like age, Haemoglobin, Hematocrit, RBC Distribution Width, MCV, MCH, MCHC, RBC, Fetal Haemoglobin, HBAo, HBA2 and Diagnosis as a response variable which was categories as N-0, S-1, & T-2 as mentioned above. After the data processing the data was divided into train data and test by 80-20 pattern. The ML algorithms like SVM, KNN, LR, DT, and Random Forest were trained on train data and accuracy was calculated on the basis of test data. From the accuracy it was observe that RF shows 96% accuracy and DT shows 95% accuracy whereas other algorithms shows less accuracy. Finally, author concludes that the RF algorithm is best for predicting the sickle cell anaemia.

Tuba Karagül Yıldız et. al. (2021) have classified anaemia by using artificial learning methods. Anaemia is the most frequent disease worldwide. An author focused on anaemia diagnosis in normal clinical practice. The data was used in this study was drawn from Düzce University Research and Application Hospital which contain 1663 sample, out of that 1109 females and 554 male patients. From the data set anaemia was classified into three main categories mycrocytic, normocytic and macrocytic anaemia. Within that classes again divided into 12 classes such as Anaemic, Non-Anaemic, IDA, IFDA, Iron and Vitamin B-12 Deficiency Anaemia, FDA, Folate and Deficiency Anaemia, Hemolytic Anaemia, Anaemia of Chronic Disease, Thalassemia, Thalassemia, Vitamin B12 Deficiency Anaemia. The independent variables used to predict the class of anaemia are gender, age, long-term disease, symptoms, and various CBC parameters. After the data collection data analysis was done. The classification was done by using MATLAB R2020 software. The various classification methods such as ANN, SVM, Naïve Bayes, and decision tree were built to classify anaemia. Ensembling methods were used for enhancing the performance of the fitted model. Also 10- fold cross validation was used while building a model. Model performance was evaluated by using confusion matrix and ROC curves, F ratio, Precision. After that

author concludes that the bagged decision tree shows highest accuracy among all the models.

Bikal Adhikari et. al. (2021) The author uses an ensemble ML model to predict cardiac disease. That was accomplished by utilizing the heart disease data set from UCI. Data was separated into two sets: train and test. Models were trained by using train data and their performances were tested on test data. Several supervised machine learning algorithms such as LR, SVM, DT, KNN, and Gaussian Naive Bayes, were built in the first stage, with accuracy of 82.46 percent, 87.34 percent, 97.67 percent, 89.94 percent, and 78.57 percent, respectively. In the next stage voting based ensemble model and average based ensemble model were build which shows accuracy 96.10% and 96.43% respectively. According to results author concluded that the ensemble model shows better performance than other machine learning models.

Mohammed F. Alrifaie et.al. (2021) have developed ML models to forecast the heart disease. Machine learning techniques have the ability to diagnose a disease based on the prior information and which will be helpful for primary diagnosis of the disease. In this research author used Random Forest and Naïve Baye's classifiers for prediction of classification purpose. The data was retrieved from the UCI Machine Learning repository and includes 14 features. There were three data sets namely Cleveland, Hungary, Switzerland. In data pre-processing the attributes those have more than 60 % missing values were removed from the data. The two methods such as filter and wrapping were used for feature selection. After the data pre-processing the Random forest and naïve Baye's classifiers were built for each of three data sets by using filter and wrapping method. After that performance was measured in the form of accuracy, precision and recall. Naïve Bayes demonstrated superior accuracy for all three datasets when compared to RF for the filter and wrapper methods. Finally author concluded that the Naïve Baye's by wrapping method was appropriate technique for prediction of heart disease.

Enav Yefet et. al. (2021) have predict anaemia at delivery by using logistic regression model. For this data of 1527 women who delivered vaginally with greater than 36 gestational weeks. Complete Blood Count (CBC) was conducted immediately after delivery and anaemia status was defined from Hb level i.e. $Hb < 10.5$ g/d considered as anaemia and $Hb > 10.5$ g/d no anaemia. Mild anaemia also considered as a anaemia. Additional data regarding health status, Use of iron supplement and vegetarianism were collected by designing questionnaire. The question related to anaemia like fatigue,

dizziness, palpitations, shortness of breath and pre-syncope etc. also included in questionnaire. Here author used Cronbach alpha to evaluate the reliability of the questionnaire. Best Hb cut off value was identified by using AUC ROC curves. According to that Hb < 10.6 g/d is best cut off value for predicting Anaemia. Statistical analysis was done in SAS software. Stepwise multiple logistic regression model was build according to that author concludes that the factors like socioeconomic status, primarity and Hb at 24–30 gestational weeks oral and intravenous iron supplement use were found to be significant factors. Author concludes that Iron deficiency was the main cause of anaemia at delivery. To avoid this kind of anaemia author suggested to focuses on iron supplements\

G Renugadevi et.al. (2021) created a hybrid ML model to foretell the occurrence of cardiac problems. The data contains questions related to ages, genders, chest pain, BP, cholesterol, fasting glucose level, electrocardiographic findings, maximum heart rate, exercise-induced angina, slop, the number of major vessels, thal, and whether heart disease was present as a pretend response or output variable. Train data and test data were formed from the original set of data. The ability of the models was evaluated using test data after they had been trained using train data. Machine learning models such as DT, RF and Hybrid were created. The DT model and the RF model were combined to form a hybrid machine learning model. Several measures were employed to assess the model's performance, including accuracy, MSE, MAE, R-squared parameter, and RMSE. The author discovered that hybrid machine learning and Random Forest gave the greatest performance for predicting heart disease from the investigation.

SulagnaDutta et.al. (2021) predicted anaemia from Anthropometric Markers and finds their best cut off. The author found that relying solely on the haemoglobin cut off was insufficient for predicting anaemia. Therefore, they used several anthropometric characteristics, including BMI, height, weight, waist circumference, waist-to-hip ratio, and waist-to-height ratio. The data was gathered by making use of a questionnaire. The study comprised 132 young adult women and 113 young adult men between the ages of 18 and 30. The anaemia was categorised according to Who criteria. Descriptive statistic and student's t test were done for continuous predictors. For all continuous variables sensitivity (SS), specificities (SP), Youden's indices, the area under the curves (AUCs) and cut-off values were determined and compared. The ROC curve analysis revealed that body mass index (BMI) is the most accurate anthropometric indicator for predicting Anaemia, with an appropriate cut-off value of 20.65 kg/m². After tallying

the anaemia prevalence across all Malaysian ethnic groups, the author discovered that 26.22% of Indians, 21.54% of Malays, 20% of others, and 20% of Chinese had the condition.

Mosieur Rahman (2021) et al. investigated the associations between WRA who engage in high-risk reproductive behaviour and the probability of anaemia, severe undernutrition, and the co-existence of these disorders. This research made use of secondary data collected from the BDHS, which ran from 8 July to 27 December 2011. 2197 WRA were ultimately chosen for the investigation after using sampling techniques. In the current study, three outcomes were examined: chronic undernutrition, anaemia and the coexistence of anaemia and undernutrition. According to WHO, undernutrition was defined by using BMI and from Hb values anaemia was categorised into mild, moderate and severe. There were number of sociodemographic factors that are theorised to be related to maternal high-risk fertility behaviours and undernutrition in women of reproductive age.

For the appropriate study, descriptive analysis and a Poisson regression technique were employed. The three regression models were developed. In order to examine collinearity between the independent variables' Spearman correlation coefficients were used. Numerous high-risk reproductive behaviours were found to be significantly correlated above three factors of interests. Significant correlations were found between certain high-risk categories, such as mothers with birth orders of three or more, maternal ages over 34 at birth, and mothers with birth orders of three or more and chronic undernutrition, anaemia, and the coexistence of anaemia and undernutrition. Additionally, a strong correlation between anaemia and birth orders of three or more and birth intervals of fewer than 24 months was found.

The author came to the conclusion that maternal high-risk behaviours among Bangladeshi women have an impact on the health of the women. The increased chance of women's chronic undernutrition, anaemia, and the coexistence of anaemia and undernutrition are all strongly correlated with high-risk fertility behaviour. The author's findings highlight the need to avoid high-risk reproductive behaviours, which mainly take the form of fewer total live births, shorter spacing between births, and too-early or too-late childbearing cycles, in order to lower the risk of chronic undernutrition, anaemia, and the co-existence of anaemia and undernutrition among women of reproductive age. For the purpose of creating treatments to enhance women's nutritional status, which is a key area of public health research, it is imperative to

investigate the causal relationship between high-risk fertility behaviour and women's nutritional outcomes.

Achamyeleh Birhanu Teshale et. al. (2020) evaluates the rate of anaemia among the WRA and identified the associated factors. To achieve the aim of the research author used secondary data from DHS of 10 eastern African countries. A weighted sample of all WRA, totalling 101524, was obtained. It was decided to create a multilevel mixed effects GLM, similar to the PR model. In order to assess clustering and variability, the intraclass correlation coefficient, proportional change in variance, and median odds ratio were also calculated. A number of models, including one with no variables, one with just variables at the individual level, one with only variables at the household level, and one with both variables at the individual and household levels, were fitted. Deviance testing was employed in order to compare the models. Anaemia in WRA was found to be significantly influenced by age of WRA, education, marital status, occupation of WRA, household wealth status, sex of the household head, toilet facility type, drinking water source, miscarried pregnancy history, parity, size of household, distance from the health facility, and pregnancy status. Age and area of residence at the community level were highly correlated with anaemia. Anaemia rates are high in Eastern Africa, according to recent observations. Anaemia was linked to both individual and community-level variables in women of reproductive age. According to the research, women who fall into certain categories like younger women, women with low education level, women from lower income households, women with outdated toilet facilities, and women without access to drinking water have a higher prevalence of anaemia and should receive special attention.

Marouane Ferjani et.al. (2020) examined performance measures to identify patterns among various supervised machine learning model types for disease detection. A few disorders that affect the heart, kidneys, breast, and brain will be the focus of the examination of machine learning (ML) models. Numerous methodologies will be tested for the disease's detection, including K-Nearest Neighbour, Naive Bayes, Decision Tree, Convolutional Neural Network, Support Vector Machine, and Logistic Regression. The top machine learning models for each disease will be identified at the conclusion of this literature review. The most popular prediction algorithms were SVM, RF, and LR. When it came to forecasting common diseases, the CNN model performed the best. Additionally, the SVM model consistently demonstrated superior accuracy for

kidney diseases. The RF performed better at predicting breast cancer. The LR algorithm ultimately turned out to be the most accurate at predicting heart diseases.

Zhang J. et. al. (2020) have establish a prediction model to estimate the possible risk of iron deficiency anaemia in babies. The relevant data were collected by well-structured questionnaire. Total 528 infants were selected from Fenglin Community Health Service Centre in Shanghai, China for this research. For each infant age (in days), gender, weight at birth (in gram), gestational age, gravida, para, mode of delivery, pattern of feeding, etc. variables were taken into account. For the anaemia diagnosis blood testing of infants also done. According to WHO guidelines the anaemia is classified into IDA-anaemia and non-IDA anaemia. After the data pre-processing various models were built including neural network, Logistic regression, Bayes Net, etc. And these models were compared by their accuracy. The multilayer perceptron model of neural network found to be better model than other models, since it has maximum accuracy than other models. Further for enhance the accuracy boosting algorithm was used. According to the prediction model developed in this study it was discovered that exclusive breastfeeding, maternal anaemia during pregnancy, and improper timing of complementary food intake were found to be the top three risk factors for IDA in infants.

Sahar J. Mohammed et. al. (2020) developed various models to predict anaemia by using rule techniques (RT). Zero R, OneR and PART algorithms were used in this research to predict anaemia. The data set that was gathered included 539 individuals and 6 distinct anaemia groups with 10 related features. According to 6 anaemia types the relationship between MCZ and RBC was detected by using scatter plot. After fitting the machine learning models, it was found that Zero R shows 40% accuracy, One R shows 81.081% accuracy and the PART shows 85.1% accuracy. The data for the three kinds of anaemia was used to compare each approach that was used. After using the PART algorithm, it demonstrates significant gains in class prediction. The results says that, to transport oxygen to all parts, there should be an appropriate amount of red blood cells. Any decrease in the number of these cells results in a decrease in the oxygen ratio and can induce anaemia. By applying data mining techniques to identify this issue early on, the patient can avoid more risk issues and sinking.

Berhanu Woldu et. al. (2020) estimated rate of prevalence of anaemia and identified factors that influence on WRA in the Sayint Adjibar town, South Wollo zone, Northeast Ethiopia. Primary data was used in this study by designing pretested questionnaire. The

single population proportion formula was used to determine the sample size, which came out to be 359. Three sections made up the questionnaire used in the data collection process. The sociodemographic information is in the first section, questions about reproductive health are in the second, and the food security condition of the home is in the third. Age, level of education, employment, menstruation history including frequency and length—use of contraception, and the number of pregnancies are among the sociodemographic questions. Women's height and weight were measured in order to compute their BMI. An equipment called a portable haemoglobinometer was used to measure the concentration of Hb. Additionally, the stool samples are examined. Data was entered into EPI Info 7 for data cleansing and verification.

When it came time to analyse the data, SPSS version 20 was used. It was used to calculate standard deviations, interquartile range (IQR), medians, means, and percentages, among other descriptive statistics. Anaemia factors were identified using multivariate and bivariate BLR models. A woman's socioeconomic status (including factors like her household's wealth index and food security), her anthropometric measurement (BMI), her religious beliefs, her occupation, the size of her household, and her level of education were all considered independent variables, while her anaemia status was considered a dependent variable.

From the study it was found that the total prevalence rate of anaemia among the WRA was 87 that is 24.2 %. Median of Hb in RAW was found to be 12.7. It was found that women who were anaemic had no formal education. The findings of the MLR showed that anaemia had a significant relationship with age, marital status, education level, food security at home, bleeding history, and the existence of intestinal parasites. Anaemia was discovered in the stool samples of the intestinal parasite-infected women. Anaemia was found to have a negative correlation with physical activity, which had a detrimental impact on health and productivity. Severe anaemia was not seen in this investigation. Low levels of schooling, BMIs greater than 25 kg/m², food insecurity, and the 36–49 age group of research participants were all linked to anaemia.

Jutatip Jammok et.al. (2020) assessed the contributions of different risk variables to anaemia and ID in this group. 399 women between the ages of 18 and 45 who were at reproductive age participated in this cross-sectional study. Those who were pregnant, displayed any indication of a chronic illness, underwent surgery, lost blood from an accident, or used iron supplements within three months of the survey were not included

in the study. The previous study's discovery of anaemia in reproductive-aged women in Northeast Thailand served as the basis for the establishment of the study's sample size. Convenience sampling was employed to find study participants. A self-administered questionnaire was required of those who satisfied the eligibility requirements.

Along with sociodemographic details like age, education, employment, and family income, the questionnaire also asked about a person's history of blood loss, which included blood donation history, surgeries, accidents, menstrual bleeding duration, and daily blood loss totals. The amount of blood lost each day was estimated by the individual. A self-administered questionnaire was required of those who satisfied the selection criteria. A complete blood count (CBC) and a serum ferritin (SF) concentration, which measures iron status, were among the laboratory tests carried out. Kolmogorov-Smirnov goodness-of-fit test verified data normality. Kruskal-Wallis one-way ANOVA for continuous variable comparisons. Comparing two independent groups was done using Mann-Whitney U test. MLR was used to identify anaemia and ID risk factors. According to statistical research, 28% of women had anaemia. Anaemia was identified in 21% of reproductive-age university students in central Thailand. Although anaemia and ID were common, WRA with and without thalassemia had different prevalence rates. Only homozygous Hb E and two-gene disorders of thalassemia types and ID caused anaemia. ID risk factors in this cohort were recent blood donation and moderate to high menstrual blood loss.

Shawni Dutta et.al. (2020) have predict diabetes by using machine learning models like Multi-layer Perceptron, NB Classifiers, DT Classifiers and KNN. From this study author detects diabetes from the patient's history. For this data was collected from UCI Machine Repository. The patient's age, total number of pregnancies, blood pressure, insulin dosage, BMI, glucose level, skin thickness, family history of diabetes, and response variable (diabetic/non-diabetic) are all part of the dataset. 80-20 pattern was used to selecting training and testing datasets. Model building procedures was done in two phases in the first phase ANN and KNN models were trained using adjusted parameters which improves the classifier's performance. After the first phase performance evaluation was done and model performance was measured by using MSE, accuracy, F1-Score, and Cohen-kappa score. Best model DT Classifier was selected. In the second phase the ensemble learning was done in two ways such as voting ensemble and stacking ensemble for enhancing the prediction performance.

Again, performance of both methods was measured by using MSE, accuracy, F1-Score, and Cohen-kappa score. It was found that stacking ensemble method shows best results than the voting ensemble method. At the last author conclude that stacking ensemble classifier gives best diabetes prediction.

Raj H. Chauhan et.al. (2020) Proposed machine learning models like Gaussian NB, DT and RF to detect various diseases without visiting a doctor or physician for the diagnosis. For this purpose, the medical data was collected from New York Presbyterian Hospital. Data contains disease, the number of discharge summaries like current and mention of disease and respective symptoms. Based on these notes, associations for the 150 most common diseases were calculated, and the symptoms were graded according to the strength of the relation. For the coding purpose MedLEEnatural language processing system was used. For finding associations the statistical methods based on frequencies and co-occurrences were used. After data pre-processing the machine learning models likes NB, DT and RF were build. It was observed that the RF Model shows highest accuracy (95.28%) among the all models. Author concludes that the RF model was best model for this study. The authors of this study predicted the disease based on the patient's discharge summaries without undertaking any medical diagnosis.

Joon-myung Kwon et al. (2020) conducted a retrospective, multicentre investigation to detect anaemia using the ECG. They employed a deep learning approach for this task, which was subsequently validated both internally and outside. Hospitals A and B were the location of the data collection. For the purposes of model construction and internal validation, data from Hospital A was utilized, while data from Hospital B was utilized for external validation. author removed participants who did not have complete demographic, electrocardiographic, or haemoglobin data and included those who had at least one ECG with a haemoglobin measurement within one hour of the index ECG. In order to identify anaemia, three distinct Deep Learning Models were created for ECGs: 12-lead, 6-lead, and single-lead. The following levels of haemoglobin are used to determine the existence of anaemia: normal to mild anaemia, moderate anaemia and severe anaemia. To create the DLA, a CNN was fed 500 Hz of raw electrocardiogram data, along with demographic information such as age and sex. Internal and external validation data were used to create the complex neural network models. Results showed that the DLA from a 12-lead ECG had an area under the receiver operating characteristics curve of 0.923 for internal validation and 0.901 for external validation

when it came to diagnosing anaemia. According to the author, DLA accurately detected anaemia using raw ECG data. Author concluded that Artificial intelligence applied to ECGs will allow for the detection of anaemia.

F. M. Javed Mehedi Shamrat et.al. (2020) Have developed various ML algorithms for detection of Breast Cancer. The detection process runs in four phase. The first phase focused on extracting and integrating data from a variety of health-care systems using devices. Phase 2 focused on the storage of a large volume of medical data. In the phase 3 machine learning models were build.

The outcome of the breast cancer detection technique for the clients was exemplified in phase 4. In this research author focused only on phase 3. For this data was collected from the 'University of Wisconsin Hospitals, Madison, Wisconsin, USA' which contains 699 breast cancer patients. Total 11 parameters were selected for this study. After the data pre-processing 6 ML models such as NB classifier, RF, SVM, DT, KNN, and LR were developed. Model performance was measured from confusion matrix by estimating accuracy, sensitivity, precision, specificity and f1 measure. And ROC curves also drawn for performance measure. At the last author concluded that the all six machine learning models shows best performance for predicting Breast cancer. According to accuracy SVM shows highest performance as well as Naïve Bayes classifier and decision tree shows considerable accuracy. It was found that SVM, Naïve Bayes and Decision Tree models can be helpful while early detection of breast cancer.

Apurb Rajdhan et.al. (2020) Have proposed various machine learning models for prediction of heart disease. Identifying heart illness in its early stages was the primary objective of this research. Preventing harmful side effects while providing patients with the right treatment will become much easier as a result. For that heart disease data from UCI Cleveland dataset was used. There were 76 attributes in the original data. The 14 attributes in the processed data, which included the patient's age, gender, level of chest pain, BP, cholesterol, blood sugar during fasting, ECG results, maximum heart rate, and whether or not exercise-induced angina, were used by the author. Using the 80-20 pattern, the final data was partitioned into 2 sets: the train set and the test set. ML models, such as RF, DT, Naive Bayesian classifier, and BLR, were trained on train data and their performance was evaluated using a confusion matrix on test data. The author found that the RF method had the highest accuracy (90.16 percent) when compared to other algorithms. The authors come to the conclusion that Random Forest is a better predictor of heart disease.

Priyanka Anand et. al. (2020) have predict the anaemia in children by using comparative approach of several Machine learning algorithms such as LDA, CART, K-NN, RF and BLR. For this study primary data were collected by designing the questionnaire. Total 2010 children of age 6-36 months were included in sample. Anaemia was classified as anaemic and non- anaemic form the Haemoglobin of the children. The children with Hb less than 11 g/dl were classified as anaemic otherwise non-anaemic. The following factors were taken into account when classifying anaemia: the child's age, place of living, tehsils, sex of the child, mother's age, family monthly income, maternal education, working status, parental education, weight of the child at birth, growth status of the child, exclusively mother's feed, any medication for parasites, mode of delivery, and child morbidity (fever). 80% of the total data were utilized for training and twenty percent were used for testing. R program was used for analysis. Using this data, four machine learning models were constructed, and the accuracy, specificity, and sensitivity of each model were used to gauge its performance. After comparing the models' performances, the author finds that the RF model outperforms other ML models in terms of accuracy (67.18%). It indicates that out of the three models, Random Forest predicts anaemia the best.

Sujan Gautam et.al. (2019) et.al. identified the variables for the prevalence of anaemia in child-bearing age women. NDHS 2016 data was used for this research. The anaemia status taken as a dependent variable which was classified into four according to WHO guidelines. The socio-demographic factors, reproductive factors were used to check the association with anaemia. The nutritional and behavioural factors such as addiction, BMI, domestic violence etc were included in this study. Descriptive statistic, frequency tables, proportions were used to do exploratory analysis. Chi-square test was used to check the association between independent factors and anaemia status. BLR analysis was used to establish each component's distinct impact on the anaemia status. A MLR analysis model was then used to determine the adjusted influence of each factor on the dependent variable. Furthermore, an alternative multivariate regression analysis model was developed to ascertain the ways in which women's choices and experiences with IPV influenced their anaemia status.

Women between the ages of 15 and 49 had a 41% total prevalence of any anaemia. In particular, 33%, 7%, and 0.3% of the women had mild, moderate, or severe anaemia, respectively. The author came to the conclusion WRA that having a well as a source of drinking water was strongly associated with a higher incidence of anaemia.

The use of contraceptive was found to be associated with anaemia. Factors including ethnic background, financial status, reproductive status, having given birth within the last three years, and breastfeeding did not reveal any significant link with the chance of developing anaemia, despite their high relationship in the univariate analysis. In this study, two out of every five women were found to be anaemic.

Those who did not smoke had a higher prevalence of anaemia than those who did, and this connection was statistically significant. According to this study, anaemia is very common in underweight women. This study examined the role of women's decision-making autonomy over their healthcare and experience of IPV with anaemia in addition to examining the prevalence of anaemia among women of reproductive age. After adjusting for a number of factors, the relationship between women's decision-making and IPV was somewhat diminished. Married women who made their own healthcare decisions independently were shown to have a decreased prevalence of anaemia. This can be explained by the women's relative status as decision-makers in the household.

Manish Jaiswal et.al. (2019) have developed various machine learning models to predict anaemia. The main purpose of this study is to investigate anaemia at early stage and this objective was fulfilled by using machine learning algorithms such as Naïve Baye's, Random Forest and Decision tree. The data originated from neighbouring laboratory test centers and pathology centers. 200 CBC test samples with 18 characteristics make up the data set; however, only the features crucial for identifying anaemia were selected for this study. The factors involved include Gender, Age, MCV, HCT, HGB, MCHC, and RDW. Following data preparation, the data is subjected to Random Forest, Decision Tree, and Naïve Baye classifiers. The accuracy and MAE of three models were found in order to evaluate the performance of the models. Subsequently, the accuracy of RF, Naïve Baye's, and DT with C4.5 was determined to be 95.3241, 96.0909, and 95.4602, respectively. The Mean Absolute Error (MAE) for the two models was found to be 0.0332, 0.0333, and 0.0347, respectively. While it was noted that all three classifiers had higher accuracy, the Naive Bayes model performed better when viewed from a comparison perspective. In the future, automated technologies might be created to improve prediction outcomes and recommend further diagnoses. The early detection of more serious diseases may benefit from the application of such automated procedures. Such a disease prediction system can also be used to recommend a course of treatment.

Nahiyan Bin Noor et.al. (2019) Predicted the Haemoglobin level by using different regression techniques. The aim behind this research was to predict the Hb by the painless , quik and cost effective tool. Data was obtained from the hospitals of Chittagong, and Cox's Bazar Medical College, Chittagong Medical College. The data set contains 104 individuals; 81 are considered as train data and 23 are test data. A smartphone camera with 12 megapixels was used to take pictures of the participant's conjunctiva. The haemoglobin level information from CBC was also acquired. The image processing, which included determining the percentage of green, red, and blue pixels, was done using the MATLAB software. The response variable, the Hb value, and the percentages of the red, green, and blue pixels make up the three independent variables that make up the statistics in the end. After the data was pre-processed, the Hb predictions were made using Decision Tree (Medium) Tree, Linear SVR, and Multivariate Linear Regression (MLR). The percentage of errors between the actual and predicted values was computed after all three models had been trained. These error percentages were used to compare models. In comparison to the other two, the decision tree model was determined to have a lower percentage of errors. Ultimately, the author came to the conclusion that Hb as measured by this method will be rapid, easy, and affordable.

A.K.M Sazzadur Rahman et. al. (2019) have developed a machine learning models for the prediction of Liver disease. The author's major goal for this research was to use machine learning techniques to predict liver disease and to reduce the cost of diagnostics. For this research author was collecting the data from UCI machine learning Repository. There were 583 liver patients out of which 75.64% were male and 24.36% were female patients. Data consist of 11 variables out of that 10 variables were considered as features and one variable as a target variable or response variable (i.e. liver patient of non-liver patient). Data analysis was done in Jupiter notebook. In the data pre-processing the variables that shows multicollinearity was omitted from the study. After the data pre-processing machine learning models such as LR, , DT, RF, , NB, SVM and KNN were built. The performance evaluation of these six models was done by estimating accuracy, sensitivity, specificity, precision and ROC curves. From these parameters author concludes that the Logistic regression model shows greater accuracy than other 5 models. This study will be helpful for predicting liver disease with a low cost.

Razat Agarwal et. al. (2019) did classification and prediction of fruit images by using five supervised ML classifiers such as SVM, RF, KNN, Naive Bayes, and SoftMax. The resource of the dataset is Kaggle. The data was divided into smaller and larger sets after 65,429 RGB photos of 95 different fruits were included. There were 18 fruits in the smaller set, with 2,961 photos in the test set and 8,846 images in the training set. Additionally, there were 95 fruits in the huge data set, with 16,421 photos in the test set and 48,905 images in the training set. Data loading and pre-processing were completed. The pre-processing step involved loading the image and converting it to grayscale in order to reduce its size. The characteristics were standardized using standard scalar, which removed the mean and scaled to unit variance. The data's dimension was reduced by using the principal component. Using training sets of both small and large data sets, five data mining models were trained. The accuracy of the model was evaluated. Following model evaluation, it was shown that SVM work best with both small and large data sets.

Dithy M.D et.al. (2019) used the Random prediction (Rp) classification model to forecast the iron deficient anaemia in pregnant women. The data used in this study was 2120 samples with 19 features. The feature/variable selection was done by using Improved Median Vector Feature Selection (IMVFS). After the data pre-processing and feature selection the prediction was done by using Random prediction (Rp) algorithm. In this paper author also reviewed various research work related to this topic according to that author says that iron deficiency anaemia was most frequently caused. Author also concluded that the Random prediction algorithm with Improved Median Vector Feature Selection (IMVFS) shows best results instead of previously used algorithms like Artificial-Neural Network (ANN) ,Gausnominal and VectNeighbour classification algorithms.

Hamid Mukhtar et.al. (2019) have created machine learning models for malaria and anaemia prediction. This study also focuses on social and economic factors that contribute to these disorders. The Demographic Health Survey (DHS), conducted in Senegal between 2015 and 2016, provided the data for the present study. The survey was conducted from February to August which is dry season and from September to January which is rainy season. The original dataset has 6935 instances and 986 variables. A household questionnaire and two individual questions for WRA and men (15 to 49 years old) were the three types of questionnaires used in the survey (15 to 59 years). The questionnaire contained questions about anthropometric measurements,

sociodemographic traits, pregnancy and postpartum care, women's occupation and status, immunization history, and nutrition. In order to evaluate haemoglobin levels and determine the prevalence of malaria, blood samples from children under the age of five were also obtained. After the data collection the data pre-processing was done in which coding and data cleaning was done. According to the WHO criteria Anaemia was categorized as severe, moderate, mild and non-anaemic in this study the severe, moderate and mild considered as positive class (i.e. 1) and non-anaemic considered as negative class (i.e. 0). Similarly, Malaria was categorized as positive and negative class. Before going to model building the feature engineering was done by using correlation, Gradient Boosting and Recursive Feature Elimination methods. After that Five ML models such as ANN, SVM, KNN, RF and NB were developed for the prediction of Anaemia and malaria. After the Model evaluation it was observed that ANN shows 94.74% accuracy for malaria and 84.17 for anaemia. At the last author concluded that the non-medical factors also affect the malaria and anaemia so for reducing these diseases we have to focused on these social factors also.

Young et al.el. (2018) reviewed a number of studies on maternal anaemia and risk. Some key findings of his review are: About 500 million WRA are affected by maternal anaemia. Additionally, maternal anaemia increases the mother's risk of death during and after childbirth. Maternal anaemia and mortality were shown in observational studies to be linearly correlated, with each 10 g/L increase in maternal haemoglobin being associated with a 29% decrease in maternal mortality. The scientists came to the conclusion that in that community, severe anaemia during pregnancy or postpartum increased the chance of maternal death. The findings highlight the need of including maternal anaemia in a comprehensive maternal health plan to lower maternal mortality.

Phuong Hong Nguyen et.al. (2018) have been identified the trends and factors that influence the prevalence of anaemia. The National Family Health Surveys (NFHS-3 and NFHS-4) provided the secondary data. Out of the two sets of data, 245 kids from NFHS-3 and 346 kids from NFHS-4 were selected. 37, 165 expectant mothers and 760,460 non-pregnant mothers, in that order. The factors that alter the trends of anaemia were chosen using the UNICEF and Nutrition series. Immediate determinants, nutrition and health treatments, and underlying determinants were the three groups of criteria used. Factors pertaining to diets, illness load, and maternal undernutrition are included in the intermediate determinant. The household level parameters linked to no. of

children within 5 years of age, socioeconomic status of household, sanitation, religion, place of residence (rural/urban) and scheduled caste/tribe.

The data analysis was carried out independently for women who were not pregnant, pregnant women, and children. The purpose of the multiple linear regression analysis was to predict the Hb and determine the components' respective contributions. MLR analysis was performed in order to confirm the Hb prediction and ascertain the influence of various variables. The findings indicate that between 2006 and 2016, pregnant women and children saw significant advances in lowering anaemia and haemoglobin levels, while NPW did not. The author found trends in the factors affecting variations in haemoglobin concentration and the frequency of anaemia in children and expectant mothers. While improvements in nutrition and health treatments had the biggest impact on the reduction of anaemia in children, increases in maternal education and socioeconomic status were the primary drivers of the anaemia reduction in pregnant women. Gains in haemoglobin and anaemia between 2006 and 2016 were also mostly linked to greater maternal BMI, increased consumption of food derived from animals by mothers, improved hygiene, and fewer small children per household.

Deepika Bansal et.al. (2018) have developed several ML algorithms for the prediction of dementia. To detect dementia four supervised ML algorithms such as J48, Naïve Baye's, Random Forest and Multilayer Perceptron were used. For this research the secondary data was collected from OASIS-Brains.org. There were two types of data available cross-sectional MRI data of size 416 and longitudinal data MRI data of size 373. Missing values are replaced by corresponding their averages. After fitting the models on both the data sets it was found that J48 gives best accuracy 99.52% than other models on both the data sets. Feature selection was done using CFSSubsetEval. The author comes to the conclusion that dementia is a significant worldwide health issue and that there should be more emphasis on risk reduction, early intervention, and timely detection of the disease in older persons than on finding a cure.

Molly A. Moor et. al. (2017) determined personal and community factors contributing to anaemia among women in rural Baja California Mexico. This research set out to answer the question, "What causes the unusually high frequency of anaemia in women in this region?" by looking at both individual and community factors. The data was collected by completing the survey of 118 women (9 PW and 109 NPW) at child-bearing age from rural Colonia in Baja California, Mexico. Descriptive statistics, Chi-square tests, and multivariate LR have been used for statistical analysis. A multivariate

logistic regression analysis includes the variables that showed a significant correlation with anaemia status in the bivariate study. After analysis it was discovered that the 22 % women were anaemic.

The main cause of anaemia is a lack in vitamin B-12. Women who participated in the government aid programme Prospera and those from low socioeconomic level households had a much higher likelihood of becoming anaemic. Author concluded that the nutritional deficiency was the main cause of anaemia, vitamin supplementation is the temporary solution. The greater prevalence of anaemia in the community might be decreased by promoting government programmes like Prospera, establishing additional programmes to enhance the nutrition and health literacy, and providing access to wholesome meals.

Rushali R. et.al. (2017) aimed to determine how common anaemia is and what variables contribute to it in reproductive-aged women. From a Mumbai slum, 315 WRA included in this study. A population contains migrated peoples from different parts of India. Stratified sampling is used to select samples from various sectors. A random sampling method was used to choose one sector from a total of eleven. Houses were chosen from that sector using stratified random sampling. The study comprised women who resided in the study region for the preceding 6 months and expressed their willingness to participate. The study excluded women who resided in the study area for a duration of less than 6 months, as well as pregnant women who were below the age of 18. Primary data by using questionnaire were collected. The survey encompassed socio-demographic inquiries such as age, education level, family structure, socio-economic standing, dietary habits, and obstetric details. The urban health centre conducted blood haemoglobin measurement using Sahli's method. The anaemia was further classified into three categories: mild, moderate, and severe, based on the standards provided by the WHO. Mild anaemia was characterised as having a Hb level between 10-12 gm %, moderate anaemia was characterised as having a Hb level between 7-10 gm %, and severe anaemia as having a Hb level below 7 gm %. The statistical analysis was conducted utilizing the SPSS software. Quantitative variables were subjected to descriptive statistics, while the connection between two qualitative factors was assessed using the chi-square test.

49.5% WRA found to be anaemic. This study's findings suggest that anaemia is more prevalent in the younger age group and among women who have completed only primary school. There was association between Anaemia and lower socio-economic

status. In this study author found that a woman who takes mixed diet was less anaemic than vegetarian women. There was evidence of association between anaemia and iron rich food. It was also found that anaemia is high in women who had less than one year difference in their pregnancies. It was discovered that the rate of prevalence of anaemia was higher in women with parity 3 or more than in women with only parity 1.

Kassandra L. Harding et.al. (2017) have studied the rate of anaemia in the women and children in Pakistan and Nepal and also find the associated factors. The secondary data was used from the NDHS 2011 for Nepal and the Pakistan NNS 2011 for Pakistan. 5794 WRA were included in data from Nepal and 8324 from Pakistan respectively. The children whose age less than 5 were included in analysis. 2088 children from Nepal and 8968 from Pakistan were included in the sample. Anaemia was defined from Hb concentration of less than 120 g/L for WRA and less than 110 g/L for Children whose age was less than five years. In this study the factors assessed in Nepal and Pakistan differed slightly. In the statistical analysis author first used log binomial model but it was not good fitted so after that author fitted modified Poisson regression model. It was found that the modified Poisson regression model was appropriate than Logistic regression model. From the output the factors associated with the anaemia among WRA and children <5 years were identified and it was observed that the factors like month of interview, lack of facility at home, education level, child's age in months and status of stunted are significantly affects the anaemia among the children< 5 years in Nepal. And the factors associated with Anaemia among the WRA in Nepal were month of interview, ecological zone, age, BMI corresponding to the age less than 18.5 and greater than 25 years, status of breastfeeding. Similarly, the factors associated with anaemia in the WRA and Child< 5 years in Pakistan are identified. For children< 5 years the factors like province, household status like poorer, poorest, mother's BMI, status of worms, age of child in months, status of stunned were shows significant association with anaemia. For WRA the factors like month of interview specially March-May, June-September, province, BMI corresponding to Less than equal to 18.5 years and greater than equal to 25 years, more births, status of supplement of iron and vitamins were shows significant association with anaemia. From the findings of the research it was observed that the 46% children and more than one third of the sample of WRA in Nepal were affected by anaemia. In Pakistan 63 % children found to be anaemic and 51% WRA found to be affected by anaemia.

Olani Debelo et.al (2016) identified the factors associated with anaemia. For this data was obtained from EDHS. 15,567 WRA were selected for this study. The prevalence statistic of anaemia was measured. When looking for an association between anaemia status and other factors, the chi-square test was employed. LR was used for identifying the factors which affects the anaemia. For the analysis purpose anaemia is categorised in mild, moderate, and severe such as if Hb level is 11.0 to 11.9 g/dl then it is categorised as mild anaemia, if Hb level is 8.0 to 10.9 g/dl then it categorises as moderate anaemia and if Hb level is below 7 g/dl then it is categorised as severe anaemia but in expecting women categorisation was different. In expecting women if Hb level is 10 to 10.9g/dl then it was categorised as mild anaemia, if Hb level is 7 to 9.9g/dl then it is categorised as moderate anaemia and if Hb level is 4.0 to 6.9g/dl then it is categorised as severe anaemia. From the results of Chi-square test variables which shows significant association with anaemia were selected in the logistic regression model. Anaemia was prevalent in women at 19.9%. Various regions had varying rates of anaemia. Anaemia is very common among women living in rural areas, according to this study. The author concluded by saying that anaemia and wealth index were significantly related. Women have low risk of anaemia if contraceptive method is used. The women who had antenatal visits during pregnancy have low risk of anaemia. The women who take iron tablet/syrup had low risk of anaemia. By LR it is clear that anaemia is depends on total no. of children born, BMI, residential region, education, wealth index, pregnancy, months of breast feeding, place of delivery, number of antenatal visit, taking iron tablets, postnatal check-up, contraceptive use and drug use for intestinal parasite. Also found that there was significant association between maternity health care services and anaemia.

Chi Huu Hong Le et. al. (2016) et al. examined the incidence of anaemia in the United States populace. Between 2003 and 2012, data were extracted from the NHANES cycle. 776 women whose pregnancy status was known to be positive were excluded from the overall analysis and analysed individually. 6.4% were other, 11.80% were non-Hispanic black, 14.70% were Hispanic, and 67.10% were non-Hispanic white within the analysed sample. Data from the NHANES were anonymized prior to their analysis and access in this study. The categorization of the anaemia was based on WHO criteria. The following table details the WHO classifications for anaemia.

Table 2.1 Anaemia Categories

Population Group	Non-Aneamia	Anaemia	Moderate-Severe Anaemia
Children (age=0.5-4.9 Years)	Hb>11 g/dL	Hb<11 g/dL	Hb<10 g/dL
Children (age=05.0-11.9 Years)	Hb>11.5 g/dL	Hb<11.5 g/dL	Hb<11 g/dL
Children (age=12.0-14.9 Years)	Hb>12 g/dL	Hb<12 g/dL	Hb<11g/dL
Non-Pregnant Women (age>15 years)	Hb>12g/dL	Hb<12g/dL	Hb<11g/dL
Pregnant Women	Hb>11 g/dL	Hb<11 g/dL	Hb<10g/dL
Men(age>15 years)	Hb>13g/dL	Hb<13g/dL	Hb<11g/dL

Statistical analysis was done by using SAS Software. Age categories, sexes, racial/ethnic groupings, and the cohort survey year were used to examine the anaemia and moderate-severe anaemia. Using the chi-square test, differences in demographic statistics were calculated. Women who were pregnant are independently were examined. The sample weight was used to determine the percentages of participants who had moderate to severe-anaemia and anaemia overall.

In this research the incidence of anaemia was determined according to age, sex, and race/ethnicity. Level of Serum haemoglobin (Hb) distributions by sex were also graphed. The author of the most recent study determined that the aggregate percentage of anaemia in the United States population was 5.60%, with a 95% confidence interval 5.1% to 6.1%. Furthermore, the percentage of moderate-severe anaemia was 1.5%, with a 95% confidence interval of 1.4% to 1.7%.

In general author discovered that, non-pregnant females had a much greater prevalence of anaemia than men. Without considering those who were pregnant, there were twice as many anaemic females as anaemic males. Compared to men, NPW had

five times the prevalence of moderate to severe anaemia. The age range of 80–85 years is an exception to this tendency, as the proportion of males with anaemia was twice as high as that of females. But for this age range, the rates of moderate-severe anaemia were approximately equal for both sexes.

It was found that NPW had a much greater average prevalence of anaemia than men. The prevalence of anaemic women was double that of anaemic males when pregnancy status was excluded. In terms of severity, women who were not pregnant were 5 times more likely than males to have moderate-severe anaemia. The age range of 80 to 85 years is the exception to this tendency, where two times as many males as females developed anaemia. However, rates of moderate-severe anaemia were approximately equal for both sexes in this age group.

Author discovered that the children in school age i.e. 5–11 years had the lowest frequency of anaemia out of all age groups. Additionally, the prevalence rate of moderate to severe anaemia was lowest among preschool-aged kids (0.5-4 years) and kids in school (5-11 years). The proportion of people with moderate-severe anaemia and anaemia was highest in the 80- to 85-year-old age group. The incidence of anaemia and moderate-severe anaemia both rose bimodally, reaching peaks at 40–49 and 80–85 years old, respectively.

The research states that severity of anaemia varied with races. Black people had the highest rate of anaemia for both sexes across all age categories. For males, the prevalence of anaemia was also shown to follow a consistent tendency across many racial/ethnic groupings. The age range 15 to 29 saw the biggest increases, and Hispanics experienced rises six times greater than whites. The prevalence of moderate-severe anaemia and overall anaemia increased from 4.0% to 7.1% and 1.0% to 1.9%, respectively, from 2003–2004 to 2011–2012. Males experienced the greatest increase in prevalence during this 10-year period.

Among 776 pregnant women, 8.8% had anaemia and 3.5% had moderate to severe anaemia. In contrast, black expectant mothers exhibited the greatest prevalence of moderate to severe anaemia. Anaemia was prevalent among individuals of other ethnicities, Hispanics, and Whites at rates of 3.1%, 9.2%, and 15.6%, respectively. The overall cohort exhibited a distribution of serum haemoglobin (Hb) levels with mean and median values of 14.2 and 14.1 g/dL, respectively. Feminine Hb values were, on average, lower than male Hb values.

The findings presented an up to date assessment of anaemia across the entire United States population, as well as in subgroups categorized by age, race/ethnicity, duration, severity, and gender. During the study's time period of 2003 to 2012, there was an observed increase in the prevalence of anaemia. Non-Hispanic Blacks, Hispanics, elderly women in their reproductive years, and expectant women are more susceptible to developing moderate to severe anaemia.

Moloud Abdar et. al. (2015) compares various data mining algorithms for prediction of heart diseases. The most recent statistics from the WHO indicate that cardiac disorders garner considerable attention in medical research due to the significant implications they have on human health. In this study researcher predict heart diseases by five data mining algorithms including C5.0, Neural Network, SVM, KNN and Logistic Regression. For this study data was took from the University of California, Irvine (UCI).Data consist of 270 sample with 13 features classified into 'with' and 'without heart diseases. Using Logistic regression the variables those significantly correlated with target variable were selected as predictors. From this the variables Sex, CP, RBP, EIA, NUM, Thal were used for prediction purpose. For model building data is divided into train and test. (70% and 30%). After Model building model fitness and model accuracy was calculated by using confusion matrix. Based on Training data model fitness was calculated and model accuracy was calculated from test data. From the model accuracy results, C5.0 decision tree shows greatest accuracy (93.02%) rather than KNN, SVM, Neural network. Also from ROC curves C5.0 decision tree classifier shows better performance. Therefore, from overall analysis C5.0 decision tree shows best performance for predicting heart diseases.

Ralph Green et. al. (2015) make a review on the microcytic anaemia. From the review author wrote some statements about Macrocytic anaemia. Adults with a mean cell volume (MCV) of 100 fL are said to have macrocytic anaemia. Microcytic (MCVo80 fL), normocytic (MCV1480-100 fL), and macrocytic (MCVZ100 fL) anaemias are categorised separately. Since the normal range for MCV is often lower in children, with the exception of the neonatal period and the first six months of life, when it is higher, it is always crucial to evaluate the MCV in respect to age-appropriate reference ranges. There is a physiological rise in MCV of roughly 4 fL during pregnancy. Although macrocytic anaemias can have a variety of causes, they are typically split into non-megaloblastic and megaloblastic anaemias. It was crucial to consider and confirm or rule out B12 and folate deficiencies when treating macrocytic anaemia since, if

present, these disorders can be treated, but if neglected, they can result in serious and occasionally permanent consequences, especially in the case of B12 deficiency.

Measuring the plasma or serum levels of vitamin B12 and folate as well as red cell folate is necessary for the evaluation of megaloblastic anaemia. There was list which contains additional causes of megaloblastic anaemia in addition to B12 and folate deficiency. These include a wide range of medications that either obstruct DNA synthesis or the assimilation, metabolism, or processing of one or both vitamins. Other antifolates, such as methotrexate, can cause functional folate deficit and cause megaloblastic alterations.

The most common cause of macrocytic anaemia is alcohol consumption. Alcohol, with or without liver disease, was in second place behind drugs and pharmaceuticals as a cause of macrocytosis in a study involving hospitalised patients in New York City published in 2000. In addition to contributing to megaloblastic macrocytic anaemia, alcohol also causes non-megaloblastic macrocytic anaemia. In this case, stopping drinking will frequently cause the macrocytosis to go away. Alcohol-related liver disease and non-megaloblastic macrocytic anaemia are frequently linked in alcoholics.

There are numerous primary haematological conditions that can exhibit macrocytic anaemia. While myelodysplastic syndrome (MDS) is a more frequent cause of macrocytic anaemia in the clinical practise of haematology and hematopathology, some individuals with aplastic anaemia have macrocytosis. An initial differential diagnosis was made when macrocytic anaemia is discovered. Analysing the peripheral blood smear should be the first step in the evaluation in order to determine whether the anaemia is most likely megaloblastic or non-megaloblastic. It is crucial to assess the patient's medical history, including past medication usage. A patient with macrocytic anaemia can be evaluated using a number of different approaches.

Bilkish N. Patavegar et.al.(2014) focuses on the prevalence of anaemia and factors affecting on anaemia. For this study researcher use primary data. Data was collected in rural areas of Maharashtra for 6 months. Sample size was calculated by using prevalence of anaemia 46%, allowable error 10% at 95% level of significance and it was 416. study was conducted in O.P.D. of primary health centre. The women attending O.P.D. and of age 15-49 years were taken in sample. By using systematic random sampling every 5th women of reproductive age was chosen for study. Data was collected by using a well-designed questionnaire which contains socio-demographic characteristics, clinical examination, and dietary factors significant for anaemia. Blood

haemoglobin measurement was done by Sahli's method in urban health centre. Anaemia Categorization like mild, moderate and severe was done from WHO guidelines. From the 416 reproductive women 216(51.92%) were found to be an anaemic and Majority of them having a mild anaemia (63.3%)and 2.31% had a severe anaemia. There were total 17 socio demographic factors out of that 10 factors were associated with anaemia. That factors are Type of House, Education, Socio economic Class, Hand washing before meal, Awareness regarding Anaemia, Intake of Iron rich food, Intake of sprouted Food, Amount of blood loss in menses, Duration of menstrual bleeding (days), H/O worms in stools. According to this study it was found that most of the women in rural area were suffering from mild anaemia (i.e.63.37%), 34.26% were suffering from moderate anaemia, and only 2.31% women were suffering from severe anaemia. Results show that, a large number of women suffering from anaemia who lived in at Katch house and it was 59.41%. The women lived in joint family have more chance to be anaemic than nuclear family. Most of the women who were illiterates were suffering from anaemia. Also the women belongs to lower socio-economic class were found to be anaemic. The women who didn't have any knowledge or information about anaemia were found as anaemic. Women with a menstrual cycle more than five days were found to be anaemic. Females with a history of worm passage in stools had a greater rate of anaemia than non-anaemic women (77.78 %). From this study author says that for reducing anaemia we have to focus on female literacy, living conditions and personal hygienic conditions, awareness and screening for anaemia etc.

Ramesh Verma et.al.(2014) find out the prevalence of anaemia among women of reproductive age group in rural block of Haryana. And also find the effect of anaemia on mean height and weight of women of reproductive age. This cross sectional study was done in Sampla block which is in south-eastern part of the district Rohtak. In this study total 18402 women of age 15-49 years were registered. Those women who did not report for examination after calling/contacting three times excluded from the study. The cut of value of haemoglobin for detecting anaemia is different therefore pregnant women also excluded from this study. Total 8590 non-pregnant women of reproductive age were analysed in this research. Information about study participants were collected by using predesigned and pre-tested questionnaire and clinical examination also done for Haemoglobin measurement. Statistical analysis was done by using SPSS software. The average weight of severely anaemic women found to be 48.525 ± 9.361 , for moderate anaemic women was 51.349 ± 11.090 and for mild anaemic women it is

52.651±11.001. The average weight of non -anaemic women was 53.147±11.315. From the one factor ANOVA it was found that mean weight in non-anaemic women of reproductive age group was more than that of varying degree (severe, moderate, mild) of anaemic females. The average height for severely anaemic women found as 153.21±6.857, for moderate anaemic women it was 154.5±6.577 and for mild anaemic women 154.6±6.194. The average height of non-anaemic women was 154.8±6.166. This study also focusses high prevalence of anaemia in child bearing age. Finally author recommended that there is need to include iron rich food in the diet of women.

Falgunikumar Laha (2014) make a review of number of articles on prevalence of anaemia in India. Anaemia found to be a global burden. Proteins, amino acids, vitamins A and C, and other vitamins of the B-complex group, such as niacin and pantothenic acid, are also involved in the maintenance of haemoglobin level, which is why the majority of anaemias are caused by inadequate supply of these nutrients. Anaemia is estimated to affect 50% of people in India. One in every two Indian women (56%) has suffered from anaemia, and it is estimated that 20% to 40% of maternal mortality in India are related to anaemia. According to the District Level Household Survey (DLHS) surveys, anaemia is highly common among young children, pregnant and breastfeeding women, and adolescent girls. Young children, women who are pregnant, and newborns with low birth weights are at particular risk for anaemia.

According to the NFHS-III study conducted in 2005–2006, India has one of the highest rates of anaemia in the world. The causes include the high expense of healthcare services, the low status of women, and poor food quality. In India, anaemia continues to be a key factor in both maternal mortality and low birth weight. Even among educated households with greater incomes have seen moderate to severe anaemia was common. According to the ICMR district nutrition research within the duration 1999–2000, anaemia prevalence in pregnant women was 84.2 percent, with 13.1 percent having severe anaemia. These results therefore imply that women may continue to experience anaemia throughout their lives.

There are several consequences of anaemia such as lower physical capacity, delayed cognitive development in young children, higher risk of infection, and compromised foetal development during pregnancy. After doing several reviews author came to the conclusion that ‘the key to reducing anaemia still lies in anaemia screening, treating

anaemic women, and making iron-fortified foods (such as wheat flour with folic acid and iron, milk sugar, and salt) readily available’.

JM AlQuaiz et. al. (2012) detected iron deficiency anaemia among healthy women of child bearing age by using parameters obtained from complete blood count. For this data was collected from King Khalid University Hospital, Riyadh, Soudi Arabia. The females of age 15 to 49 were selected for this study. Laboratory investigations were done such as complete blood count, serum ferritin and haemoglobin electrophoresis. According to criteria of WHO Anaemia was defined. If MCV (mean Cell Volume) less than 80fL then there was Microcytic anaemia. If ferritin level ≤ 15 ng/ml then there was Iron deficiency anaemia. Data analysis was done in SPSS software. MedcalcTM software was used for finding the accuracy of the iron parameters by plotting receiver operating characteristic (ROC) curves. From the ROC it was observed that the parameters MCV, MCH and RWD are significant parameters for detecting iron deficiency anaemia (IDA) in women at child bearing age. Author finally concludes that the CBC indices were good alternative predictor for iron deficiency anaemia in women at child bearing age.

Prabhaker Mishra et.al. (2012) developed a cross sectional study to determine the prevalence of anaemia among the women at reproductive age (WRA). For this study 598 reproductive aged women (15-45 years) were included as a sample from Barara village of the Ambala district. The two-stage cluster sampling technique was used here in the first stage the one rural PHC namely Barana selected out of three PHC using simple sampling. In the second stage one village from the entire villages were selected. Anaemia was defined by using haemoglobin concentration as Hb<11 gm/dl in pregnant women and Hb<gm/dl in non-pregnant women According to International Nutritional Anaemia Consultative Group (INACG). In the statistical analysis one-way ANOVA was used to determine whether the mean difference in haemoglobin levels among three anaemic groups was significant. After the data analysis it was found that the prevalence of anaemia was 96.8%.

The most of the women were mild anaemic (75.3%). There were 16.9% women were moderate anaemic and 7.8% women found to be severe anaemic. Author found that the most affected age group was 21-25 years but their difference was not significant. The author finally suggested that we have to emphasize the importance of iron supplementation for all pregnant women, particularly during the prenatal period, with specific attention to be paid on the most affected areas.

Fabian Rohner et. al. (2012) have been checked whether a possession score or a poverty index best predicts undernutrition and anaemia in women of reproductive age and children aged 6-59 months. Cross-sectional study was conducted from Cote d'Ivoire in July 2007. The general census of 1998 used as a strategy for sampling. Each eco-region's representative district was picked at random, and within that district, the district capital (urban) and a number of rural regions were chosen. The proportional-to-population-size technique was used to determine the number of clusters for each ecoregion. There were sixty clusters altogether, with an equal number of clusters from urban and rural locations. Fourteen households were chosen from each cluster. The study concentrated on women of reproductive age (15–49) and children aged 6-59 months within the households. The Kish table was used for random selection, and just one person from each age category was chosen, yielding 840 WRA and pre-schoolers as the expected sample sizes.

Standard methods were employed to take anthropometric measurements. Using the WHO 2006 growth criteria, the Z-scores for each child's weight-for-height, weight-for-age, and height-for-age were derived. A child is considered stunted if HAZ is less than -2.00, underweight if WAZ is less than -2.00 and wasted if WHZ is less than -2.00. BMI was calculated for WRA and from that obtained BMI four categories were generated such as under-weight, normal weight, pre-obese and obese. On the basis of the Multi-dimensional Poverty Index (MPI), a possession score was created. The MPI possession score ranges from 0 (no possessions) to 4 (four possessions), and it is based on owning a radio, television, mobile phone, bicycle, or motorcycle. The poverty index was determined by factors related to household assets and characteristics, such as dwelling quality (roof and wall types), access to power and water, ownership of technological devices, and transportation.

SPSS software was used for the purpose of further statistical analysis. Skewness was checked for continuous variables by using Cox test. Association between two categorical variables were examined by using chi-square testing. Also, independent sample t-test and one-way ANOVA was used for continuous variables. The association between Hb concentration, residency, and the poverty index or possession score in WRA and children was examined using sequential multiple regression analysis. After eliminating effect of residency, there were no any noticeable differences in the Hb means for either the possession score or the poverty index for WRA. But when the poverty index was used to measure the sample as a whole, the Hb concentration showed

a consistent, significant downward trend from the richest to the poorest quintile. Whereas in children there was considerable effect on mean Hb but no obvious trend found. After controlling for age and place of residence, a stepwise multinomial logistic regression analysis revealed that the poverty index had no discernible impact on BMI category. It was revealed that there was no significant association between possession score and BMI category.

It was discovered that 50.2% of WRA reported anaemic, with 1% having severe anaemia whereas in children 74.9 % reported as anaemic, with 9.5% were suffered from severe anaemia. The present research has demonstrated that both preschool children and WRA from Cote d'Ivoire had high rates of undernutrition and anaemia. Residency and the poverty index were reliable indicators of mean Hb concentration and anaemia in both WRA and children. Similarly, residency and poverty index are the significant predictors for the BMI. While poverty index, possession score, and residency were found to be highly associated with mean HAZ and stunting in children. **Waseem Sharieff et.al.(2008)** examine the effects of two types of iron pots on haemoglobin (Hb) and serum ferritin (SF) concentrations in young children (6–24 months), adolescent girls (11–15 years), and women of reproductive age (15–44 years) whose households received iron pots for cooking food over a 6-month period. This research was conducted between October 2004 and March 2005 in the town of Porto Novo on Benin's Atlantic coast. For the data randomly assigned cast iron pots, blue steel pots, or oral iron supplements to 161 households, including 339 individuals from the three subgroups (control). In the control group, children received micronutrient Sprinkles™, and adolescent girls and women received iron tablets daily for 6 months. Hb, SF, and C-reactive protein concentrations were measured at baseline and 6 months. This is a clinical trial with a cluster-randomization design. In the statistical analysis first examine the data using descriptive statistics and histograms and done random effect linear regression models for continuous variables and random effects logistic regression models for binary variables. At the conclusion of the study, it was discovered that there were no significant differences in mean Hb concentrations or anaemia prevalence between any two groups, whereas Sf concentration differ significantly between groups. Finally, author concluded that controlling anaemia with iron pots (blue steel or cast iron) was ineffective.

Kayihan Pala et.al. (2008) developed a multiple logistic regression model to identify the risk factors of anaemia. This study focussing on prevalence of anaemia and

identifies the various risk factors associated with the anaemia among the women of reproductive age. For this purpose researcher used primary data by designing a questionnaire. The questionnaire was developed by Uludag University Faculty of Medicine, Department of Public Health. This cross sectional study was done in NPHTRA in Bursa between June 2004 and June 2005. There were total 6506 women in reproductive age here reproductive age is 15-49 years living in NPHTRA. Sample size was estimated as 530. The sample was chosen by using stratified random sampling. In this study pregnant women and those women who were not sure of pregnancy are excluded. Also menopause women and lacting phase women were excluded from the study. Finally total 488 women were agreed to participate in study. The questionnaire consists of 28 questions. The questions were depends on socio-demographic characters of women, their fertility information, family planning method, menstrual information. Haemoglobin of women was measured and those having haemoglobin level less than 12.0 g/dl were characterised as anaemic. For calculating BMI height and weight also measured. For the analysis purpose SPSS software was used. T test and chi-square test were used to compare means and percentages. To identify risk factors associated with anaemia multiple logistic regression was used. For the logistic regression presence of anaemia was considered as dependent variable and age, education, marital status, job, parity, body mass index, and menstruation characteristics (regularity of cycle, length of cycle, length of flow, sanitary pad usage were used as independent variables. It was found that there was 32.2% prevalence of anaemia. The women with at least one pregnancy had 30.3% anaemia prevalence and 36.4 % those for no pregnancy. The prevalence of anaemia was significantly ($P < 0.001$) higher in women who had 6-10 days of flow during menstrual cycle and used more sanitary pads than in women who had less days of flow and used 1-2 pads. From the logistic regression is was found that there was no association between age, education, income, marital status, occupation, parity, body mass index, regularity of cycle and length of cycle, and anaemia in this study. But more than 5 days menstrual bleeding and uses of more than 2 sanitary pads during menstruation are factors associated with anaemia. Finally, author concludes that nearly one out of every three women in the research area was found to be anaemic. This emphasises the importance of adopting a public health programme aimed at preventing and detecting anaemia in women of childbearing age.

2.3 Findings:

In general, previous research has yielded good findings in terms of predicting anaemia. The scientists also believe that utilising machine learning and data mining to create an ensemble methodology can aid in accurately predicting anaemia and reducing previous diagnostic errors. As a result, the WRA receives high-quality services as a result of the ensemble predictive technique. Henceforth, these algorithms like random forest, Ada boost, bagged decision tree and stacking ensemble model may give the best prediction and higher accuracy in prediction of disease in clinical research.

CHAPTER 3 METHODOLOGY

3.1 Introduction:

Data is the foundation of empirical research and critical analysis, making them an absolute necessity in all research. It is a crucial tool since it offers the starting point from which insights, patterns, and conclusions can be drawn. Data collection and analysis in a study allow researchers to test hypotheses, support or refute accepted theories, and come to wise judgements. Without data, research would remain simply theoretical and be unable to provide the empirical proof needed to confirm or deny claims. Data also strengthens the rigour and legitimacy of research by enabling replication and peer review, which guarantee the accuracy and reliability of results. The following figure shows various types of data that have been segmented.

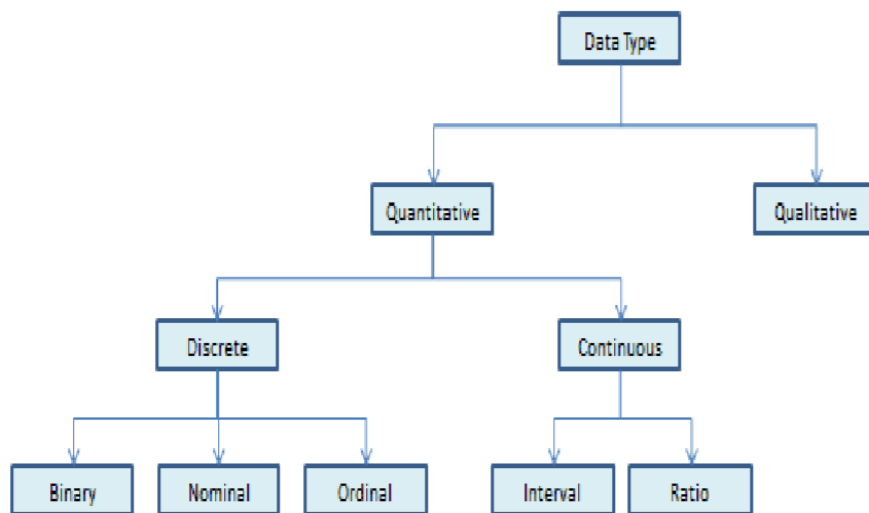


Fig. 3.1 Types of data [1]

3.2 Data Pre-processing and its need:

Data pre-processing is a crucial first step in any project involving data analysis or machine learning. It has set of procedures designed to clean, transform, and arrange unstructured data into a format appropriate for analysis or modelling. Real-world data is frequently disorganised, lacking, and inconsistent, necessitating data pre-processing. Data may be in multiple units or scales, have outliers, redundant information, or missing numbers. By filling in missing data, identifying and handling outliers, standardising units, and converting categorical variables into a numerical representation, pre-processing aids in resolving these problems.

The efficiency and accuracy of the model can be increased by using data reduction techniques to lower the dimensionality of high-dimensional input. Because the quality of the input data has a considerable impact on the outcomes and functionality of analytical models, effective data preparation is essential. It makes sure the data is trustworthy, consistent, and prepared for further analysis, enabling researchers and data scientists to draw forth important conclusions and take sensible action.

3.3 Data Visualisation and its importance:

To make difficult information easier to understand and analyse, data visualisation is the technique of displaying data in graphical or visual formats, such as charts, graphs, maps, and infographics. In many different industries, it is crucial for data analysis, communication, and decision-making.

In conclusion, data visualisation is an effective method for drawing conclusions from data, effectively presenting information, and assisting in data-driven decision-making. It is a crucial component of contemporary data analysis and has uses in a variety of industries, including business, science, education, and media.

3.4 Analysis of Data:

Data analysis is the procedure of looking through, purifying, manipulating, and interpreting data in order to draw forth important conclusions and make wise judgements. It is the foundation for all types of research, commercial operations, and problem-solving. The enormous amounts of data produced in the current digital era create the necessity for data analysis. Without analysis, raw data is frequently too much to handle and can result in information overload.

3.5 Pilot study:

In order to evaluate and improve research techniques, methods, and data gathering tools prior to the major research project, a pilot study is a small-scale preliminary investigation. It acts as an essential pre-research process, giving researchers perceptions into the viability, practicality, and probable problems of their study.

A pilot study is necessary for a number of reasons. First off, it assists researchers in recognising and correcting any defects or restrictions in their research design, guaranteeing that the methods used for data collecting and analysis are reliable. Second, a pilot study enables the fine-tuning of experimental protocols, questionnaires, or surveys, reducing ambiguity and guaranteeing that the instruments accurately measure the variables of interest.

Thirdly, it gives researchers a preview of any difficulties and roadblocks that can appear during the primary study, allowing them to make the required adjustments beforehand. Overall, a pilot study improves the quality and integrity of research, improving the chances that the major study will produce significant and trustworthy findings.

In this research the pilot study was done to assess the factors affecting to the status of anaemia. For the pilot study DHS data [India 2015-16](#) was used. The DHS is a well-known programme that carries out household surveys in over 90 countries, mainly in low- and middle-income countries. These surveys collect thorough information on a range of demographic and health variables, including nutrition, HIV/AIDS, family planning, maternal and child health, and other topics.

In order to inform health and development policies and programmes, DHS surveys seek to collect high-quality data. They are used to study demographic and health trends, assess how well the world is doing towards its development goals, and provide evidence-based solutions. DHS surveys are normally carried out by conducting personal interviews with women of reproductive age (WRA) and, occasionally, with men in particular homes. The data collected are representative of the entire country and give an overview of the demographic and health status.

Accessibility of the DHS data: The DHS Programme website makes its data available to the general public and researchers without charge. Datasets, reports, and analytical tools are available to researchers. The National Family Health Survey (NFHS), which is conducted in India as the Demographic and Health Survey (DHS), is the main repository for statistics on the country's population and health. The National Family Health Survey (NFHS) is a comprehensive, nationally representative survey that collects vital data on various aspects of India's population, health, and nutrition.

DHS data Description:

The data with 46 variables were taken from DHS 2015 these variables are as follows:

1. Anaemia: This variable likely represents the presence or absence of anaemia in the survey participants. Anaemia is a condition characterized by a deficiency of red blood cells or haemoglobin in the blood.
2. URBAN: This variable may indicate whether the survey participant resides in an urban or rural area. It is often used to classify the location of the participants.
3. RESIDEINTYR: This variable could pertain to the number of years a participant has lived in their current residence.

4. AGE: The age of the survey participants, usually measured in years.
5. RESIDENT: This variable might be related to the type of residence, such as 'urban' or 'rural.'
6. PREGNANT: It likely indicates whether a female participant is pregnant at the time of the survey.
7. DURCURPREG: This variable may represent the duration of the current pregnancy for pregnant participants.
8. RELIGION: The religious affiliation or belief system of the survey participants.
9. MARSTAT: Marital status, which could include categories such as 'married,' 'single,' 'divorced,' etc.
10. AGEFRSTMAR: The age at which a participant got married for the first time.
11. HUSAGE: Household usage or characteristics related to the household.
12. CHEB: This variable's meaning depends on the specific survey, but it could refer to a type of child health or well-being indicator.
13. PREGTERMIN: It might indicate the outcome of a pregnancy, whether it resulted in a live birth, stillbirth, abortion, etc.
14. AGEMENARCHE: Age at menarche, which is the age at which a female participant had her first menstrual period.
15. AGEAT1STBIRTH: Age at the time of a woman's first childbirth.
16. HHMEMTOTAL: The total no. of household members.
17. HHKIDLT5: Total children in the household under the age of 5.
18. HHELIGWOMEN: The number of eligible women in the household (e.g., women of reproductive age).
19. ELECTRC: Likely represents access to electricity in the household.
20. COOKFUEL: The type of fuel used for cooking in the household.
21. TOILETTYPE: The type of toilet facilities used by the household.
22. BPLCARDHH: Whether the household possesses a below poverty line (BPL) card, often used for government welfare programs.
23. CURRWORK: Current work status of the survey participant.
24. WKCURRJOB: Current job or employment status.
25. HUSJOB: The job or employment status of the household's head (husband or head of the household).
26. WEALTHS: A measure of household wealth or socioeconomic status.
27. EDUCLVL: Educational level or attainment of the survey participants.

28. EDYRTOTAL: The total number of years of education completed by participants.
29. NEWSBRIG: Exposure to news or information through a source like television, radio, or newspapers.
30. FPKNOTYP: Family planning knowledge or method types.
31. FPMETHNOW: The current family planning method in use by participants.
32. ALDRINK: Alcohol consumption behaviour.
33. TOSMOKE: Smoking behaviour or tobacco use.
34. ATEDAIRYFQ: Frequency of dairy consumption.
35. ATEEGGFQ: Frequency of egg consumption.
36. ATEFISHFQ: Frequency of fish consumption.
37. ATEFRUITFQ: Frequency of fruit consumption
38. ATEGRNVEGFQ: Frequency of green vegetable consumption.
39. ATELEGUMFQ: Frequency of legume consumption.
40. ATEMEATFQ: Frequency of meat consumption.
41. ATEFRIEDFQ: Frequency of fried food consumption.
42. DRANKSODAFQ: Frequency of soda or soft drink consumption.
43. BIOFHHAGE: Age of the head of the household.
44. ASTHMA: The presence or absence of asthma as a health condition.
45. GOITER: The presence or absence of goiter, often related to iodine deficiency.
46. HEARTDIS: The presence or absence of heart disease.

Decision tree and Random forest algorithms were developed on 45 predictor variables to predict the status of anaemia ('No', 'Mild', 'Moderate', 'Severe'). After the pilot study the main analysis was done on the primary data which was collected by designing well designed questionnaire. Therefore, in the next section the primary data was explained in detail.

Primary Data information: The dataset contains information about women at reproductive age (WRA) between the ages of 14 to 49. The data was collected from the hostel girls, pregnant women and non-pregnant women. Therefore, the main data itself has three sub datasets. The parameters or questions are slightly different for the Unmarried women, married women and married women with pregnancy. It encompasses a diverse set of 56 variables that capture a wide range of characteristics related to the participants' health, lifestyle, and socioeconomic status. Out of these 56 some are not used or applicable for unmarried women like pregnancy related questions, contraceptive related questions, etc. Similarly, some of questions not applicable for

nonpregnant WRA such as pregnancy related questions. Among these variables, 'Anaemia' represents the presence and possibly the severity of anaemia in the participants, a condition of low red blood cell count. 'HIV status' indicates whether the participants are HIV-positive or HIV-negative, offering insights into the prevalence of HIV within this specific age group. 'Are you feeling weak or dizzy?' is a binary variable that may provide a glimpse into the overall health and well-being of the participants. Age, a fundamental demographic variable, is recorded to understand the age distribution within the dataset. 'Education(years)' quantifies the number of years of formal education each participant has received, shedding light on their educational backgrounds. 'Occupation' is likely to describe the participants' employment or profession. Economic status is reflected in 'Income of the family (Rs. Annual),' which documents the annual income of the participants' families in Indian Rupees. 'Weight(kg)' and 'Height (meter)' provide essential health indicators, representing the weight and height of the participants, respectively. These metrics are integral for calculating the 'BMI,' which is a measure of body composition and overall health. 'Eating Habits' and 'Food type' shed light on their dietary patterns and preferences, providing a window into their nutritional habits. 'Daily Tea intake' reveals their daily tea consumption behaviour, potentially reflecting cultural and lifestyle choices. 'Acidity Problem' indicates whether participants experience acid reflux or related issues, while 'Alcohol Consumption' provides information about their alcohol consumption patterns, which can impact health and lifestyle decisions.

Additionally, the dataset includes variables related to health and habits. 'Any Addiction' and 'Type of Addiction' capture any addictive behaviours among the participant while 'Suffer from any long-term disease' highlights the presence of chronic or long-term health conditions. 'Suffer from stress' delves into participants' mental well-being, while 'Use Iron supplementation' informs about their nutritional practices. Finally, 'Suffers from Diabetes' is a binary variable indicating whether participants are diagnosed with diabetes, a significant medical condition.

These variables collectively cover a wide spectrum of factors, including health-related parameters such as anaemia and HIV status, as well as socio-economic indicators like education, income, and occupation. They offer a comprehensive overview of the lives and health of unmarried WRA in the dataset, making it possible to conduct in-depth analyses and gain insights into various aspects of their well-being and circumstances.

‘Acidity Problem’ informs about any acid reflux or related health issues, while ‘Alcohol Consumption’ details their alcohol consumption patterns, influencing their lifestyle and health decisions. ‘Any Addiction’ and ‘Type of Addiction’ reveal information about addictive behaviours, while ‘Suffer from any long term disease’ highlights the presence of chronic health conditions. ‘Suffer from stress’ delves into participants’ mental well-being, while ‘Use Iron supplementation’ informs about their nutritional practices. ‘Suffers from Diabetes’ is a binary variable indicating the presence of diabetes, a significant medical condition.

Furthermore, the dataset includes variables related to socio-economic and environmental aspects. ‘Household Wealth status’ reflects the economic well-being of their households, while ‘Number of family members’ offers insights into family size. ‘Toilet facility,’ ‘Drinking water source,’ and ‘Cooking fuel’ are essential for assessing the living conditions and access to basic amenities. ‘Exposure to domestic violence’ delves into the sensitive issue of domestic violence, ‘Avg.of rest in day (per Hr)’ indicates their daily rest patterns, and ‘Regular visit to doctor’ provides information about healthcare-seeking behaviour. ‘Daily eat fresh fruits/Vegetable /Milk’ reflects their dietary choices, and ‘Menstrual cycle 1’ and ‘Menstrual cycle 2’ are likely related to their menstrual health. ‘No of pads (per day)’ and ‘days of blood flow’ are crucial for understanding menstrual patterns, and ‘Pain on menstrual period’ details their experiences. ‘Age at menstrual cycle begins’ is essential for understanding reproductive health, and ‘Region’ may provide insights into geographic variations. ‘Number of years lives in residential area’ reflects their residential stability, while ‘Mass media exposure’ is indicative of media consumption patterns. Finally, ‘Community women education’ likely relates to their exposure to education and empowerment initiatives.

All the data variables were coded for further statistical analysis. The coding of the variable is as follows: Here is the information about the variables and their coding: Hb (gm/dl): Haemoglobin levels, an essential blood parameter, are measured in grams per decilitre (gm/dl) according to WHO guidelines anaemia was categorise into “no anaemia”, “mild anaemia”, “moderate anaemia” and “severe anaemia” by using Hb cut off values.

HIV status: Participants are coded as 1 for ‘Yes’ if they have HIV and 0 for ‘No’ if they don’t.

Weakness or Dizziness: This variable is binary, with 1 indicating ‘Yes’ for experiencing weakness or dizziness and 0 for ‘No’.

Disease Suffering: This variable was nominal which contains names of the disease of WRA suffering.

Age: This variable represents the age of participants which is numeric variable which include age of WRA in years.

Education of WRA: Participants' education levels are coded from 0 to 4, corresponding to primary, secondary, higher secondary, graduate, and post-graduate levels of education, respectively.

Occupation: Coding includes 0 for housewives, 1 for working women, 2 for those in agriculture, and 3 for students.

Income: It's a quantitative variable which shows annual income of the family.

BMI : Calculated by using weight in kilogram and height in meter.

Eating Habits: These habits are coded from 0 to 14 to represent various combinations of spicy, sweet, salty, and sour preferences.

Food Type: Coded as 0 for vegetarian and 1 for non-vegetarian.

Daily Tea Intake: A binary variable coded as 0 for 'No' and 1 for 'Yes' for daily tea consumption.

Acidity Problem: It's binary, with 0 for 'No' and 1 for 'Yes'.

Marital Status: Participants' marital status is coded as 0 for 'No' and 1 for 'Yes' (married).

Age at Marriage: This variable is quantitative which represents age of WRA at the time of marriage in years.

Husband's Age and Age at Marriage: Both are quantitative variables.

Husband's Occupation: Coded as 0 for farmer, 1 for laborer, 2 for job, and 3 for business.

Husband's Education: Education levels of husbands are coded from 0 to 4, with 5 representing 'No education'.

Alcohol Consumption: Coded as 0 for 'No' and 1 for 'Yes'.

Any Addiction: A binary variable with 0 for 'No' and 1 for 'Yes'.

Type of Addiction: Coded from 0 to 4 for different types of addictions.

Long-Term Disease: Coded as 0 for 'No' and 1 for 'Yes'.

Suffering from Stress: Binary, with 0 for 'No' and 1 for 'Yes'.

Iron Supplementation: A binary variable, coded as 0 for 'No' and 1 for 'Yes'.

Suffers from Diabetes: Coded as 0 for 'No' and 1 for 'Yes'.

Household Wealth Status: Coded as 0 for 'Poor', 1 for 'Middle Class', and 2 for 'Rich'.

Toilet Facility: Binary coding, 0 for 'No' and 1 for 'Yes'.

Drinking Water Source: Coded as 0 for surface, 1 for tap, and 2 for well.

Cooking Fuel: Various combinations of cooking fuel sources are coded from 0 to 26.

Exposure to Domestic Violence: Binary, with 0 for 'No' and 1 for 'Yes'.

Regular Visit to Doctor: Coded as 0 for 'No' and 1 for 'Yes'.

Daily Consumption of Fruits/Vegetables/Milk: Binary, with 0 for 'No' and 1 for 'Yes'.

Menstrual Cycle: Coded as 0 for 'Below 25', 1 for '25-30', and 2 for 'Above 25'.

Regular Menstrual Cycle: Binary, with 0 for 'Irregular' and 1 for 'Regular'.

No of Pads and Days of Blood Flow: These are quantitative variables.

Pain on Menstrual Period: Coded from 0 to 3 for different pain levels such as 'No', 'mild', 'moderate', 'severe' respectively.

Age at Menstrual Cycle Begins: A quantitative variable.

Pregnancy Status: Binary, with 0 for 'No' and 1 for 'Yes'.

Gestational Month: This variable categorizes the gestational month into three groups, with 0 representing the first trimester (1-3 months), 1 for the second trimester (4-6 months), and 2 for the third trimester (7-9 months) during pregnancy.

Number of Children/kids Ever Born: This is a quantitative variable, representing the total number of children a participant has given birth to during her lifetime.

Premature Delivery: A binary variable, coded as 0 for 'No' and 1 for 'Yes' to indicate whether participants have experienced premature deliveries.

Miscarriage History: Similar to premature delivery, this variable is binary, with 0 for 'No' and 1 for 'Yes' to denote whether participants have a history of miscarriages.

Age at First Birth of Child: This is a quantitative variable, capturing the age at which participants gave birth to their first child.

Age of Last Child: Also a quantitative variable, indicating the age of participants' most recent child.

Total Number of Births in recent 5 Years: A quantitative variable that represents the number of childbirths in the last five years.

Use of Contraceptive: A binary variable, coded as 0 for 'No' and 1 for 'Yes' to indicate whether participants use contraceptives.

Method of Contraceptive: Coded from 0 to 5 to represent different contraceptive methods, including 'No' contraception, 'Condom' (1), 'Copper T' (2), 'Pill' (3), 'Injection' (4), and 'Other' (5).

Region: A binary variable coded as 0 for 'Rural' and 1 for 'Urban,' representing the region where participants reside.

Number of Years Lived in Residential Area: This is a quantitative variable, indicating the total years participants have lived in their current residential area.

Mass Media Exposure: A binary variable, coded as 0 for 'No' and 1 for 'Yes' to denote whether participants have exposure to mass media.

Community Women Education: Similar to mass media exposure, this variable is binary, with 0 for 'No' and 1 for 'Yes,' indicating whether participants are part of a community where women receive education.

3.6 Statistical Analysis:

3.6.1 Supervised Machine learning:

Supervised-learning is a ML technique that predicts the output through the use of appropriately labelled training data. The term 'labelled data' denotes input data to which the corresponding output has been previously assigned. Supervised learning operates under the guidance of training data, which functions as the supervisor and instructs the computers on how to generate accurate predictions of the output. It utilizes the identical concept that a learner would acquire knowledge from an instructor. Supervised learning is a process that entails providing the machine learning model with accurate input and output data. The principal objective of a SL algorithm is to distinguish between the input variable (x) and the output variable (y) by locating a mapping function.

Working of supervised learning algorithms.

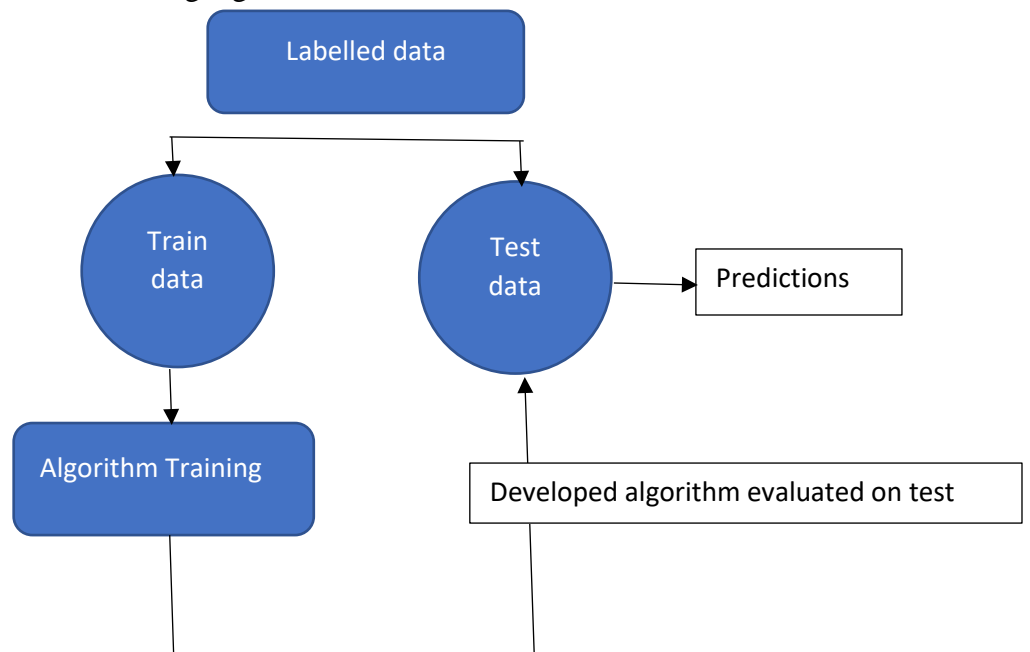


Fig. 3.2 Flow of Supervised learning algorithm.

The above fig 3.2 is about SL algorithm. The primary goal of SL methods is to identify the relationship that exists between an input variable (independent variable) and an outcome variable (dependent variable). A structure known as an algorithm which contains a representation of the identified relationship. When the values of the input attributes are known, algorithm can be used to predict the value of the outcome variable by describing and explaining phenomena that are hidden in the dataset. Algorithms are trained using labelled datasets in supervised learning, where the model learns about various types of input. After the training phase is over, the model is tested with test data and then it generates predictions for the output.

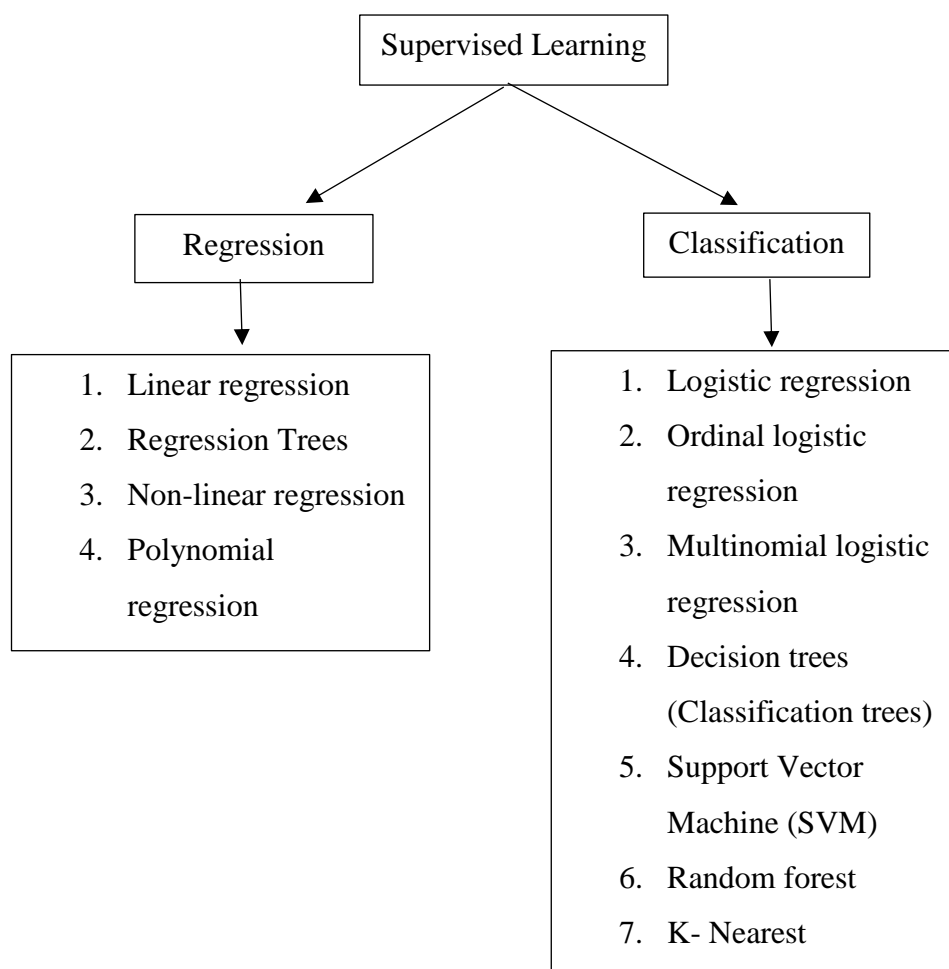


Fig 3.3 Types of supervised learning models

Ordinal logistic regression:

One type of categorical variable is known as an ordinal variable, and it is distinguished by the different ordering of the category levels. In order to represent the relationship that exists between an ordinal dependent or response variable and one or many independent variables, a statistical analysis method known as ordinal logistic

regression can be utilized. Continuous or categorical variables may serve as explanatory factors. OLR is an extension of LR method that establishes a relationship between the independent/explanatory variables and the logit of a binary response. Instead, $k-1$ logits are present when the response variable possesses k levels. Ordinal logistic regression is predicated on the proportional odds assumption, which states that the effect of an independent variable remains constant as the degree of the response increases. In consequence, the OLR will produce an output consisting of an intercept for all response levels other than one and a solitary slope for each explanatory/independent variable.

Decision tree (Classification tree):

A DT is a tree like algorithm which is used for decision making. It is made up of nodes indicating decision points and branches indicating the outcomes of those decisions. The decision points can be defined by the input variable values, and the results are possible classifications or predictions.

Classification and Regression Tree Algorithm (CART):

There are a great number of methods that may be used to predict numerical as well as categorical response/dependent variables based on continuous or categorical predictors. For instance, in general linear models and general regression models, we have the ability to describe a linear combination of continuous and categorical/dependent predictors in order to make a prediction about a continuous dependent variable. In the context of general discriminant function analysis, we are able to provide a design for predicting categorical variables, which is to address the problem of classification.

Regression type problem regression:

Problems of the regression kind typically involve estimating the values of a continuous variable from one or more categorical or continuous predictor factors. For example, we may want to predict the selling prices of houses which is considered as continuous dependent variable from various other continuous predictors, like area in square foot of house and as well as categorical variables like zip code, area, style of the home, etc.

Classification-type problems:

Predicting values of a categorical dependent variable from one or more continuous and categorical predictor variables is the general goal of classification-type tasks. For instance, we might be interested in forecasting who will or won't buy a specific item from the store or whether or not to renew a membership. This would be example of simply binary classification problem, where the categorical dependent variables can

only assume two distinct and mutually exclusive classes. There are various methods for analysing classification-type problems and to make prediction. CHAID also analyses classification-type problems and produces results that are similar to those computed by CART.

Classification tree:

Generally speaking, the goal of analysis using the tree building approach is to identify a set of logical split conditions that allow for precise class prediction. Take the well-known Iris data categorization problem, for instance, which Fisher presented. The lengths and widths of the petals and sepals of three different types of iris such as Setosa, Versicolor, and Virginica are reported in the data called "Iris data." Finding out how to distinguish between the three types of flowers using the four measurements of petal and sepal width and length is the aim of the investigation. To calculate classification scores and provide the user with the projected classification for each observation, discriminant function analysis will estimate several linear combinations of predictor variables. Rather than using linear regression to forecast or categorise the classes like Setosa, Versicolor, and Virginica, a classification tree will establish a series of logical-if-then conditions.

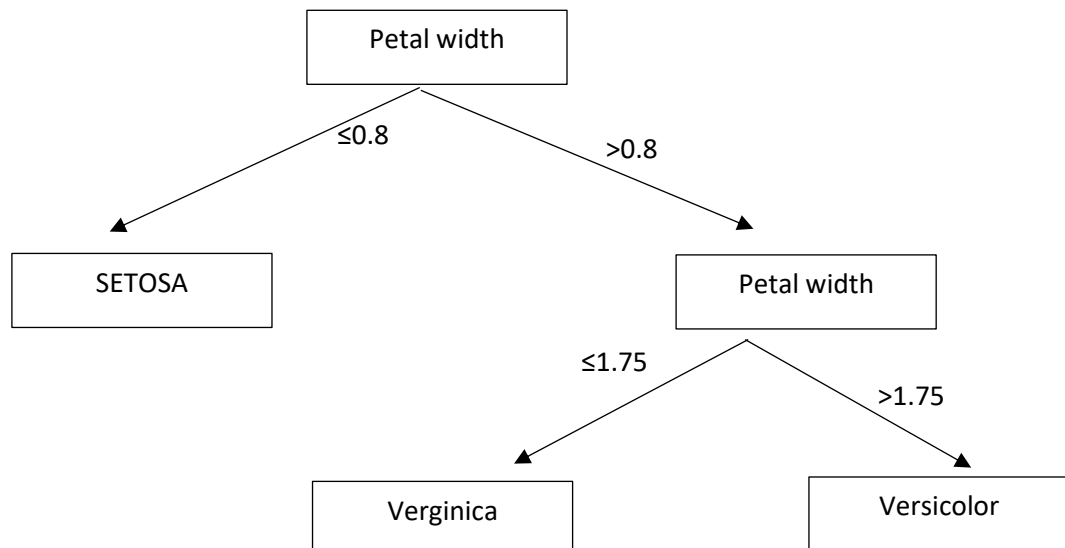


Fig 3.4 Classification tree

Regression Tree:

The general method for creating predictions from a few straightforward if-then statements can also be used to solve regression problems. These illustrations are based on data on poverty that include census data from 1960 to 1970 for 30 randomly chosen nations. Finding the best variable to predict the percentage of families living below the poverty line in a nation was the research topic for this example.

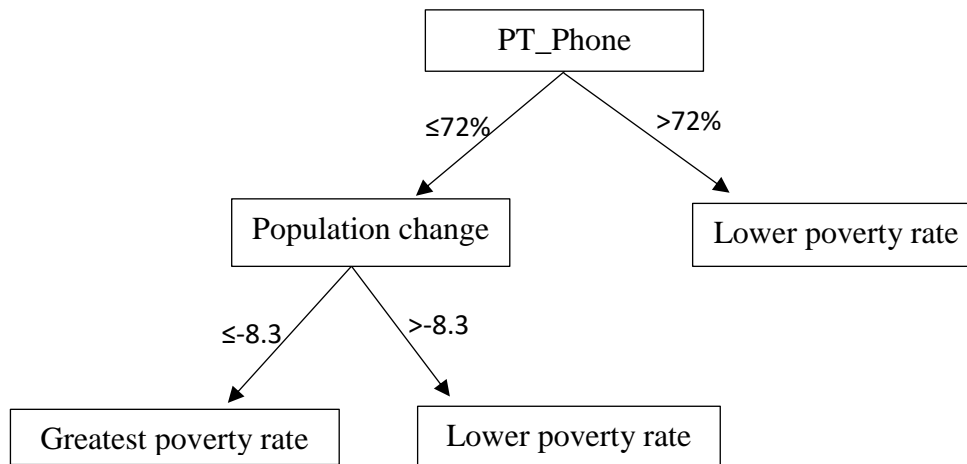


Fig 3.5 Regression tree

This finding can be understood in a fairly simple way. Generally speaking, poverty rates are lower in nations where more than 72% of households own a phone. The countries with the highest rates of poverty are those where the population change is less than -8.3 and less than or equal to 72% of households have a phone. These findings are simple to understand and present. The majority of households in countries with low rates of poverty own telephones.

Classification and regression tree (CART):

CART is one of the popular method of building decision tree in the ML community. Decision trees algorithm suffers from different problems like missing value overfitting, etc. To overcome such problems CART can be good solution. This algorithm was developed by Leo Breiman, Jerom Friedman Richard Olshen and Charles Stone. CART divides data at each node based on a single attribute function to create a binary decision tree. CART determines the optimal split using a gini index. We now try splitting each of the two nodes that the first split produced in the same way as the root node. Once more, in order to identify the potential splitters, we look over every input field. We designate a node as a leaf node if no split that considerably reduces its diversity can be discovered; eventually, only the leaf node remains, and the entire decision tree is depicted.

Due to overfitting, the entire tree should not be treated as it does not perform the best job of classifying a fresh batch of records. Every training set record has been assigned to a leaf of the entire decision tree by the time the tree has reached its mature state. Each leaf can now be assigned a class. The error rate of leaf node is the percentage of incorrect classification at the node. The weighted sum of the error rates of each leaf in the class decision tree represents the overall error rate. Each leaf contribution to the

total is the error rate of that leaf multiplied by probability the record will be end up there. tree-based approach are straightforward yet effective; they divide the feature space into a collection of rectangles, and then fit a basic model into each one.

Think of a regression problem where the inputs (X_1 , X_2) each take values in the unit interval and the output (Y) is a continuous response.

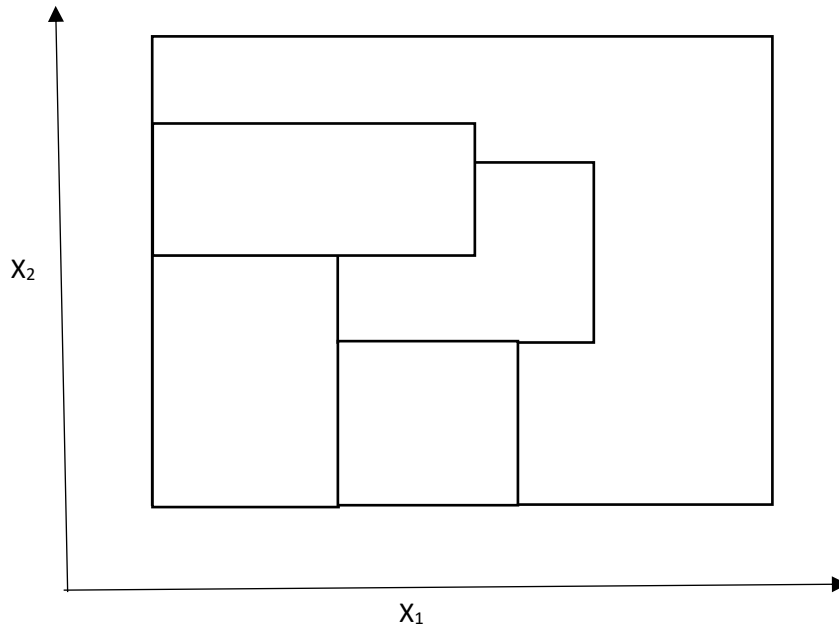


Fig 3.6 partition of feature space

A feature space split by lines parallel to the coordinate axis is depicted in Figure 3.6. We can use a different constant to model Y in each split. Each partitioning line has simple description, like X_1 is equal to C . There is problem because of some regions are complicated to describe. To simplify this, we do recursive binary partitions like that shown in below.

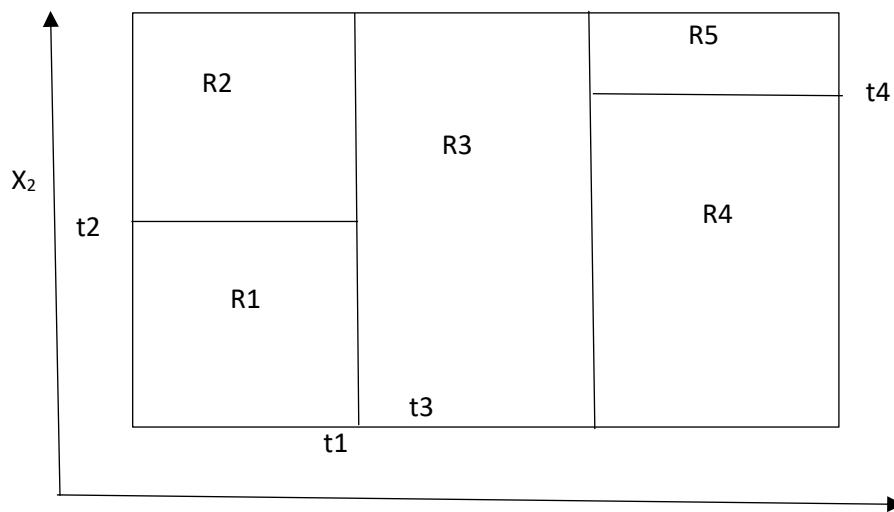


Fig 3.7 Dividation of sample space

From this partitioning we can draw decision tree such as:

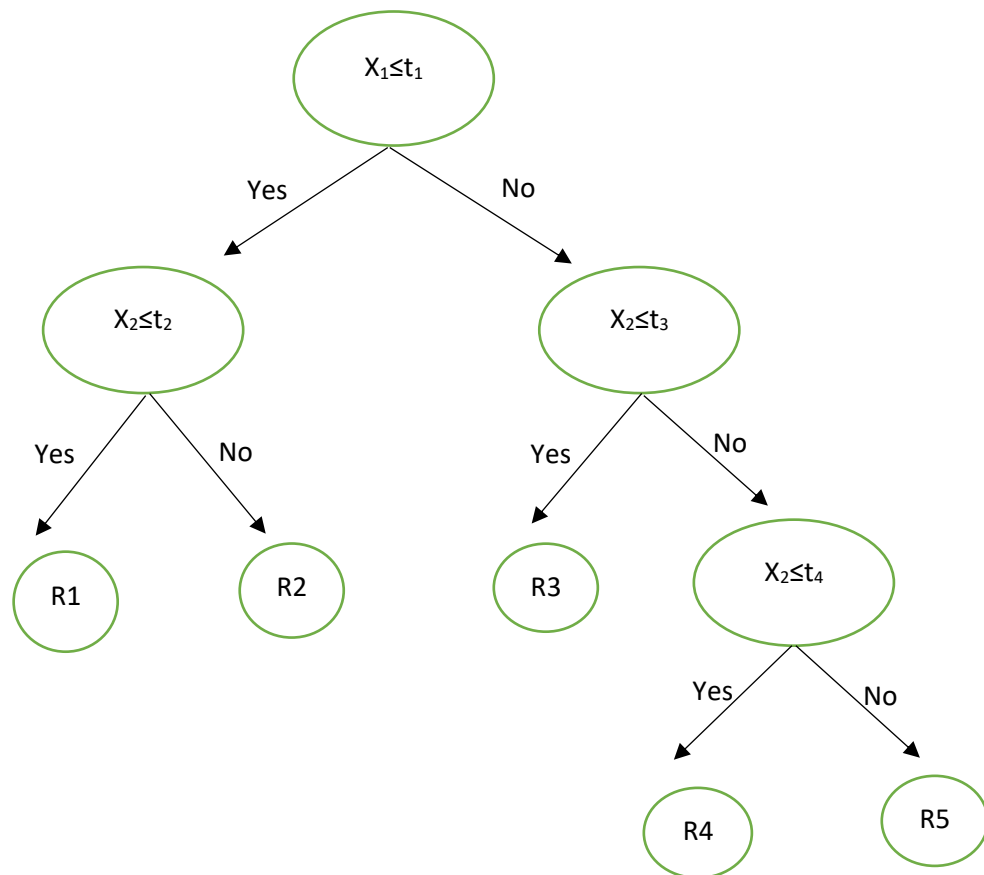


Fig 3.8 Decision tree

Gini Impurity:

Gini impurity is also known as Gini index. Within the context of a decision tree algorithm, the Gini index is a metric that evaluates the quality of a split. In other words, it determines the likelihood of a random sample being incorrectly classified. The Gini index is a statistical measure that determines the frequency with which a randomly selected element from the set would become incorrectly classified if it were randomly classified in accordance with the distribution of labels in the subset.

The Gini index/impurity of a node can be calculated as follows:

$$Gini = 1 - \sum_{i=1}^c p_i^2$$

Where,

c is the number of classes/categories

pi is the probability of a randomly chosen element in the node being labelled as class i.

The splitting attribute is determined to be the one that maximises impurity reduction or, alternatively, has the lowest gini index. This selected attribute (variable) and either its splitting subset or split point together forms a splitting criterion.

Support Vector Machine:

Support vector machines (SVMs) are algorithms for categorising linear as well as nonlinear data. An SVM is a basic algorithm that functions in this way. Nonlinear mapping is used to raise the original training data to a higher dimension. Within this new dimension, it looks for the linear optimal separation hyperplane—a "decision boundary" that divides tuples of one class from those of another. Every time, two classes of data can be divided by a hyperplane with a suitable nonlinear mapping to a high enough dimension. This hyperplane can be located by applying margins (as indicated by the support vectors) and support vectors.

We'll go through these new ideas in greater detail later. The discovered support vectors also provide a concise description of the learned model. SVMs can be used for both numerical prediction as well as categorical classification. They have been used in a variety of applications, including handwritten digit recognition, object recognition, speaker identification, and benchmark time-series prediction tests.

SVM with Linearly Separable data:

A support vector machine (SVM) is a method for the classification of both premier and non-linear data. SVM algorithm needs work as follows. It uses the non-linear mapping to transform the original training data into higher dimension within this new dimension. It searches for the linear optimal separating hyperplane (i.e. a decision boundary separating the tables for 1 class from another.) With an appropriate nonlinear mapping to sufficiently high dimension data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane by support vectors and margins.

The SVM algorithm works in two cases. In first case if the data is linearly separable or in the second case, the data is linearly not separable.

A case when the data are linearly separable:

Let us consider a two-class situation in which the classes can be separated linearly. Let the data set D be given as $(X_i, Y_i), i=1,2,\dots,n$.

Where X_i is the set of training tables with associated class labels Y_i . Each Y_i can take one of two values, either +1 or -1, corresponding to the respective classes.

To aid individualisation. Let's consider an example based on two attributes such as X_1 and X_2 .

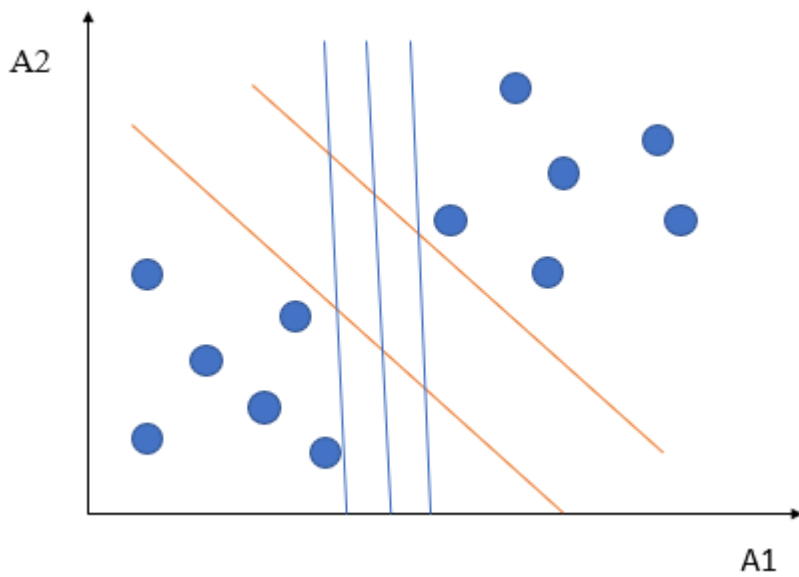


Fig 3.9 Linearly separable 2D data

From the graph, we can see here the two-dimensional data is linearly separable. One could draw an unlimited number of lines to separate things. Our goal is to identify the optimal candidate, which is the one with the lowest classification error on tuples that haven't been seen before. In the event that the provided data is three dimensional, we will have three attributes for each class. In general, we would like to determine the optimal separating plane in N dimensions. Our goal is to identify the optimal hyper plane. The decision boundary will be referred to as the hyper plane.

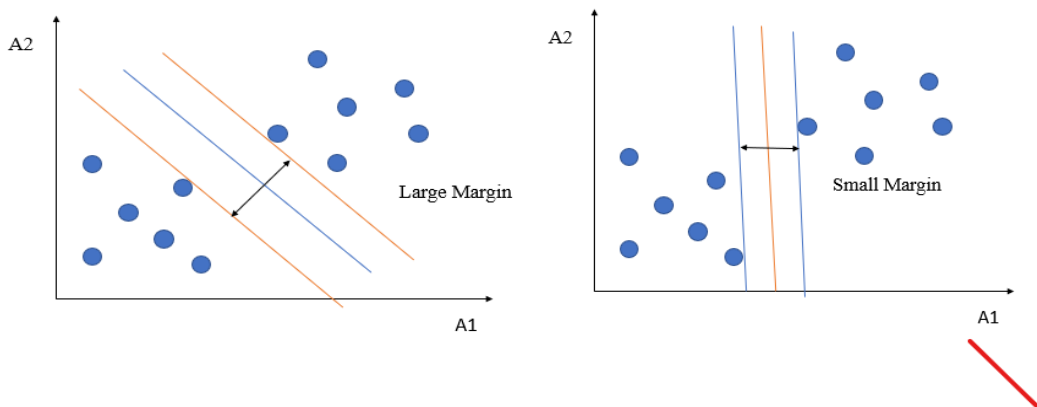


Fig 3.10 Margin of Hyperplane

Examine the above figure, which displays two distinct separating hyperplanes along with the margins that go with them. All of the provided data tuples are accurately classified by both hyperplanes. In contrast to the hyperplane with the lower margin, we anticipate the hyperplane with the bigger margin will be more accurate in identifying

future data tuples. For this reason, the SVM looks for the maximum marginal hyperplane, or the hyperplane with the biggest margin. The associated margin, or MMH, provides the greatest class separation. Developing a common definition of margin. When the margin's sides are parallel to the hyperplane, we can state that the shortest path from the hyperplane to one of its margins is equal to the shortest path from the hyperplane to the other side of the margin. In the context of MMH, distance really refers to the shortest path between the MMH and the closed set training pair on either side of the classroom.

A separating hyperplane can be written as,

$$W * X + b = 0 \quad \text{-----}(1)$$

Where,

W is weight vector, $W=(w_1, w_2, \dots, w_p)$

p is number of attributes

b is scalar often considered as bias

To aid individualisation. Let us consider two attributes, X1 and X2 training tuple of three-dimensional data. The values of attributes for X1 , X2 are x1 and x2 respectively for vector X.

If we consider b as an additional weight w0, we can rewrite equation (1) as,

$$w_0 + w_1x_1 + w_2x_2 = 0 \quad \text{-----}(2)$$

Thus, any point that lies above the separating hyperplane satisfies,

$$w_0 + w_1x_1 + w_2x_2 > 0 \quad \text{-----}(3)$$

Similarly, Any point that lies below the separating hyperplane satisfies,

$$w_0 + w_1x_1 + w_2x_2 < 0 \quad \text{-----}(4)$$

The weights can be adjusted so that the hyperplane defining the sides of the margin can be written as,

$$H_1: w_0 + w_1x_1 + w_2x_2 \geq 1 \quad \text{for } Y_i = +1 \quad \text{-----}(5)$$

$$H_2: w_0 + w_1x_1 + w_2x_2 \leq -1 \quad \text{for } Y_i = -1 \quad \text{-----}(6)$$

That is any tuple that falls on or above H₁ belongs to class +1 and any tuple that falls on or below H₂ belongs to class -1.

Combining the hyperplanes (5) and (6) we get,

$$y_i(w_0 + w_1x_1 + w_2x_2) \geq 1, \quad \forall i \quad \text{-----}(7)$$

Any training couple that fall on hyperplane H₁ or H₂ satisfies equation (7) are called support vectors. That is, they are equally close to the MMH. The support vectors are

the most difficult tuples to classify and give the most information regarding classification.

From this, we can obtain a formula for the size of maximal margin, the distance from separating hyper plane to any point on H1 is $1/\|W\|$, Where $\|W\|$ Is the Euclidian norm of W by definition. This is equal to the distance from any point on H2 to the hyper plane. Therefore, the maximal margin is $2/\|W\|$.

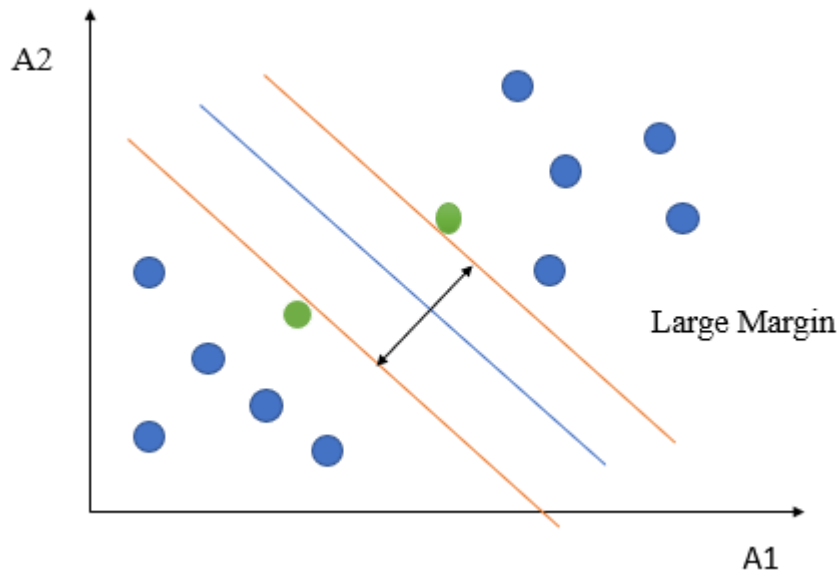


Fig 3.11 Maximum margin Hyperplane

Support vectors are displayed in the figure as green. We have a trained support vector machine after locating the support vectors and MMH. Since the MMH is a linear class boundary, data that can be classified linearly will be classified using the matching SVM. We refer to such a training SVM as a linear SVM Based on Lagrangian formula, the MMH can be rewritten as the decision boundary

$$d(X^T) = \sum_{i=1}^l y_i \alpha_i X_i X^T + b_0 \quad \text{---(8)}$$

Where,

y_i is the class label of support vector x_i

X^T is the test tuple.

α_i and b_0 is numerical parameters that where obtained automatically by the optimization of SVM algorithm.

l is the number of support vectors.

We input test data point X^T into equation (8) and determine the sign after that. This indicates the side of the hyperplane where the test data point is located. If the sign

is positive, SVM predicts that X_T is a member of class one since it falls on or above the MMH. In the event that the sign is negative, X_T is predicted to slip into class 1 and sit below the MMH. It can be observed that there is a dot product between the test X_T and the support vector X_i in the Lagrangian Formula for our problem (8).

Rather than the dimensionality of the data, the total of support vectors was what defined the complexity of the learnt classifier. Therefore, compared to certain other techniques, SVM likely to be less prone to overfitting. Support vectors are the training pair nearest to the decision boundary (MMH), which makes them critical and important. The same separation hyper plane would be discovered if all other training tuples were eliminated. Additionally, the number of support vectors discovered can be used to estimate and limit the SVM algorithm's expected error rate, which is independent of the dimensionality of the data. Even with a high degree of data dimensionality, an SVM with few support vectors can achieve good generalisation.

When the given data is linearly inseparable:

We learned about linear SVMs for classifying linearly separable data. But what if the data are not linearly separable? As shown in below figure. In this case, no straight line can be found that would separate the classes. For the purpose of classifying linearly inseparable data, non-linear SVMs can be created using the same methodology as linear SVMs. SVMs are examples of algorithms that can locate non-linear decision boundaries in input space. By expanding the method for linear SVMs, we may get the non-linear SVM in the following way. The first stage consists of two primary steps: first, we use a non-linear mapping to transform the original input data into a higher dimensional space. In this stage, a number of popular nonlinear mappings can be applied. Following the data transfer to the new location, we are once more faced with a quadratic optimisation problem, which may be resolved with the help of the linear SVM formalisation. We again end up with a quadratic optimization problem that can be solved using the linear SVM formulation. The MMH found in the new space corresponds to a non-linear separating hyperplane in the original space.

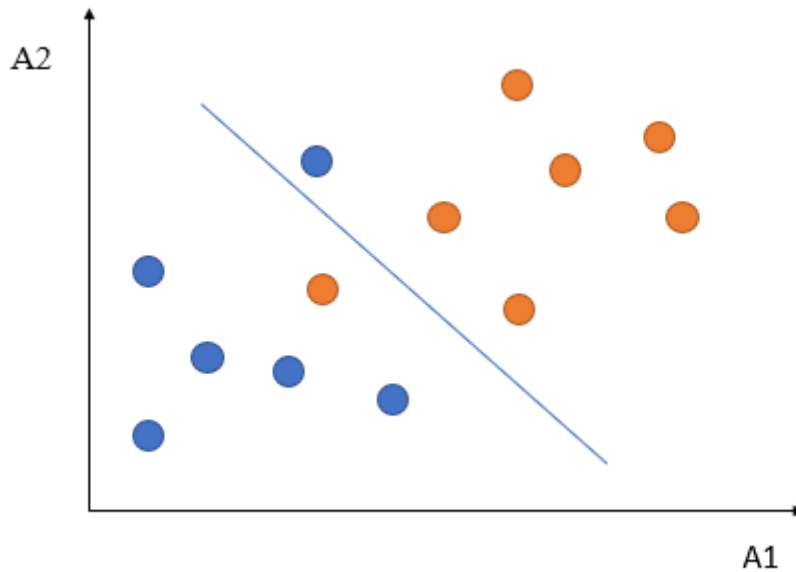


Fig 3.12 SVM decision

Tuning parameters in SVM:

Kernel:

In linear SVM, the problem is transformed using some linear algebra in order to learn the hyperplane. The kernel is involved in this. Equation (8) illustrates how to derive the equation for a linear kernel that predicts a new input by taking the dot product of each support vector (x_i) and the input (x).

Polynomial kernel of degree a : $K(X_i, X_j) = (X_i \cdot X_j + 1)^a$

Gaussian/radial kernel: $K(X_i, X_j) = e^{-\|X_i - X_j\|^2 / 2\sigma^2}$

Sigmoid kernel: $K(X_i, X_j) = \tanh(X_i \cdot X_j - \delta)$

Regularization

To the SVM optimizer, the Regularisation parameter (abbreviated C) tells it how much you want to avoid misclassifying every training example. In cases where a smaller-margin hyperplane performs better at reliably categorising all of the training points, the optimisation chooses it for large values of C . Conversely, an extremely small value of C motivates the optimizer to look for a hyperplane with a higher margin of separation, even if it misclassifies more points.

Margin:

SVM tries to search the good margin hyperplane. A margin is the distance between the nearest class points on a line. A good margin is one with a greater gap between the classes.

K-nearest neighbours:

One well-liked SL technique that may be applied to both regression and classification is KNN. KNN attempts to predict the correct class for the test data by computing the distance between the test data and each training point. Based on learning by analogy, nearest-neighbour classifiers compare a given test tuple with training tuples that are similar to it. There are p number of attributes are used to describe the training tuples. A point in an p -dimensional space is represented by each tuple. This creates an p -dimensional pattern space in which to store all of the training tuples. A k -nearest-neighbour classifier looks for the k training tuples that are most similar to an unknown tuple when it is given an unknown tuple. The k 'nearest neighbours' of the unknown tuple are these k training tuples. The concept of 'closeness' is described in terms of a distance metric, such as the Euclidean distance, Manhattan distance (for continuous distance), or Hamming distance (for categorical distance).

Selection of K value:

There are no established statistical techniques for determining the most advantageous value of K . Deciding on a low value of K results in unclear decision boundaries. The smoother the decision boundaries are, the better for classification the larger K value is. Make a plot between the error rate and K , which stands for values within a specified range. Then select K as the value with the lowest error rate.

3.6.3 Ensemble methods:

An ensemble is a combined model that uses several machine learning methods for categorization. Each ML algorithm casts a vote, and the ensemble predicts the class label based on the total number of votes. The accuracy of ensemble algorithm is usually higher than that of the individual algorithm. The well-distribution of the data classes is a basic assumption of traditional machine learning algorithm. However, the data in many real-world data domains are class-imbalanced, with the principal class of interest being represented by a small number of tuples. The problem is called as class imbalance. Ensemble approaches can be used to solve these kinds of problems.

Ensemble algorithm includes bagging, boosting, and random forests, for instance. The goal of an ensemble algorithm is to develop a composite classification model M by combining a set of k trained models (or base classifiers), M_1, M_2, \dots, M_k . From the given dataset D we produce k training sets, D_1, D_2, \dots, D_k , and classifier M_i is produced by D_i , where $i=1,2,\dots,k$. The base classifiers cast their votes by each providing a class prediction in response to a new test tuple to categorise. Based on the basic classifiers'

votes, the ensemble returns a class prediction. In general, an ensemble algorithm shows accuracy is higher than that of its base classifiers.

The examination will commence with each classifier 'voting' for the class designation of a new data tuple. Following the aggregation of the ballots, the ensemble algorithm generates a class prediction.

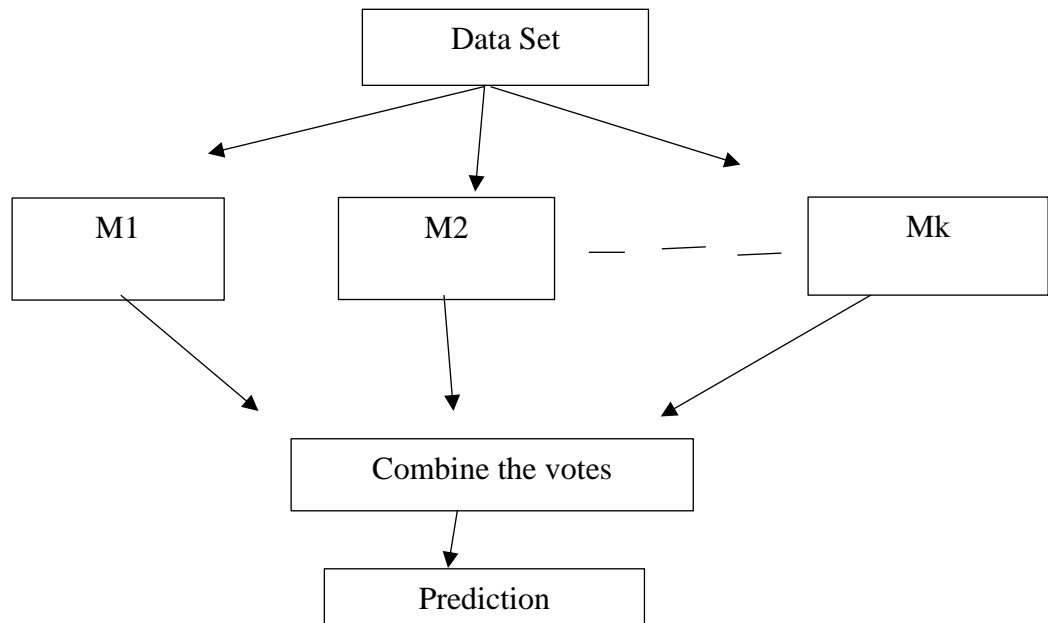


Fig 3.13 Ensemble learning

1. Random Forest:

Within the supervised learning technique, Random Forest is the most often used machine learning algorithm. Random Forest can be used to handle problems related to both regression and classification. The methodology underlying it is known as the ensemble machine learning algorithm, and it is a way to solve a challenging problem by combining multiple algorithms to enhance the overall performance of the algorithm in question.

As the name suggest, 'Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.' The random forest predicts the final result based on the majority of votes from the predictions rather than relying solely on one decision tree. Rather, it takes into account the predictions from each tree. A higher accuracy is achieved as a result of the increased number of trees in the forest, which also helps to prevent the issue of overfitting. The Random Forest algorithm is able to carry out duties related to classification as well as regression. Large datasets with high dimensionality

can be handled by it. In addition to preventing overfitting, it improves the model's accuracy.

Bagged Decision tree:

The accuracy and capacity of machine learning algorithms are increased by the ensemble learning technique such as bagged decision trees, Bootstrap Aggregating. Bagging is a generic ensemble method that can be used to many different base learners. Let's examine Bagged Decision Trees in more detail:

Bootstrap Sampling:

The bootstrap samples used by bagged decision trees include selecting several subsets (with replacement) at random from the initial training dataset. Each subset, often referred to as a bootstrap sample, has the same size as the original dataset but includes some instances that were replicated and leaves out others.

Training Multiple Decision Trees:

By training several decision tree models using the same hyperparameters and splitting criteria as a single DT, bagging produces an ensemble of DTs. One of the bootstrap samples generated in the previous phase is used to train each decision tree.

Parallelization:

Bagging is a good candidate for parallelization because each decision tree can be trained separately. Bagged decision trees are computationally effective.

Predictions:

The final prediction for regression tasks is the average (mean) of predictions from each individual decision tree, whereas for classification problems, the final prediction might be decided by majority vote. Each tree offers a prediction, and the ultimate prediction is made by the class that obtains the most votes among all trees.

Reduction of Variance:

The main aim or purpose of bagging is to reduce the model variance. Multiple decision trees are trained on various subsets of the data, adding diversity to the ensemble and contributing in minimising overfitting. Bagging is particularly useful in making the model more stable and robust since decision trees are prone to large variation (overfitting).

Out-of-Bag (OOB) Error:

One benefit of bagging is that each bootstrap sample typically uses roughly 63.2% of the initial training data. In each repetition, this leaves about 36.8% of the data unused. Without the requirement for a separate validation data set, the performance of the model

can be estimated using this unaltered data. The out of bag (OOB) error is a valid indicator of the model's generalisation abilities.

Hyperparameters:

Hyperparameters like the number of decision trees in the ensemble and the depth of each decision tree can be used to adjust bagging. Up until a certain point of diminishing returns, performance is frequently improved by adding more trees.

Advantages of bagged decision Tree:

Implementing Bagged Decision Trees is straightforward.

They are sturdy and stable due to their ability to effectively reduce overfitting.

They can effectively handle noisy or high-variance data.

Disadvantages of bagged decision Tree:

Bagging lowers variance, however it could not significantly lower bias. Therefore, bagging might not help if the basic decision tree is biased.

Comparing bagging to a single decision tree, the computing cost can increase.

In conclusion, compared to a single decision tree, bagged decision trees are a strong ensemble method that make use of bootstrap sampling and aggregation to build a more reliable and precise model. They can offer a trustworthy approximation of model performance through the out-of-bag error and are especially helpful when working with high-variance models.

2. ADA Boost:

AdaBoost is nothing but Adaptive boosting is a prevalent ensemble learning technique utilized in machine learning to address classification and regression tasks. AdaBoost is a member of the boosting algorithm family, which creates stronger, more accurate learners by combining the predictions of several weak learners (usually basic models). Here is a thorough description of AdaBoost:

Weak Learners:

Multiple weak learners (often decision trees or shallow models) are combined to create AdaBoost. These underperforming learners are also known as 'base classifiers' or 'base models.' Models known as weak learners outperform random guessing on the given problem by a small margin.

Weighted Training Data:

AdaBoost assigns weights to the training data in each iteration. All of the data points are first given equal weights. The method gives more weight to the data points that the

prior base classifier incorrectly classified. This concentrates the attention of the next classifiers on the harder-to-classify samples.

Sequential Training:

Each base classifier in AdaBoost is trained to reduce the weighted classification error in a sequential manner. The algorithm modifies the weights of the training examples after each iteration to give the incorrectly identified samples from the previous iteration greater weight.

Combining Weak Classifiers:

Based on its performance, AdaBoost allocates a weight to each classifier by aggregating the predictions of every base classifier. The weighted total of each base classifier prediction makes up AdaBoost's final prediction. Depending on how accurate a base classifier is, different weights are assigned.

AdaBoost Algorithm:

Here's are steps to running of AdaBoost algorithm:

Step I: Initialize sample weights uniformly for all data points.

Step II: For each iteration (T):

- a. Train a weak model/classifier on the training data with the current weights.
- b. Calculate the weighted-error rate of the weak classifier.
- c. Compute the weight for the weak classifier in the final ensemble.
- d. Update the sample weights to give more importance to misclassified data points.

Step III: Combine the weak classifiers to form a strong classifier.

Strengths of AdaBoost:

AdaBoost is well known for its ability to considerably enhance the performance of underperforming models.

It is adaptive and concentrates on samples that were incorrectly classified, making it resistant to noisy data and outliers.

It is adaptable and works with a variety of base classifiers.

Weaknesses of AdaBoost:

Because AdaBoost tries to fit outliers and noisy data during training, it can be sensitive to these types of data.

If the base classifiers are too complicated or there is not enough data, it can be prone to overfitting.

The performance of AdaBoost may suffer if the weak classifiers are either too weak or too powerful.

Hyperparameters:

The number of base classifiers (iterations), type of weak learner, and learning rate are important hyperparameters to modify in AdaBoost since they affect how much each classifier contributes to the final prediction.

In conclusion, AdaBoost is a robust ensemble learning algorithm that builds a strong classifier by repeatedly combining the predictions of weak classifiers. It can increase classification accuracy and is adaptable, reliable, and efficient, especially when used with weak models. It may, however, be sensitive to outliers and noisy data, and careful hyperparameter tuning is necessary.

Model Evaluation and selection:

3.7 Confusion matrix:

Compared to the case of dichotomous classification, the confusion matrix and its associated parameters are more complicated in the case of a four-class multi-class classification problem. The confusion matrix for a multi-class problem is constructed such that an instance from an actual class is displayed in each row, and an instance from a predicted/estimated class is displayed in each column. The following variables comprise a four-class confusion matrix:

Assuming four classes are labelled A, B, C, and D, the confusion matrix might look like this:

Table 3.1 Confusion matrix for 4 classes

	A	B	C	D
A	TP_A	FP_A	FP_A	FP_A
B	FN_B	TP_B	FP_B	FP_B
C	FN_C	FN_C	TP_C	FP_C
D	FN_D	FN_D	FN_D	TP_D

Here are the parameters associated with a 4-class confusion matrix:

True Positives (TP) for each class:

The True Positives (TP) for each class are calculated by TP_class.

False Positives (FP) for each class:

The formula FP_A represents the number of instances incorrectly classified as class A when they actually belong to other classes (B, C, or D), FP_B represents the number of instances incorrectly classified as class B when they belong to other classes, and FP_C represents the number of instances incorrectly classified as class D.

False Negatives (FN) for each class:

FN_Class represents the count of instances of class that were incorrectly classified as other classes.

True Negatives (TN) for each class: For a multi-class problem, the number of true negatives is not explicitly mentioned in the confusion matrix because it's often not as informative as TP, FP, and FN.

In addition to these parameters, we can also calculate various metrics to assess the performance of the multi-class classification model:

Accuracy: The calculation of accuracy, which quantifies overall correctness, involves dividing the total number of sample points by the sum of the diagonal members of the confusion matrix.

Precision for each class: Precision is determined as TP for the class divided by the product of TP and FP for that class. Precision assesses the accuracy of positive predictions for each class separately.

Recall (Sensitivity) for each class: Recall is calculated as TP for the class divided by the product of TP and FN for that class, and it assesses the model's capacity to identify all relevant instances for each class individually.

F1-Score for each class: The harmonic mean of recall and precision for each class separately is nothing but the F1-Score. In conclusion, the F1-score is a very crucial metric for assessing the effectiveness of a classifier/ ML model, particularly when you wish to take precision and recall into account at the same time. Making educated choices regarding model tuning and optimisation allows you to evaluate how well your model is performing for each class.

These measures allow for a thorough evaluation of a multi-class classification model's performance across all classes and aid in pinpointing areas for development.

CHAPTER 4 PILOT STUDY

4.1 Introduction:

A pilot study is an investigation that is carried out before starting a larger research effort. It entails a scaled-down version of the primary study and seeks to identify potential problems, improve research techniques, and judge the viability of the research plan. This essay clarifies the critical part a pilot study plays in improving the reliability and validity of larger-scale research projects. The IDHS 2015 was used to source the secondary data for this study. The IPUMS-DHS contains thousands of accurately coded variables on the well-being of women, children, pregnant women, males, and all other household members. There were 6,99,686 samples at first. 29,460 samples were extracted for the state of Maharashtra. Only 179 samples were collected for analysis after the data pre-processing was completed. The final data contained a total of 45 independent variables, including menstruation information, anaemia status, individual level factors including age, weight, education, and occupation as well as husband's age and occupation. variables at the community level like region and community education etc. Household level variables such as number of household members, number of children, cooking fuel, toilet facility etc. The first objective of this research is to find prevalence of anaemia. Therefore, in the next section anaemia prevalence in India was find and explained.

4.2. Prevalence of anaemia in India:

Table 4.1 Anaemia distribution in India

No Anaemia	Mild	Moderate	Severe	Grand Total
331619	263130	82490	6950	684189

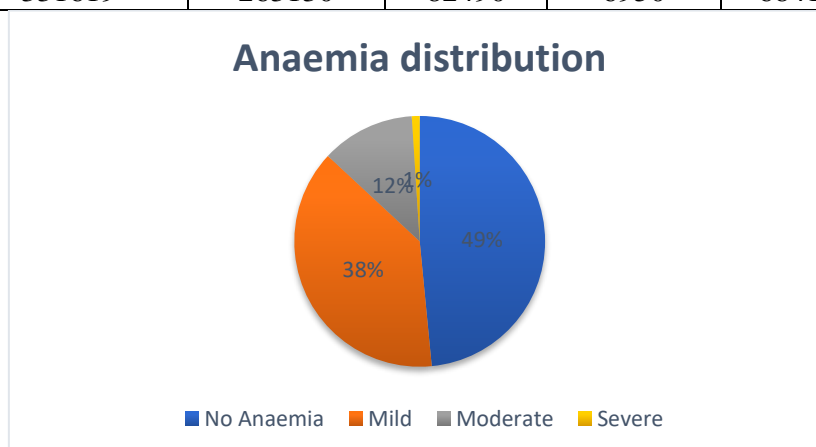


Fig 4.1 Anaemia Prevalence in India

The distribution of anaemia cases in India is shown in the table under the categories ‘No Anaemia,’ ‘Mild,’ ‘Moderate,’ and ‘Severe.’ The majority (331,619) of the 684,189 cases are classified as ‘No Anaemia,’ followed by ‘Mild’ with 263,130 cases, ‘Moderate’ with 82,490 cases, and ‘Severe’ with 6,950 instances as the least common category. But if we consider two categories that is anaemia and no anaemia we can say that approximately 61% WRA found to be anaemic this condition is highly detrimental to their health.

After determining the prevalence of anaemia throughout India, one of the objective of this investigation was to determine its prevalence in Maharashtra. The Maharashtra state prevalence investigation was therefore conducted in the following section.

Prevalence of anaemia in Maharashtra:

```
> idhs_00008 <- read_sav('D:/User Profiles/Staff3/Desktop/PSZ/idhs_00008.sav')
> View(idhs_00008)
> t=table(idhs_00008$GEO_IA1992_2015,idhs_00008$BIOFANAEMIALVL)
> t
```

	0	1	2	3	8
2	7800	7532	3609	384	1481
3	8346	4227	1176	116	429
4	14914	10231	2300	165	837
5	27685	35005	10561	536	1071
10	1153	422	98	13	10
11	10378	10219	3333	361	830
12	8063	9183	3817	291	300
13	4308	3765	1426	183	247
14	11142	8012	3813	439	394
16	13854	8566	2828	258	785
17	7789	3501	563	32	218
19	41196	34440	10217	901	1221
20	15234	10264	2936	214	812
21	10011	2949	534	38	61
22	4037	3389	1356	123	297
23	8745	2771	604	26	133
24	7518	2267	535	63	407
25	2263	1742	568	62	1279
26	15170	14016	3655	250	630
27	2105	1502	364	19	22
28	9291	8342	2201	105	291
29	21439	14804	4822	447	453
30	3347	1433	439	42	32
31	12837	11224	4042	428	289
32	2114	1972	531	33	154
33	55304	43027	13819	1294	1517
35	6296	8327	2343	127	575

Here above table shows the anaemia distribution over all India were first column represents the states of India which are coded. The Maharashtra state was coded by 20. So, extract the anaemia distribution of Maharashtra.

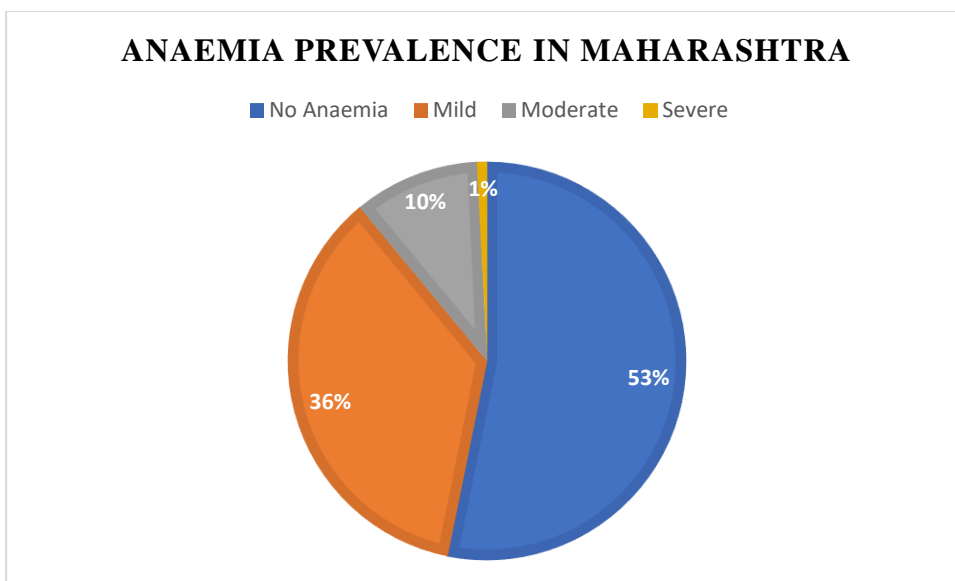


Fig 4.2 Anaemia prevalence in Maharashtra

The statistic means that most members of the group or community under study (53%) do not have anaemia. A considerable fraction (36%) of the population in the group suffers from mild anaemia, a smaller percentage of people in the group have moderate anaemia, and a very small number of people in the group have severe anaemia. The fact that 53% of women do not have anaemia is encouraging, but when comparing the anaemic and non-anaemic groups, 47% of WRA were found to be anaemic, according to DHS data.

Following table shows the state wise percentage of prevalence of anaemia in WRA.

Table 4.2 State wise anaemia prevalence

States	No Anaemia	Mild	Moderate	Severe
Andhra Pradesh, Telangana, Andaman and Nicobar Islands	40%	39%	19%	2%
Arunachal Pradesh	60%	30%	8%	1%
Assam	54%	37%	8%	1%
Bihar and Jharkhand	38%	47%	14%	1%
Goa	68%	25%	6%	1%
Gujarat, Dadra and Nagar Haveli, Daman and Diu	43%	42%	14%	1%
Haryana	38%	43%	18%	1%
Himachal Pradesh	44%	39%	15%	2%
Jammu and Kashmir	48%	34%	16%	2%
Karnataka	54%	34%	11%	1%
Kerala and Lakshadweep	66%	29%	5%	0%
Madhya Pradesh and Chhattisgarh	47%	40%	12%	1%

Maharashtra	53%	36%	10%	1%
Manipur	74%	22%	4%	0%
Meghalaya	45%	38%	15%	1%
Mizoram	72%	23%	5%	0%
Nagaland	72%	22%	5%	1%
Delhi	49%	38%	12%	1%
Odisha	46%	42%	11%	1%
Puducherry	53%	38%	9%	0%
Punjab and Chandigarh	47%	42%	11%	1%
Rajasthan	52%	36%	12%	1%
Sikkim	64%	27%	8%	1%
Tamil Nadu	45%	39%	14%	2%
Tripura	45%	42%	11%	1%
Uttar Pradesh and Uttarakhand	49%	38%	12%	1%
West Bengal	37%	49%	14%	1%

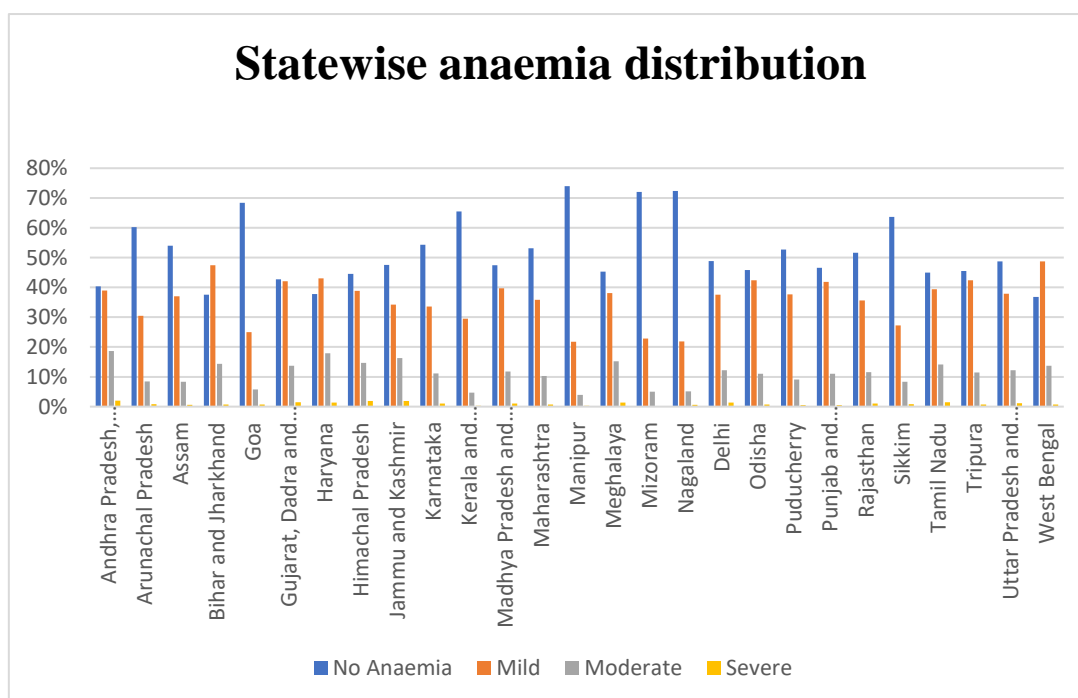


Fig 4.3 State wise anaemia distribution

From the above figure it was found that west Bengal, Bihar and Zarkhand and Haryana states have highest number of anaemic WRA.

There was curiosity to find association between leaving area of WRA (Rural/Urban) and the stage of anaemia. So, in the next section area wise prevalence of anaemia was examined.

4.3 Area wise distribution of anaemia:

```
> table(idhs_00008$URBAN,idhs_00008$BIOFANAEMIALVL)
```

```

0 1 2 3 8
1 101821 72501 21764 1779 6870
2 230518 190631 60726 5171 7905

```

here 1 represents the Urban and 2 represents rural.

Table 4.3 Area-wise anaemia percentage

	No Anaemia	Mild	Moderate	Severe
Urban	51%	37%	11%	1%
Rural	47%	39%	12%	1%

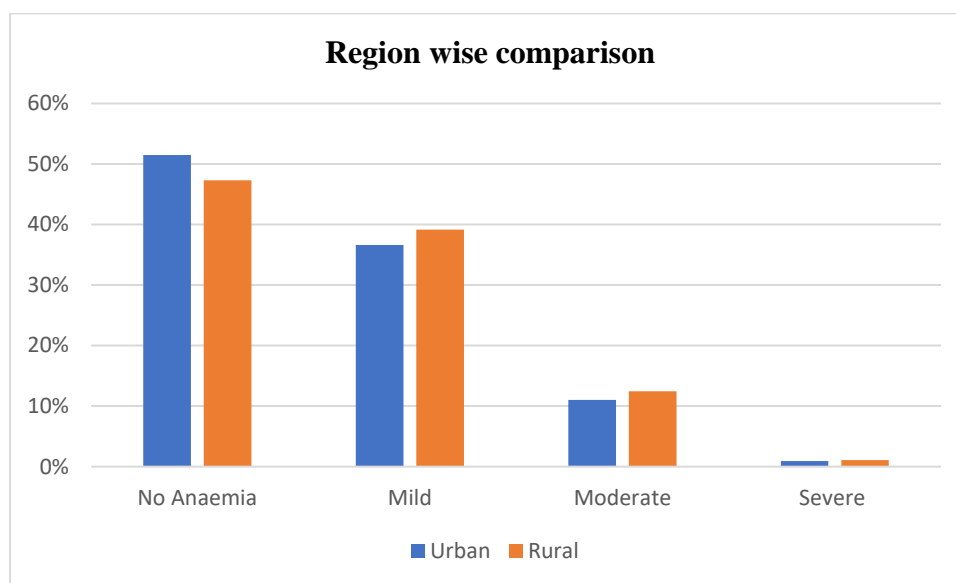


Fig 4.4 anaemia prevalence by Rural-Urban status.

From the above figure here we can see that there was decreasing trend of anaemia severity. If a comparative approach was taken into consideration, the percentage of WRA who were anaemic was found to be higher in rural areas than in urban areas.

Pregnancy is the crucial part of the women's life. In this phase her all body drastically changes. It was interested to find anaemia prevalence according to pregnancy status of WRA. The section 4.4 discovers the anaemia trend or pattern in the WRA according to pregnancy status.

4.4 Anaemia according to pregnancy status:

```
> table(idhs_00008$PREGNANT, idhs_00008$BIOFANAEMIALVL)
```

```

0 1 2 3 8
0 316297 255598 74679 6489 14195
1 16042 7534 7811 461 580

```

Table 4.4 Anaemia distribution according to pregnancy status

	No Anaemia	Mild	Moderate	Severe
Non-pregnant	48%	39%	11%	1%
Pregnant	50%	24%	25%	1%

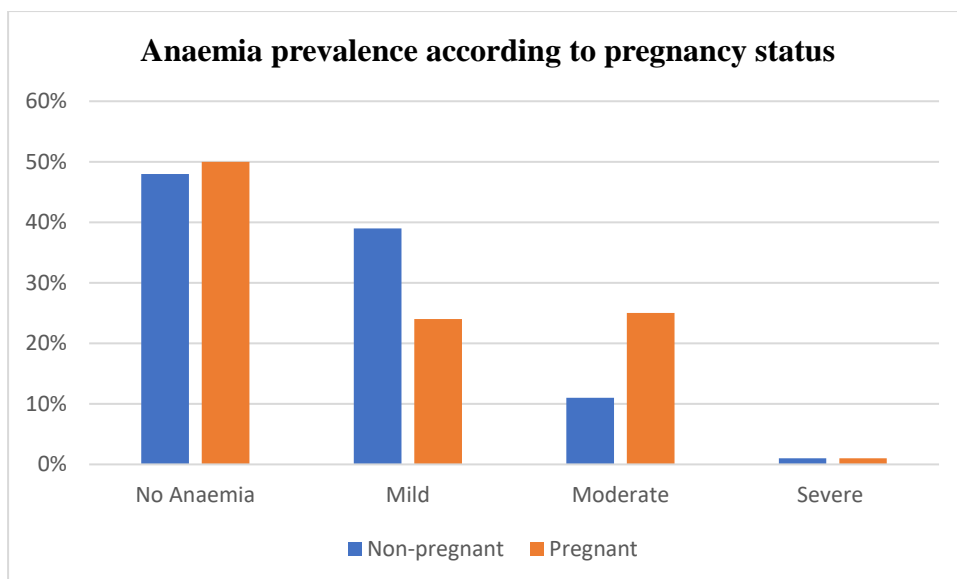


Fig 4.5 Anaemia prevalence according to pregnancy status.

As seen in the above figure, the percentage of pregnant WRAs without anaemia was higher than that of non-pregnant WRAs. However, percentage of mild anaemia was higher in non-pregnant WRA. Percentage of moderate anaemia was higher in pregnant WRA.

Although emotional stability and a sense of security can be obtained from a supportive marriage, there are other aspects of marriage that might impact on a woman's health. Therefore, anaemia prevalence according to marital status was examined in section 4.5.

4.5 Relationship of anaemia and marital status of WRA:

Here DHS variable MARSTAT is categorical variable which shows the marital status of WRA. Where 10 stands for never married, 11 stands for Unconsummated marriage, 21 stands for married, 31 for widowed, 32 for divorced, 33 for Separated/not living together and 34 for deserted. Following table shows the frequency distribution.

```
> table(idhs_00008$MARSTAT, idhs_00008$BIOFANAEMIALVL)
```

	0	1	2	3	8
10	82268	63352	17685	1665	4844
11	880	784	255	23	41
21	236188	188350	60962	4885	9242
31	9273	7827	2589	273	446
32	1615	1027	356	34	80
33	1712	1418	488	53	98
34	403	374	155	17	24

Table 4.5 Marital status with anaemia

	No Anaemia	Mild	Moderate	Severe
Never married	50%	38%	11%	1%
Unconsummated marriage	45%	40%	13%	1%
Married	48%	38%	12%	1%
Widowed	46%	39%	13%	1%
Divorced	53%	34%	12%	1%
Separated/not living together	47%	39%	13%	1%
Deserted	42%	39%	16%	2%

The above table shows the distribution of anaemia was uniform over all categories of marital status. It can be concluded that the marital status does not have any effect on presence of anaemia in reproductive women.

The purpose of the research is to identify risk factors for anaemia in WRA. In the sections that follow, machine learning approaches were used for this.

4.6 Prediction of anaemia with Machine learning techniques.

For the prediction purpose supervised machine learning techniques were used. Since the aim is to predict anaemia which was categorical variable the supervised classification techniques were employed. In the final data there were total 46 variables. Dataset covers a wide range of information, including health-related factors like anaemia, pregnancy status, and various dietary habits, such as consumption of specific food items. Socio-economic indicators like household wealth, education levels, and employment status are also included. Demographic details like age, marital status, and family size are key components. Moreover, the dataset encompasses lifestyle choices like alcohol and tobacco consumption. Decision tree and Random Forest algorithm were developed for the classification of anaemia in the next section. In the following section decision tree and Random forest were developed to predict anaemia.

4.6.1 Decision Tree with 10-fold Cross validation:

A total of 179 rows and 46 columns are included in the dataset. A split of 80 percent to 20 percent is used to divide it into training and testing sets. The training set, also known as train data, is comprised of eighty percent (143) of the data, while the testing set, also known as test data, has the remaining twenty percent (36).

After the 10-fold cross validation the results are as follows:

CART

179 samples

45 predictor

4 classes: 'mild', 'moderate', 'no anaemia', 'severe'

Summary of sample sizes: 161, 161, 162, 161, 161, 161, ...

Resampling results across tuning parameters:

cp	Accuracy	Kappa
0.0930233	0.625817	0.4969145
0.1937985	0.5189542	0.3426154
0.2635659	0.3633987	0.1190476

The cross-validation results suggest that a decision tree with a complexity parameter (cp) of approximately 0.093 provides the best accuracy. Accuracy was used to select the optimal model using the largest value. The final value used for the model was cp = 0.09302326. Therefore, the CART algorithm was redeveloped by setting tuning parameter cp = 0.09302326 and the results are as follows:

The new CART (Decision Tree) algorithm was developed using 143 train sample with Complexity parameter (CP) 0.09302326.

CART algorithm:

```
rpart(formula = Anaemia ~ ., data = train, method = 'class',  
      cp = 0.09302326)  
n= 143
```

	CP	nsplit	rel error	xerror	xstd
1	0.27184466	0	1.0000000	1.0000000	0.05211267
2	0.22330097	1	0.7281553	0.7669903	0.05772957
3	0.09302326	2	0.5048544	0.5048544	0.0558492

Variable importance table:

Table 4.6 Variable importance by CART

Variable	Importance
CURRWORK	22
HUSJOB	12
AGEAT1BIRTH	11
AGEFRSTMAR	11

HUSAGE	11
MARSTAT	11
CHEB	10
AGE	1
BIOFHHAGE	1
EDYRTOTAL	1

In a CART algorithm, variable importance indicates the degree of influence or contribution of each variable towards the anaemia status. Here, the variable ‘CURRWORK’ holds the highest importance with a score of 22, suggesting that current employment status plays a significant role in the status of anaemia in reproductive age women. ‘HUSJOB’ follows with a score of 12, indicating that the husband’s employment status is also relatively influential to anaemia in WRA. Other factors such as ‘AGEAT1BIRTH’, ‘AGEFRSTMAR’, ‘HUSAGE’, and ‘MARSTAT’ have similar importance scores of 11, shows their comparable impact on the anaemia severity in WRA. These variables likely pertain to factors like age at first childbirth, age at first marriage, husband’s age, and marital status, signifying their collective significance in the classification of anaemia in women at reproductive age. ‘CHEB’ stands out as well with a score of 10, suggesting its noteworthy influence. Meanwhile, variables such as ‘AGE’, ‘BIOFHHAGE’, and ‘EDYRTOTAL’ seem to have minimal individual importance scores of 1, implying a relatively lower impact on the prediction of anaemia in WRA. In general, the variable importance scores aid in the order and understanding of the most influential components that influence the research conclusion, offering valuable information about which aspects should be prioritised or taken into account for future research or decision-making.

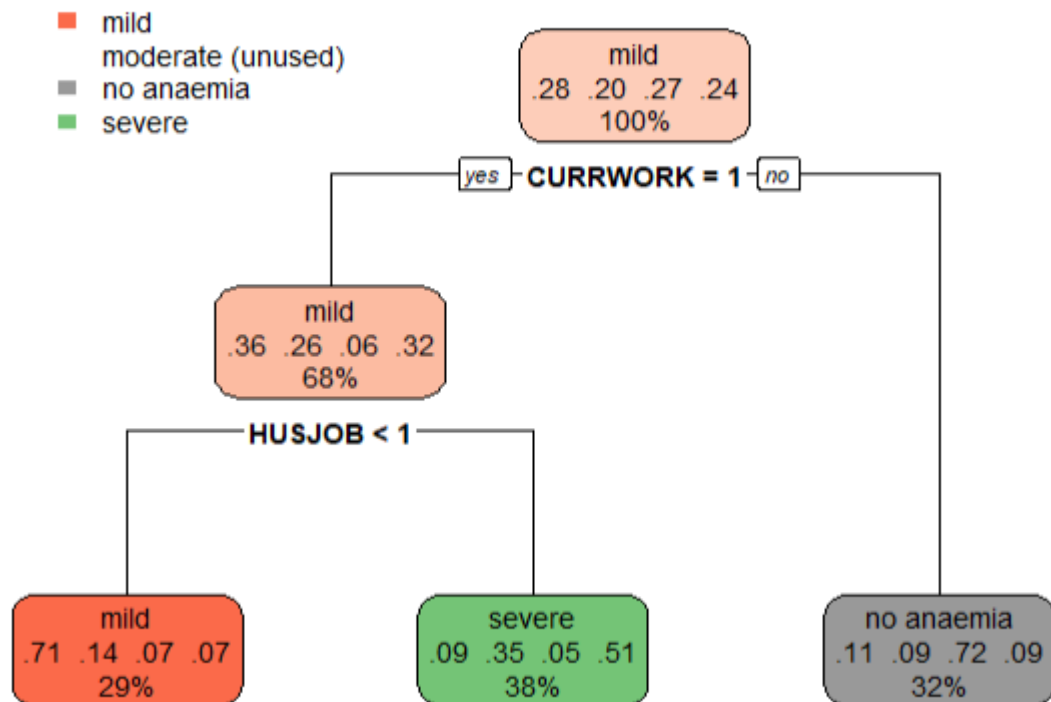


Fig. 4.6 Decision tree plot for pilot study.

From the above plot we can see that the Husbands job and working status of WRA is most important factors.

It is crucial to assess machine learning models to make sure they are trustworthy, effective, and broadly applicable. By evaluating these model's performance, we can determine how well they will function with new data, giving us a better understanding of their predictive power. Understanding if the algorithms have successfully learned patterns from the provided data or whether they may have overfitted is made easier with the help of model evaluation. The model's performance is assessed using assessment parameters such as accuracy, recall, precision and F1 score, which also identify areas that require improvement. Therefore in the next section confusion matrix with various evaluation measures were examined.

Confusion Matrix:

This Confusion matrix table represents the model's performance on the test data. It shows the predicted classes (Mild, Moderate, No, Severe) against the actual classes in the test dataset.

Table 4.7 confusion matrix by CART

		Predicted class			
		Mild	Moderate	No	Severe
Actual class	Mild	6	0	2	2
	Moderate	3	0	1	7
	No	2	0	8	0
	Severe	1	0	0	4

The confusion matrix reveals the performance of the CART algorithm in classifying different stages of ‘Anaemia.’ The algorithm shows an overall accuracy of approximately 54.5%, indicating that it correctly predicts the Anaemia class for just over half of the cases in the test data. However, there are notable variations in its performance across different Anaemia classes. The algorithm performs well in correctly identifying cases of ‘No’ Anaemia, achieving 100% true positives, but it struggles with the ‘Moderate’ class, failing to predict any instances correctly. The misclassification rate, which is around 45.5%, underscores the algorithm’s limitations in providing precise classifications, especially for the ‘Moderate’ category. This suggests that while the algorithm has a moderate overall performance, there is room for improvement, particularly in accurately classifying the ‘Moderate’ Anaemia cases. Further evaluation metrics such as precision, recall, and an understanding of the specific clinical or practical implications are essential for a more comprehensive assessment of the model’s suitability for the task at hand.

```
> # Confusion matrix
> confusion_matrix <- matrix(c(6, 0, 2, 2,
+                               3, 0, 1, 7,
+                               2, 0, 8, 0,
+                               1, 0, 0, 4),
+                               nrow = 4, byrow = TRUE)
> # Function to calculate evaluation measures
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FN <- rowSums(cm) - TP
+   FP <- colSums(cm) - TP
+ }
```

```

+ precision <- TP / (TP + FP)
+ recall <- TP / (TP + FN)
+ f1_score <- 2 * precision * recall / (precision + recall)
+
+ metrics <- data.frame(Class =0:(nrow(cm) - 1), Precision = precision, Recall = recall,
F1_Score = f1_score)
+ return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
  Class Precision Recall F1_Score
1    0 0.5000000    0.6 0.5454545
2    1    NaN    0.0    NaN
3    2 0.7272727    0.8 0.7619048
4    3 0.3076923    0.8 0.4444444

```

Interpretation of Decision tree:

For class 1 that is Mild, the precision is 0 which indicates that among the items predicted as class mild, none were correctly predicted, a recall of 0.5, and an F1-score of 0.5454545, which is the harmonic mean of precision and recall. Class 2 which is moderate anaemia shows a precision of NaN, indicates that undefined value likely due to zero true positives, and a recall of 0.0, signifying that none of the true class 2 instances were correctly predicted. Consequently, the F1-score is also undefined as it involves division by zero. Class 3 which is No anaemia demonstrates a higher performance with a precision of 0.7272727 which was the indication of nearly 72.7% of the predicted class 3 instances were correct, a recall of 0.8 highlighting that 80% of the actual class 3 instances were captured, and an F1-score of 0.7619048, suggesting a reasonably balanced performance between precision and recall. Finally, for class 4 which is severe anaemia, the precision is 0.3076923 meaning around 30.8% of the predicted class 4 instances were correct, a recall of 0.8 indicating that 80% of the actual class 4 instances were identified, and an F1-score of 0.4444444, showing a moderate balance between precision and recall but relatively lower performance compared to class 3. Overall, this examination shows both the strengths and weaknesses of the model's ability to accurately categorise each class, as well as variations in performance

between various classes. Classes 3 and 1 appear to have greater recall and precision scores, respectively, while classes 2 and 4 appear to have noticeably lower precision scores. This study identifies potential areas for model refinement or additional research, particularly for lower-scoring classes, in order to improve the model’s prediction ability for all classes.

To overcome this problem the ensemble algorithms may be helpful. Therefore, to examine Anaemia in the WRA the ensemble algorithm Random forest was developed on same data. The results are as follows:

4.6.2 Random Forest:

The RF algorithm is constructed using the randomForest function. The formula specifies the prediction of ‘Anaemia’ based on all predictor variables in the training dataset. It uses a forest of 100 trees (ntree = 100), and at each split, it tries six randomly selected predictor variables. The results are as follows:

Call:

```
randomForest(formula = Anaemia ~ ., data = train, ntree = 100)
```

Type of random forest: classification

Number of trees: 100

No. of variables tried at each split: 6

OOB estimate of error rate: 42.66%

Confusion matrix:

	mild	moderate	no anaemia	severe	class.error
mild	26	5	7	2	0.3500000
moderate	11	4	2	12	0.8620690
no anaemia	5	0	31	3	0.2051282
severe	3	9	2	21	0.4000000

Variable Importance Table:

Table 4.8 Variable importance by Random Forest

Variable	Importance
CURRWORK	10.66916853
HUSJOB	9.155
WEALTHS	6.3937
RESIDEINTYR	6.2304

AGEAT1STBIRTH	5.01851152
BIOFHHAGE	4.9967
HHMEMTOTAL	4.69571514
EDYRTOTAL	4.3483
AGE	4.3172

Table 4.7 shows variable importance by Random Forest classifier were the first column represents variables and second column represents importance score.

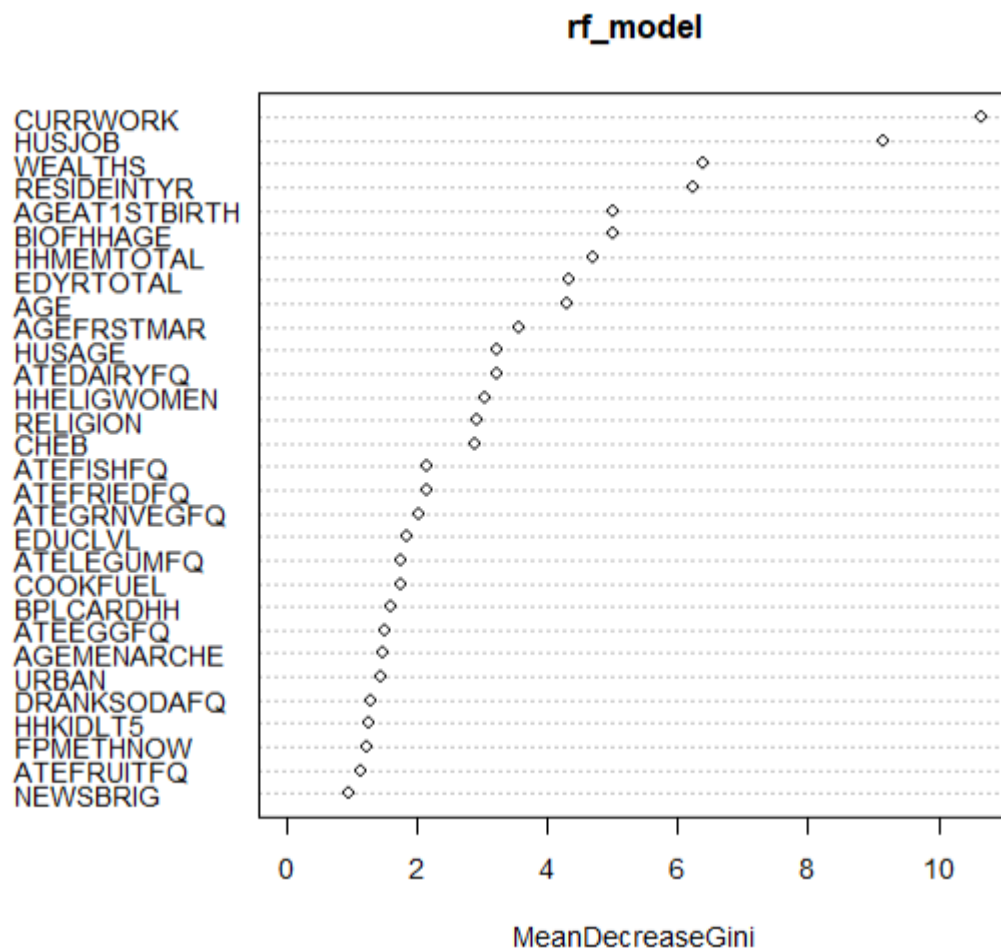


Fig 4.7 Variable Importance plot of Random Forest

The variable importance table highlights the key predictors influencing the Random Forest model's predictions for anaemia severity. Notably, 'CURRWORK' (Current Work) emerges as the most influential variable, indicating that an individual's current employment status significantly impacts anaemia predictions. 'HUSJOB' (Husband's Job) follows closely, suggesting that the occupation of an individual's spouse plays a pivotal role in determining anaemia outcomes, likely reflecting

socioeconomic dynamics within households. ‘WEALTHS’ (Wealth Status) emphasizes that economic well-being is a critical factor, while ‘RESIDEINTYR’ (Residency Years) implies that the duration of residency is relevant. ‘AGEAT1STBIRTH’ (Age at First Birth) underscores the significance of maternal age, and ‘BIOFHHAGE’ (Age of Female Household Head) suggests the potential role of household leadership. Additionally, factors like household size (‘HHMEMTOTAL’), education (‘EDYRTOTAL’), and age (‘AGE’) are deemed important in predicting anaemia. This comprehensive view of influential variables underlines the multifaceted nature of anaemia, influenced by socioeconomic, demographic, and health-related factors. Understanding these critical predictors can guide targeted interventions and public health strategies to address anaemia more effectively.

Confusion Matrix for Random Forest:

The confusion matrix for the Random Forest model’s predictions provides insights into its performance in classifying the severity of anaemia.

Table 4.9 Confusion Matrix for Random Forest

		Predicted class			
		Mild	Moderate	No	Severe
Actual class	Mild	7	1	0	2
	Moderate	3	3	0	5
	No	1	0	9	0
	Severe	1	0	0	4

The overall accuracy, calculated as approximately 64%, suggests that the algorithm correctly predicts the anaemia classes for about two-thirds of the cases in the test dataset. While the model performs well for some anaemia severity levels, it requires further improvement, particularly in classifying the ‘Moderate’ cases. Since the data is only 179 so this results may change when a sample is sufficiently large. The accuracy results will be increase if we take large samples.

As an accuracy of random forest greater than that of decision tree therefore significant factors for the status of anaemia in reproductive age women were identified according to Random forest algorithm. In above paragraph influential factors were identified but trend or pattern of these factors is the also important to assess. Therefore, in the next section relationship of anaemia with these significant factors were examined deeply.

4.7 Relationship of significant factors with anaemia:

CURRWORK i.e. Currently working status of WRA found to be comparatively most important factor according to RF algorithm. CURRWORK is the variable which has two categories 0 and 1. 0 stands for house wife and 1 for working women.

Table 4.10 Anaemia with working status of WRA

		Anaemia status			
		no anaemia	mild	moderate	severe
CURRWORK	0	72%	12%	9%	7%
	1	7%	35%	29%	30%

It is evident from the above table that, in comparison to housewives, working WRA appear to have a higher prevalence of anaemia across all severity levels. Working WRA have the highest percentage of anaemia, with the highest prevalence of severe anaemia. Compared to working WRA, housewives often had lower anaemia counts across all severity levels. The table above indicates a possible correlation between the anaemia status and employment status (housewife vs. working woman), suggesting that there may be a working woman who was at risk of developing anaemia at different severity levels.

The second next factor found was HUSJOB i.e. husband's job. The relationship of woman's anaemia status and her husband's job status may be indirect but influential. Higher socioeconomic status associated with the husband's job may influenced. HUSJOB was a categorical variable which has two categories 0 and 1. 0 stands for occupation of WRA's husband is farmer and 1 stands for WRA's husband is worker or doing a job in company or other.

Table 4.11 Anaemia and Husband's job status

		no anaemia	mild	moderate	severe
HUSJOB	0	58%	42%	0%	0%
	1	4%	14%	38%	44%

The WRA whose husbands were farmers have a 0 percentage of moderate and severe anaemia compared to WRA whose husband's occupation was worker/employee of other organisations. The WRA whose husband's occupation was working have a higher count severe anaemia compared to that of WRA with farmer husbands. The highest total count among the four anaemia statuses is for severe anaemia (44% WRA), indicating a relatively higher prevalence of severe anaemia within this population. Both farmers and working individuals show varying counts across different anaemia levels,

suggesting a potential association between occupation (farmer vs. working individual) and anaemia status.

This table implies a potential link between the occupation of individuals (farmers versus working individuals) and their anaemia status, indicating differing distributions of anaemia levels among these two groups within the sampled population.

Relationship between Anaemia status and wealth index (WEALTHS):

A composite indicator of a household’s overall standard of life is the wealth index. Easy-to-collect information on a household’s possession of certain goods, such as televisions and bicycles, building materials used to construct homes, and kinds of water access and sanitary facilities are used to create the wealth index.

Table 4.12 Anaemia status and average wealth index

	mild	moderate	no anaemia	severe
Average of WEALTHS	0.338197	0.3297705	0.4321651	0.5390748

The WRA with mild anaemia have average WEALTHS value 0.338197. For those with moderate anaemia, the average WEALTHS value was 0.3297705. Reproductive age women with no anaemia, the average WEALTHS value was 0.4321651. For those experiencing severe anaemia, the average WEALTHS value found to be 0.5390748. From these averages, it was revealed that a potential trend that is women at reproductive age experiencing severe anaemia tend to have higher average wealth levels compared to those with no anaemia, mild anaemia, or moderate anaemia. The next significant factor was RESIDEINTYR so, the relation with anaemia was examined in next section.

Relationship between anaemia status and average number of residential years lives in area (RESIDEINTYR).

Table 4.13 Anaemia status and average number of residential years.

	no anaemia	mild	moderate	severe
Average of RESIDEINTYR	22.3877551	12.34	11.475	18.75

For women with no anaemia, the average duration of residency was 22.3877551 years. For those with mild anaemia, the average duration of residency is 12.34 years. For WRA with moderate anaemia, the average duration of residency is 11.475 years. For those experiencing severe anaemia, the average duration of residency is 18.75 years. The average length of residency is longest among women who do not have

anaemia. Compared to women with mild or moderate anaemia, those with severe anaemia have an average residence time that is comparatively higher.

Relationship between anaemia status and age at first birth (AGEAT1STBIRTH):

Here all data has mixture of married and unmarried WRA so the AGEAT1STBIRTH was only for marries as well as have child. Therefore, to access the relationship between anaemia status and AGEAT1STBIRTH married WRA was extracted from the original dataset. There were total 122 married WRA but among them some WRA didn't have child so these WRA removed from analysis. Therefore total 103 The following table shows the average AGEAT1STBIRTH according to anaemia severity.

Table 4.14 Age at first birth and anaemia

	no anaemia	mild	moderate	severe
Average of AGEAT1STBIRTH	24.7	23.24	23.307692	21.121212

The average of age at first birth for women without anaemia was found to be 24.7 years. In WRA who were diagnosed with mild anaemia, the average age at the time of their first birth was 23.24 years. The mean age of WRA at time of first birth with mild anaemia is roughly 23.307692 years. The average age of WRA at time of first birth with severe anaemia was 21.121212 years. These findings suggest that there may be a relationship between the age of WRA at the time of her first birth and the severity of anaemia. WRA with more severe anaemia typically have lower average of age when she gives birth to child than those with no anaemia or milder forms of anaemia.

The next significant factor was BIOFHHAGE. The abbreviation BIOFHHAGE represents the 'Biological Age of Household Head.' It displays the head of the household's age based on biological markers or data. The purpose of this variable is to determine the head of the household's age, usually expressed in years so the variable type is numeric.

Table 4.15 Biological Age of Household Head and anaemia severity

	no anaemia	mild	moderate	severe
Average of BIOFHHAGE	32.04081633	31.66	32.875	34.45

This table indicates the average biological age of household heads within different groups categorized by the severity of anaemia. From the above relational table it was found that, WRA with no anaemia have an average age of household head was approximately 32.04 years. Those with mild anaemia have a slightly lower average age

of around 31.66 years. For moderate anaemia, the average age rises to about 32.875 years. Severe anaemia appears to have the highest average age among household heads, approximately 34.45 years. It appears to indicate that, in comparison to WRA in other anaemia groups, WRA with severe anaemia typically have a greater average biological age of the household head.

Next important factor was HHMEMTOTAL. The abbreviation HHMEMTOTAL is ‘Household Members Total.’ It shows the total number of individuals living in a household. It was numeric variable.

Table 4.16 Household Members Total with anaemia

	no anaemia	mild	moderate	severe
Average of HHMEMTOTAL	4.39	5.92	5.525	5.3

From the above table it can be seen that. WRA with no anaemia have household members on an average 4, that of mild have approximately 6 members. Moderate and Severe anaemia have nearly same average household member total (i.e. approximately 5 members). From this data we can make the statement about relationship between anaemia status and information about household member total. The WRA from nuclear family have a chance of no anaemia than that of joint family.

Next significant factor was EDYRTOTAL. EDYRTOTAL represents the total years of education of women.

Table 4.17 Education of women with anaemia

	no anaemia	mild	moderate	severe
Average of EDYRTOTAL	10.20408163	8.78	8.225	7.275

The information table displays the average years of schooling attained by WRA with several degrees of anaemia severity, which are classified as ‘no anaemia,’ ‘mild,’ ‘moderate,’ and ‘severe.’ WRA with no anaemia completed the highest average years of education, around 10.204 years. Those with mild anaemia had a slightly lower average education duration, approximately 8.78 years. For moderate anaemia, the average years of education dropped further to about 8.225 years. Severe anaemia had the lowest average years of education among the groups, around 7.275 years. Here clear downward trend was discovered between anaemia severity and educational years and educational years decreases the anaemia severity increases from mild to severe. This shows importance of women education in women health.

The next influential factor was age of WRA. Which is numeric variable which is measured in years. The following table displays the average age of WRA according to anaemia.

Table 4.18 Age of WRA and anaemia

	no anaemia	mild	moderate	severe
Average of AGE	32.04081633	31.66	32.875	34.45

WRA with no anaemia have an average age of approximately 32 years. Those with mild anaemia have a slightly lower average age of around 31.66 years. For moderate anaemia, the average age rises to about 32.875 years. Severe anaemia has the highest average age among the groups, approximately 34.45 years. It implies that those with more severe anaemia are often older than people with lesser or no anaemia.

All the significant factors associated with anaemia status were deeply examined. It was discovered that the individual level factors like age, WRA's education, husband's job, age at first birth were found to be significant. As well as household level factors such as wealth index, household member total and age of household head were key indicators of anaemia and demographic level factor like total - number of years of lived in current residential area were found to be significant. Therefore, for further research questionnaire were formed using several factors and mainly these factors. Therefore, in the next chapter primary data analysis was done to predict anaemia in reproductive aged women.

CHAPTER 5
COMPARING THE PERFORMANCE OF MACHINE LEARNING
ALGORITHMS ON UNMARRIED WRA.

5.1 Introduction:

In the previous chapter anaemia prediction and prevalence was done for Maharashtra state on DHS data. From the DHS data results the questionnaire were developed. As discussed in methodology chapter the primary data was consists of mainly there section. First section contains unmarried women, second section contains non-pregnant married WRA and third section included pregnant WRA. In this chapter unmarried WRA studied. The marriage and pregnancy related factors were excluded from this data set, finally data has 39 predictors and 182 WRA. While improving the model performance it was discovered that there is need of additional data so the data was increased from 182 to 203. In this chapter various ML techniques were employed on the data. Trial and error method was used here while developing ML algorithms. At the last best ML model was selected according to the accuracy and the significant factors related to anaemia was extracted from this best ML model.

5.2 Prevalence of Anaemia among unmarried WRA.

```
> table(data$Anaemia)
0  1  2  3
6 119 62 16
```

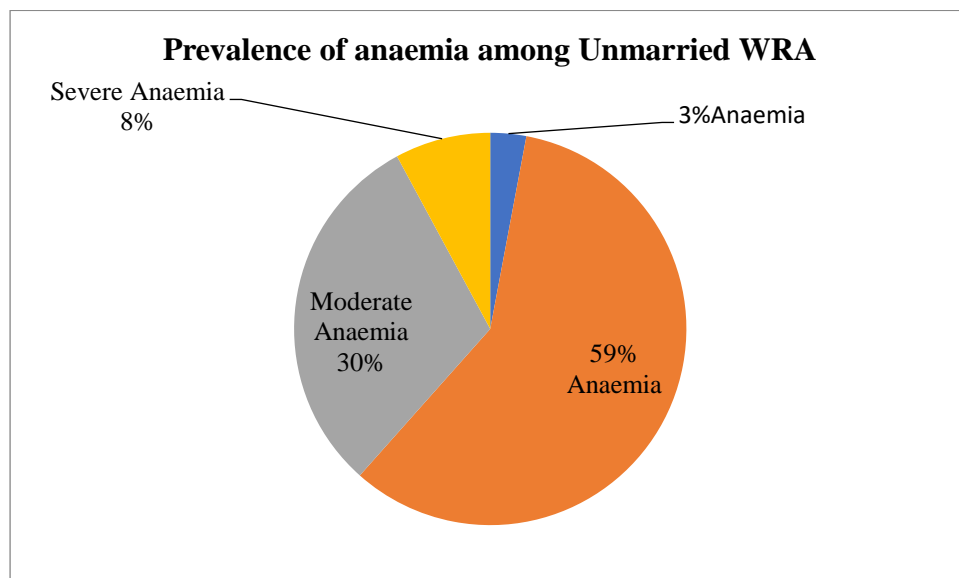


Fig 5.1 Prevalence of anaemia among Unmarried WRA

From the above pie chart it was found that 97% unmarried WRA were found to be anaemic in various categories. There was high prevalence in the mild and moderate

category. In this population severity of anaemia was found to be less but the moderate anaemia prevalence is relatively large. So there is need to implement schemes and programs so we can stop anaemia at mild stage otherwise moderate stage will go to the severe stage and it will become more hazardous.

Further analysis was done by using ML algorithms. In the section 5.2 DT model on whole data was examined and evaluated with the help of confusion matrix. Also the drawbacks of developing model on whole data was explained.

5.3 Decision tree:

The simple to hard procedure applied here so to predict anaemia in unmarried WRA. The decision tree is one of the simplest algorithms among the machine learning algorithms. Therefore, in the next section Decision tree was developed on the whole data.

5.3.1 Decision tree for whole data:

Decision tree model was built on the data set which contains the unmarried WRA. Anaemia taken as a response variable and all other variables in the data set are used as a predictor in the model. Data set used for building the model contains 182 observations. The method used for building the model is class for because Anaemia is categorical variable. The classification tree was built by using CART algorithm.

The CART method is a type of machine learning algorithm that is frequently utilized for classification applications as well as regression tasks. It was developed by Leo Breiman, Jerome Friedman, Richard Olshen and Charlie Stone in 1980. With the use of feature values, the CART algorithm recursively divides the input space into smaller segments to create a binary tree structure. Each internal node of the tree represents a splitting condition on a specific feature while. The leaf nodes contain the predicted output or class label. CART trees are binary trees, meaning each internal node has 2 child nodes corresponding to the binary decision of the splitting condition.

R Part function in R is used to build classification and regression trees using the Recursive partitioning and regression trees algorithm. It is part of the rpart package which is widely used package for tree-based algorithm.

Parameters:

formula: response~predictor1+predictor2+....

data= The data frame containing the variables specified in the formula.

method= It is optional. Which indicates a character string specifying the splitting method. The default is 'ANOVA' for continuous response and 'class' for the categorical response.

Following is the output of the CART algorithm:

```
> # decision tree by r part
> data=data.frame(data)
> m=rpart(Anaemia~.,data= data, method='class')
> summary(m)
```

Call:

```
rpart(formula = Anaemia ~ ., data = data, method = 'class')
```

n= 182

CP	nsplit	Rel error	xerror	xstd
0.06451631	0	1	1	0.1031238
0.02903226	1	0.9354839	1.112903	0.1055690
0.02688172	6	0.7903226	1.322581	0.1082629
0.01612903	9	0.7096774	1.5	0.1087700
0.01000000	12	0.6612903	1.5	0.1087700

Interpretation:

The decision tree model was built on a dataset with 182 observations. The model's performance is evaluated using various parameters, including complexity parameters (CP). The 'CP' represents the complexity parameter, which controls the size and complexity of the decision tree. The output shows a sequence of CP values, corresponding to different levels of tree complexity.

The initial CP value is 0.0645, indicating a full tree with no pruning (0 splits). At this point, the tree perfectly fits the training data, resulting in a relative error and cross-validation error of 1.000.

As the CP value decreases, the tree undergoes pruning (splits) to reduce its complexity. The summary table says the number of splits ('nsplit'), the relative error ('Rel error'), the cross-validation error ('xerror'), and the standard deviation of the cross-validation error ('xstd') at each CP level.

The tree prunes itself several times, with each step leading to a simpler tree. As CP decreases, the tree becomes less complex, and the cross-validation error increases. The goal is to find an optimal CP value that balances model complexity and predictive accuracy.

The summary output provides a useful starting point for evaluating the decision tree's performance, assisting in the selection of an appropriate CP value to control the model's complexity. This allows you to strike a balance between underfitting (too simple) and overfitting (too complex) the data. It's essential to assess the model's performance on an independent test dataset and, if needed, fine-tune the CP value to achieve the best predictive performance while avoiding overcomplicated models.

Variable Importance table:

Table 5.1 Variable importance table of CART

Variable name	Number of years lives in residential area	Weight(kg)	HIV status	Number of days of blood flow
Importance	14	14	9	7
Variable name	BMI	Annual family Income	Average rest in day(hour)	Alcohol consumption
Importance	7	5	5	4
Variable name	No. pads per day	Mass media exposure	occupation	Eating hobbits
Importance	4	4	3	3
Variable name	age	No. Family members	Household wealthy status	Age at menstrual cycle begins
Importance	3	3	3	2
Variable name	Feeling week	height	Education	Regular visit to doctors
Importance	2	2	1	1
Variable name	Exposure to domestic violence	Drinking water source	Pain in menstrual cycle	
Importance	1	1	1	

The variable importance rankings provide valuable insights into the factors that significantly influence the status of anaemia. At the top of the list, 'Number of Years Lived in Residential Area' and 'Weight (kg),' both assigned the highest importance scores of 14. This suggests that the duration of residence and an individual's weight are exceptionally influential in predicting the status of anaemia. Following closely, 'HIV Status' holds the third position with an importance score of 9, indicating that a person's HIV status strongly impacts the predictions. 'Number of Days of Blood Flow' and

'BMI (Body Mass Index)' are also notable predictors with importance scores of 7, highlighting their significant roles in classification of stage of anaemia. On the socioeconomic side, 'Annual Family Income' and 'Average Rest in Day (Hours)' share an importance score of 5, underlining their moderate influence. The rankings proceed with variables related to health and lifestyle, such as 'Alcohol Consumption,' 'Number of Pads Per Day,' and 'Mass Media Exposure,' each with an importance score of 4. As the importance scores decrease to 3, 2, and 1, the influence of the respective variables diminishes. In summary, these rankings help identify which variables mostly influenced the status of anaemia. This will allowing us to focus on the most crucial factors when considering anaemia problem in WRA.

Variable importance table only gives the information about which predictors was significantly affects the response variable or the target variable. Once we get the important variable the next objective is to in which criteria the algorithm gives the decision about the target variable. Therefore, for this objective decision tree plot will be helpful. The decision tree plot is as follows.

```
> rpart.plot(m,extra=104)
> rpart.plot(m, type = 4, extra = 104, tweak = 1.6)
```

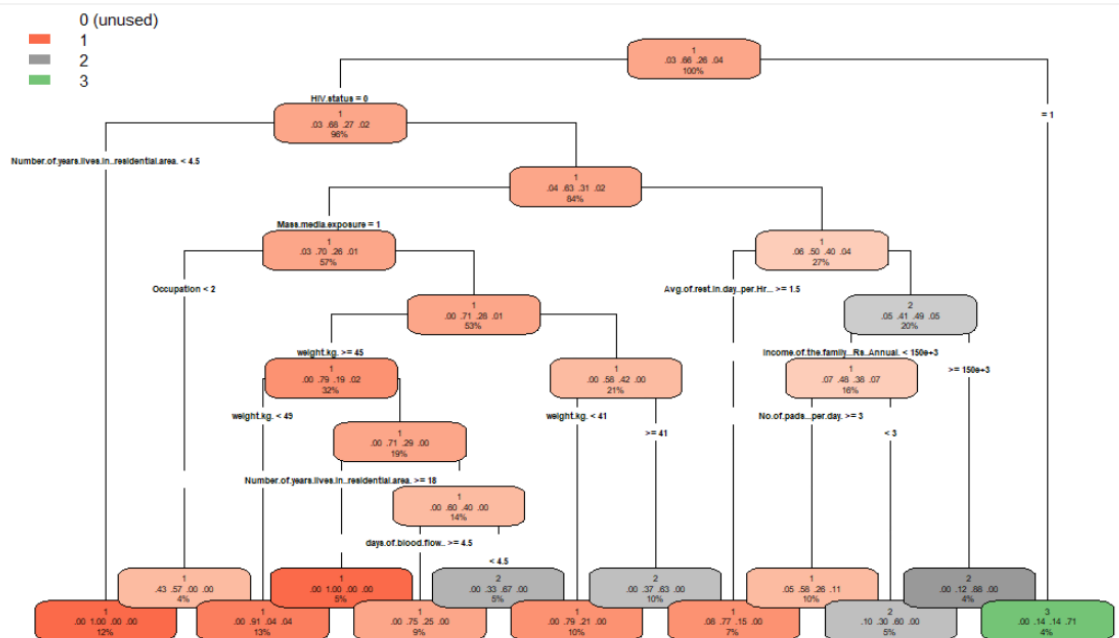


Fig. 5.2 Decision tree plot 1

```
> P=predict(m, newdata = data, type='class')
Levels: 0 1 2 3
> Table=table(data$Anaemia,P)
```



```
> Table
```

```
P
```

```
  0  1  2  3
0  0  5  1  0
1  0 105 14  1
2  0 16 31  1
3  0  3  0  5
```

```
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
```

```
[1] 0.7747253
```

The accuracy of approximately 77.47% suggests that the model's overall performance is quite respectable. But this decision tree algorithm was fitted on the whole data. There are some drawbacks of developing machine learning algorithm without train- test splitting criteria such as:

1. **Overfitting:** Overfitting: we cannot measure the model's generalisation performance without a separate test dataset. It may perfectly fit the training data but fail to generalise to new unknown data.
2. **Lack of Evaluation:** There will be no reliable way to assess the model's performance. Knowing how well your model works on a simulated test dataset is critical for making educated decisions about its applicability and efficacy in real world applications.
3. **Bias:** When training and evaluating a model on the same dataset, it is possible to overestimate the model's performance. This is because the model has previously seen the data on which it is being tested, resulting in biased and inflated performance estimations.
4. **No Insight into Generalisation:** Without a test dataset, you won't be able to determine how effectively the model generalises to new, previously unknown data, which is a key goal of machine learning. You may be unaware of potential complications that may develop when the model interacts with real-world data.
5. **Risk of Model Selection Bias:** Without a test dataset, you may be more prone to model selection bias. You may not effectively compare multiple models or model hyperparameters, resulting in unsatisfactory model selection.
6. **Inability to Optimise Hyperparameters:** Without a test dataset unable to successfully adjust the model's hyperparameters.

7. Loss of Confidence: Making decisions or predictions based on a model without validating its generalisation performance can lead to a loss of confidence in the model's reliability, resulting in costly or incorrect decisions in real-world applications.

To prevent these disadvantages, divide the dataset into two parts: one for training the model and one for testing its performance. In the next section decision tree algorithm was developed with splitting train and test data by using 70-30 pattern.

5.3.2 Decision tree by splitting train and test data:

The same model was fitted on the train data which contains 127 sample points and fitted model was tested on 55 sample observations. The R output is as follows:

```
> m=rpart(Anaemia~.,data= train, method='class')
> m
n= 127
> summary(m)
Call:
rpart(formula = Anaemia ~ ., data = train, method = 'class')
n= 127
```

CP	nsplit	Rel error	xerror	xstd
0.03488372	0	1	1	0.1240234
0.02328581	6	0.744186	1.511628	0.1310033
0.01	7	0.7209302	1.322093	0.1307104

Interpretation

The above model output shows the decision tree model were trained on 27 sample points. As the CP value decreases, the tree undergoes splitting leads to a complex tree. As the CP value decreases from 0.0349 to 0.0233, the tree's complexity increases, and the cross-validation error decreases. The tree becomes even more simplified at a CP of 0.01, with fewer splits and a cross-validation error of 1.3221. In practice, the goal is to select an optimal CP value that balances the trade-off between model complexity and predictive accuracy. Smaller CP values lead to more complex trees, potentially overfitting the data, while larger CP values result in simpler trees that may underfit.

Variable Importance Table:

Table 5.2 Significant variables by updated CART

Variable name	Number of years lives in residential area	Age	Pain in Menstrual cycle	HIV status
Importance	27	14	10	8
Variable name	Annual family Income	No. Family members	Eating hobbies	Age at menstrual cycle begins
Importance	6	5	5	3
Variable name	height	Alcohol Consumption	Household wealthy status	Acidity Problem
Importance	3	3	3	2
Variable name	Exposure to domestic violence	Use of Iron Supplementation	Education	Cooking Fuel
Importance	2	2	1	1
Variable name	Average Rest in day (Hour)	BMI	No. pads per day	Days of blood flow
Importance	1	1	1	1
Variable name	Regularity of menstrual cycle			
Importance	1			

Interpretation

With an importance score of 27, ‘Number of Years Lived in Residential Area’ was found to be the most significant factor, showing that the residence is crucial in influencing the status of anaemia ‘Age’ was the second most important variable, with a score of 14, emphasising its relevance in the the prediction of anaemia. ‘Pain in Menstrual Cycle’ and ‘HIV Status’ are both important, with significance scores of 10 and 8, indicating that these variables have a considerable impact on the status of anaemia according to this decision tree algorithm. Moving on, ‘Annual Family Income,’ ‘Number of Family Members,’ ‘Eating Habits,’ and ‘Age at Menstrual Cycle Begins’ all play reasonably significant role, with importance scores ranging from 6 to 3. These factors have a moderate influence on the status of anaemia. Further down the list, it was discovering that ‘Height’, ‘Alcohol Consumption’, ‘Household Wealthy Status’ and ‘Acidity Problem’ all of which has mild influence on the status of anaemia

according to this developed decision tree algorithm. 'Exposure to Domestic Violence' and 'Use of Iron Supplementation' also have mild significance on response variable.

The decision tree plot was draw for the same model as follows:

```
> rpart.plot(m,extra=104)
> rpart.plot(m, type = 4, extra = 104, tweak = 1.6)
> rpart.plot(m,extra=104)
```

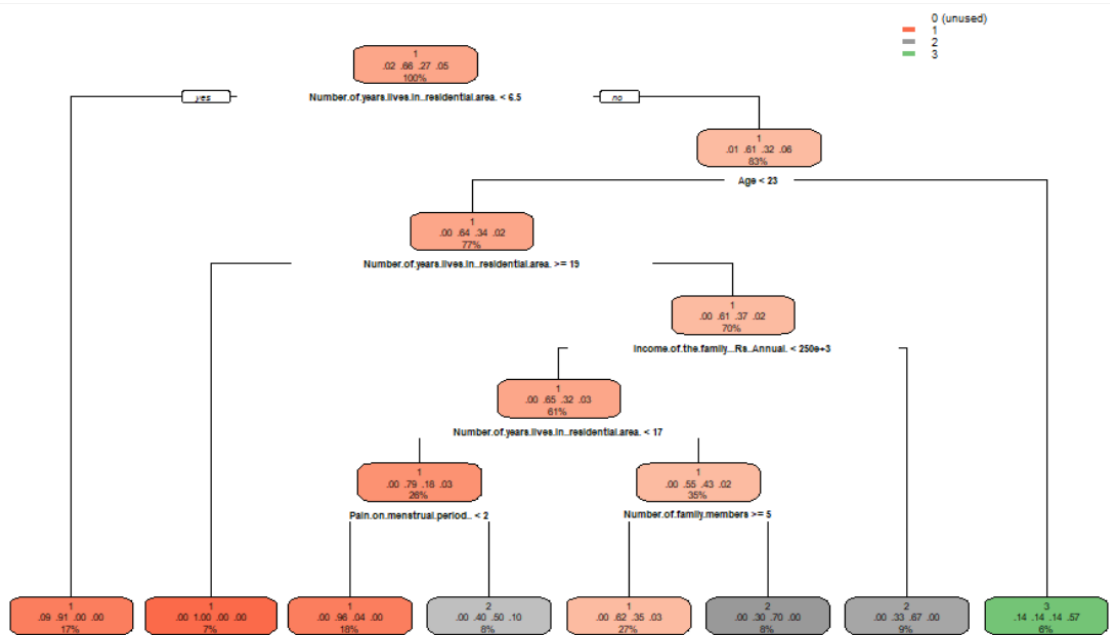


Fig 5.3 Decision tree by updated CART

The above decision tree outlines a series of conditions and the corresponding likelihood of different degrees of anaemia in WRA based on the available train data.

First, if a WRA has lived in a residential area for less than 6.5 years, there's a 17% chance of her experiencing mild anaemia. However, if the no. of years lived in the residential area exceeds 6.5 years, we consider her age as the next factor. If her age is over 23 years, the probability of severe anaemia is 6%. For WRA under the age of 23, another condition comes into play. If they've resided in their residential area for 19 years or more, the likelihood of mild anaemia is 7%. However, resided in their area more than 19 years then moving on to the family income, if it less than 250,000 Rs then there's a 9% chance of moderate anaemia. In cases where the family income is less than 250,000 Rs, and the total years lived in the residential area is less than 17 years, and menstrual pain lasts more than 2 days, the probability of moderate anaemia is 8%. However, if the menstrual pain is less than 2 days, the likelihood shifts to 18% for mild anaemia.

Lastly, when the family income is below 250,000 Rs and the total years lived in the residential area exceeds 17 years, the number of family members becomes a factor. If there are more than 5 family members, the chance of mild anaemia is 27%, otherwise, the likelihood drops to 8% for moderate anaemia. This decision tree provides valuable insights into the factors influencing anaemia among WRA in the dataset, helping to better understand and address this health issue.

Now move forward to examine model performance by using accuracy with the help of confusion matrix:

```
> P=predict(m, newdata = test, type='class')
> Table=table(test$Anaemia,P)
> Table
  P
  0 1 2 3
0 0 2 1 0
1 0 23 10 3
2 0 11 3 0
3 0 0 0 2
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.5090909
```

Interpretation:

Over fitting is a term used to describe the variance in accuracy between the two scenarios that were discussed previously. When a machine learning model performs good on train data but poorly generalises to new data, this is known as over fitting.

In the example above, the decision tree model had an accuracy of 77% after being trained on the whole dataset of 182 samples. This implies that the model was able to identify the relates to and patterns in the data that enabled it to make precise predictions. The accuracy decreased to 50% when the same model was trained on a smaller subset of data, particularly 127 samples. This difference suggests that the model might have improperly suited the training set of data. In other words, rather of learning generalizable patterns that can be applied to new, unseen data, it became overly specialised and suited to the unique features of the 127 examples.

A model that is over fit can perform poorly on fresh data because it effectively memorises the training data, including any noise or outliers. When the train data is sparse or not representative of the full population, this is very likely to occur.

We can use a number of tactics to solve this problem:

- 1) Increasing the range of training data: model can learn more generalizable patterns and decrease over fitting by receiving more diverse and representative data.
- 2) Feature engineering: The generalizability of the model can be increased by carefully choosing and engineering pertinent features. This entails locating informative aspects, eliminating those that are unnecessary, and developing fresh features that could boost the prediction potential.
- 3) Pruning or placing restrictions on the decision tree's complexity are examples of regularisation strategies that can be used to prevent over fitting. By adding penalties to the learning process, regularisation prevents the model from imitating the training data too closely.
- 4) Cross-validation: By assessing the model on many splits of the data, approaches like k-fold cross-validation can give a more reliable assessment of the model's performance. By doing so, it is possible to evaluate how well the model generalises to new data.
- 5) Implementing ensemble methods: Random forests and boosting algorithms are two forms of ensemble methods that combine numerous models to provide predictions. By averaging numerous models' predictions or changing the weights given to various models, these techniques can aid in reducing overfitting.

To overcome this problem the 10-fold cross validation technique was used. The result is as follows:

```
> #Using k-fold cross-validation to train and test the model
> library(tidyverse)
> library(caret)
> data=data.frame(data)
> data=na.omit(data)
> set.seed(123)
> trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
> mc1=train(Anaemia~., data=data, method='rpart',trControl=trctrl)
> mc1
CART
182 samples
39 predictor
4 classes: '0', '1', '2', '3'
```

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 163, 164, 163, 164, 165, 164, ...

Resampling results across tuning parameters:

cp	Accuracy	Kappa
0.02688172	0.6224286	0.04139727
0.02903226	0.6224286	0.04139727
0.06451613	0.6604747	0.00000000

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was $cp = 0.06451613$.

Interpretation:

The CART algorithm was applied to the dataset. The dataset contained 182 observations or data points with 39 predictors. The data was separated into 10 subgroups or folds for the application of the fold cross validation technique. These folds were used to train and test the CART algorithm and assess its performance. According to a summary of sample sizes, 163, 164, 163, 164, 165, 164, ... it displays the number of samples in each fold. The sample sizes appear to be roughly evenly distributed among the folds. The accuracy was utilised as a performance metric to choose the best model. The model with the highest accuracy value was deemed to be the most effective one. $cp = 0.06451613$ served as the model's final value. After comparing several 'cp' values, the model with a 'cp' value of 0.06451613 was chosen as the best model. Lower values typically lead to more complicated trees, and this value decides how complex the tree will be. So, for building decision tree model cp value 0.06451613 was used.

The result is as follows:

```
> m=rpart(Anaemia~.,data= train, method='class',cp=0.06451613.)
> m
n= 127
node), split, n, loss, yval, (yprob)
* denotes terminal node
1) root 127 43 1 (0.02362205 0.66141732 0.26771654 0.04724409) *
> summary(m)
Call:
rpart(formula = Anaemia ~ ., data = train, method = 'class',
cp = 0.06451613)
```

```

n= 127
CP    nsplit  Rel error    xerror  xstd
0.06451613  0    1          0
Node number 1: 127 observations
  predicted class=1  expected loss=0.3385827  P(node) =1
  class counts:    3  84  34  6
  probabilities:  0.024 0.661 0.268 0.047
> rpart.plot(m,extra=104)
> P=predict(m, newdata = test, type='class')
  Levels: 0 1 2 3
  > Table=table(test$Anaemia,P)
  > Table
    P
    0 1 2 3
  0 0 3 0 0
  1 0 36 0 0
  2 0 14 0 0
  3 0 2 0 0
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.6545455

```

Interpretation:

Though model’s accuracy increased but the model appears to only be predicting class 1 for all instances in the dataset, according to the confusion matrix, as there are no predictions for classes 0, 2, or 3. Various factors, such as unbalanced data or a problem with the model training procedure, could be to blame for this. Metrics like accuracy, recall, precision, and F1-score can be used to assess the performance of the model. Since there are no predictions for classes other than 1, it is impossible to appropriately calculate these metrics using the provided confusion matrix.

There are numerous actions can take to fix this issue if model only predicts one class for all instances:

- 1) Verify data. Make sure the dataset is balanced and inclusive of all the classes by carefully reviewing it. Predictions may be skewed by unbalanced data, when one class is much more abundant than the others. If your data are unbalanced, you

might need to use data augmentation techniques, oversample the minority class, or under sample the majority class.

- 2) Model tuning: Try out several hyper parameters or model methods to see which one works best for your dataset. The performance may be enhanced by changing variables like the complexity parameter (cp) in the rpart function or by investigating additional tree-based techniques.
- 3) Consider employing ensemble methods, like random forests or gradient boosting, rather than depending just on one model. These methods combine several models to provide predictions, and they frequently enhance overall performance.
- 4) Consider other models: It can be worthwhile to experiment with different machine learning models or algorithms that are more appropriate for your particular issue. Find out if other classification techniques, such as , support vector machines (SVMs), or Naïve Baye’s Classifier, KNN, etc.

To overcome this problem first additional data were collected to balance the data. Old data have 182 unmarried girls then new data contains 203 unmarries girls. And then the Decision tree algorithm was used to redeveloped. The new training data consist of 142 sample points and testing data set contains 61 sample points.

```
> cat('Training data dimensions:', dim(train_data), '\n')
```

```
Training data dimensions: 142 40
```

```
> cat('Testing data dimensions:', dim(test_data), '\n')
```

```
Testing data dimensions: 61 40
```

Decision tree :

```
> trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
```

```
> mc1=train(Anaemia~., data=data, method='rpart',trControl=trctrl)
```

```
> mc1
```

```
CART
```

```
203 samples
```

```
39 predictor
```

```
4 classes: '0', '1', '2', '3'
```

```
No pre-processing
```

```
Resampling: Cross-Validated (10 fold)
```

```
Summary of sample sizes: 183, 183, 182, 183, 181, 183, ...
```

```
Resampling results across tuning parameters:
```

```
cp      Accuracy  Kappa
```

```
0.02380952 0.5269389 0.129342716
0.03809524 0.5564627 0.088621494
0.10714286 0.5867259 0.008333333
```

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was $cp = 0.1071429$.

Interpretation:

From the above cross validation results the cp value 0.1071429 will be good to fit decision tree by CART. The following decision tree model was developed on new data result are as follows.

```
> # decision tree with cross validation:
> library(rpart)
> m=rpart(Anaemia~.,data= train, method='class',cp = 0.02432084 )
> m
n= 142
> summary(m)
```

Call:

```
rpart(formula = Anaemia ~ ., data = train, method = 'class',
cp = 0.02432084)
n= 142
```

CP	nsplit	rel error	xerror	xstd	
1	0.15873016	0	1.0000000	1.0000000	0.09397214
2	0.05555556	1	0.8412698	0.8412698	0.09148465
3	0.04761905	4	0.6666667	0.9047619	0.09271760
4	0.03703704	5	0.6190476	0.8888889	0.09243970
5	0.02432084	8	0.5079365	0.8730159	0.09214171

Variable importance

HIV.status	weight.kg.
14	13
BMI	Eating.Habits
11	9
days.of.blood.flow..	Age
9	8
Household.Wealth.status..	Alcohol.Consumption..
7	5

Number.of.years.lives.in.residential.area.	Are.you.feeling.weak.or.dizziness.
5	5
Regular.visit.to.doctor..	Income.of.the.family...Rs..Annual.
5	2
Number.of.family.members	Height..meter.
2	2
Age.at..menstrual.cycle.begins	Drinking.water.source..
1	1
Cooking.fuel	
1	

Interpretation:

From the above variable importance table, it was discovered that factors like HIV status, weight of WRA, BMI, eating hobbies, number of days of blood flow, age, Household wealth status, alcohol consumption, years lives in residential area, feeling weak, regular visit to doctors, family income, number of family members, height, age at menstrual cycle begins, drinking water source and cooking fuel.

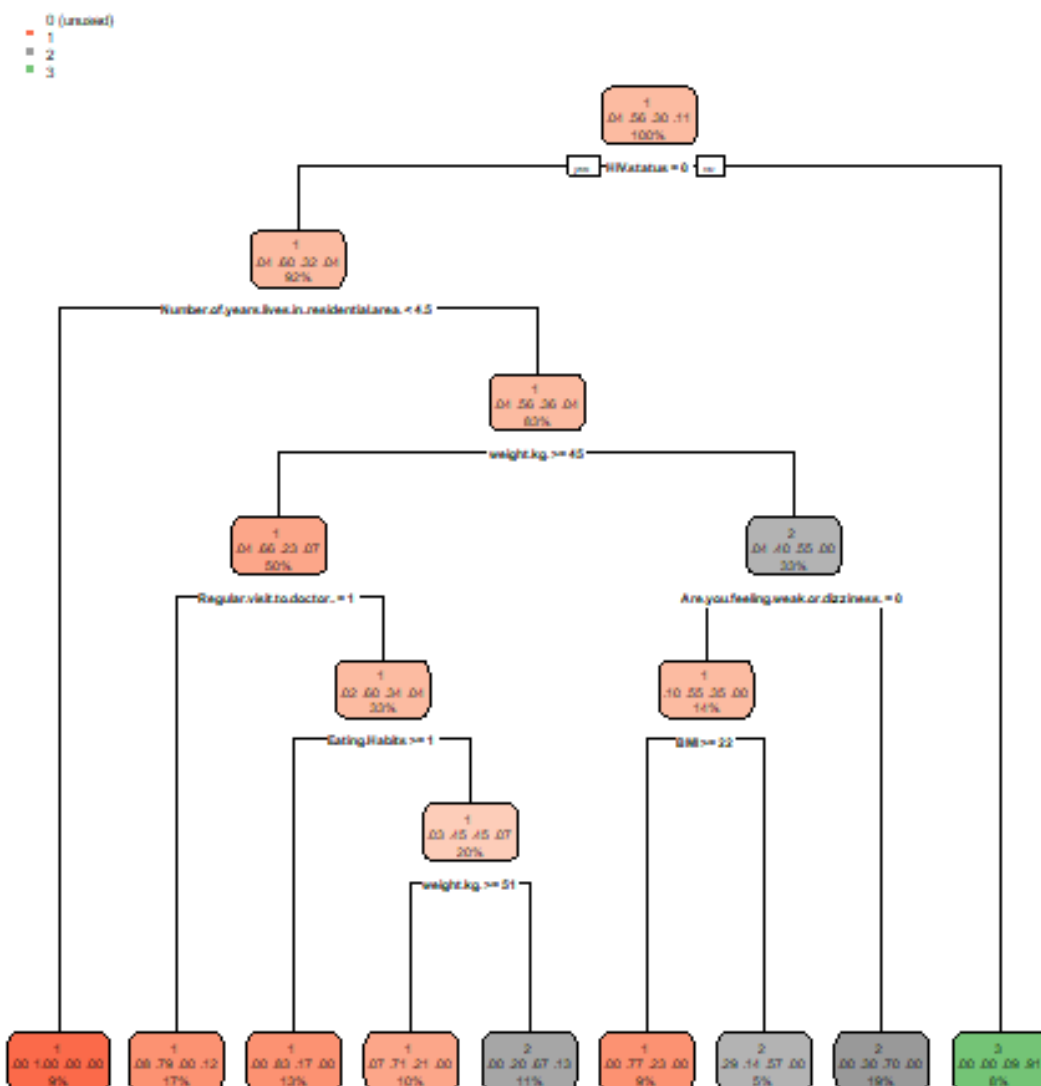


Fig. 5.4 Decision tree plot 3

Interpretation:

If WRA is HIV negative and if years lives in residential area is less than 4.5 years, then there is 9% chance that the WRA is mild anaemic. If woman is HIV negative, years lives in residential area more than 4.5 years. Weight of the women is greater than 45 kg. And if she had regular visit to doctor, then 17% chance of WRA is mild anaemic. If WRA is HIV negative, total years leaves in residential area more than 4.5 yrs. She did not go to regular to doctor, eating habits greater than or equal to 1 then 13% chance that she has mild anaemia. If WRA is HIV negative, no number of years

leaves in residential area more than 4.5 yrs. She didn't go regularly to doctor. Her weight is less than 51KG then 11% chance that she is moderately enemy. However, her weight is greater than 51 then 10% chance that she is mild anaemic. If WRA is HIV negative, years leaves in residential area is greater than 4.5 yrs. Weight is less than 45KG. And if she feels weak or dizziness, then 18% chance that she is moderately anaemic. However, if she not feeling weak or dizziness, but her BMI is less than 22 then 5% chance that she is moderately enemy. Otherwise, 9% chance that she is mild anaemic.

Lets move toward the accuracy of the model.

```
> P=predict(m, newdata = test, type='class')
> Table=table(test$Anaemia,P)
> Table
  P
  0 1 2 3
0 0 1 0 0
1 0 29 10 1
2 0 11 8 0
3 0 1 0 0
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.6065574
```

Interpretation:

We can see here initially, the model achieved a relatively low accuracy on the dataset, indicating that it struggled to make accurate predictions without any adjustments. However, after applying resampling techniques and utilizing cross-validation, the model's accuracy significantly improved to 60.65%. The improved accuracy is a reflection of the model's capacity to provide more accurate predictions and to generalize more effectively when assessed on data that has not been seen before. This improvement underscores the importance of rigorous model validation and optimization techniques, such as cross-validation, in enhancing the overall performance of machine learning models. But we cannot stop here with this 61% accuracy. Therefore, the in the next section Support vector Machine algorithm with various kernels were developed to check performance of the developed model on the same data.

5.3.3 Support Vector Machine:

SVM algorithm with different kernels were developed. Following is the R output

```

> classifier_li = svm(formula = Anaemia ~ ., data = train, type = 'C-
classification', kernel = 'linear')
> classifier_li
Call:
svm(formula = Anaemia ~ ., data = train, type = 'C-classification', kernel = 'linear')
Parameters:
  SVM-Type: C-classification
  SVM-Kernel: linear
    cost: 1
Number of Support Vectors: 105
> summary(classifier_li)
Call:
svm(formula = Anaemia ~ ., data = train, type = 'C-classification', kernel = 'linear')
Parameters:
  SVM-Type: C-classification
  SVM-Kernel: linear
    cost: 1
Number of Support Vectors: 105
( 36 59 7 3 )
Number of Classes: 4
Levels:
0 1 2 3
> P=predict(classifier_li, newdata = test)
> P=predict(classifier_li, newdata = test, type='class')
> T=table(test$Anaemia,P)
> T
  P
  0 1 2 3
0 0 1 2 0
1 0 26 6 1
2 0 12 8 1
3 0 0 0 4
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.6229508

```

Interpretation:

The target variable 'Anaemia' is a categorical variable with multiple classes, so the SVM model is employed for classification (C-classification). The linear kernel denotes that the SVM model divides the classes along a linear decision boundary. It presumes that the data can be separated linearly. The cost parameter (C) in SVM, which is set to 1 in this case, sets the trade-off between the misclassification of training samples and the ease of the decision boundary. A higher cost value emphasises correct classification but may provide a more complex boundary, whereas a lower cost value allows for more misclassifications but may produce a simpler decision boundary. The cost is set to 1 in this instance, indicating a compromise between accuracy and simplicity. The data points nearest to the decision boundary, known as support vectors, are very important in determining the decision boundary. In this instance, there are 105 support vectors, indicating that these 105 data points are the most important in establishing the decision border and class separation.

Overall, the results indicate that a linear SVM model with a cost parameter of 1 was trained. The model's decision boundary is established by 105 support vectors, and it aims to categorise the 'Anaemia' variable with predictive accuracy 62.29%

Support vector machine classifier with polynomial kernel.

The support vector machine by taking polynomial kernel was developed, the R output as follows:

```
> classifier_pol = svm(formula = Anaemia ~ ., data = train_data, type = 'C-  
classification', kernel = 'polynomial')
```

```
> classifier_pol
```

Call:

```
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel =  
'polynomial')
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: polynomial

cost: 1

degree: 3

coef.0: 0

Number of Support Vectors: 141

```
> summary(classifier_pol)
```

Call:

```
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel =  
'polynomial')
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: polynomial

cost: 1

degree: 3

coef.0: 0

Support-Vectors: 141

(41 85 12 3)

Number of Classes: 4

Levels:

0 1 2 3

```
> p=predict(classifier_pol, newdata = test_data, type='class')
```

```
> Ta=table(test$Anaemia,p)
```

```
> Ta
```

p

0 1 2 3

0 0 3 0 0

1 0 33 0 0

2 0 19 1 1

3 0 1 0 3

```
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
```

```
[1] 0.6065574
```

Interpretation:

We may interpret the SVM (Support Vector Machine) analysis based on the code and output from the preceding output. The kernel SVM Polynomial denotes the use of a polynomial kernel for classification. here Cost=1 denotes the SVM model's penalty for incorrect classification. Degree=3, which denotes the polynomial kernel function's degree. The Coef.0=0 indicating that the intercept term is set to 0. The total training data points that were chosen as support vectors, or the number of support vectors, was 141. Following is a distribution of how support vectors are distributed throughout the classes:

41 support vectors in class 0, 85 support vectors in class 1, 12 support vectors in class 2 and 3 support vectors in class 3

Support vector machine classifier with Sigmoid kernel.

```
> classifier_sig = svm(formula = Anaemia ~ ., data = train_data, type = 'C-  
classification', kernel = 'sigmoid')
```

```
> classifier_sig
```

Call:

```
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel =  
'sigmoid')
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: sigmoid

cost: 1

coef.0: 0

Number of Support Vectors: 110

```
> summary(classifier_sig)
```

Call:

```
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel =  
'sigmoid')
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: sigmoid

cost: 1

coef.0: 0

Support-Vectors: 110

(41 56 10 3)

Number of Classes: 4

Levels:

0 1 2 3

```
> p=predict(classifier_sig, newdata = test_data, type='class')
```

```
> Table=table(test$Anaemia,p)
```

```
> Table
```

p

0 1 2 3

```
0 0 3 0 0
1 0 31 2 0
2 0 20 0 1
3 0 1 0 3
```

```
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.557377
```

Interpretation:

From the above SVM (Support Vector Machine) analysis we can see that the kernel Sigmoid used for the classification.. here Cost=1 denotes the SVM model's penalty for incorrect classification. The Coef.0=0 indicating that the intercept term is set to 0. The training data points that were chosen as support vectors, or the number of support vectors, was 110. Following is a distribution of how support vectors are distributed throughout the classes:

41 support vectors in class 0, 56 support vectors in class 1, 10 support vectors in class 2 and 3 support vectors in class 3.

Support Vector Machine with kernel Radial:

```
> classifier_rad = svm(formula = Anaemia ~ .,data = train_data,type = 'C-
classification',kernel = 'radial')
> classifier_rad
```

Call:

```
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel =
'radial')
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost: 1

Number of Support Vectors: 135

```
> summary(classifier_rad)
```

Call:

```
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel =
'radial')
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

```

cost: 1
Number of Support Vectors: 135
( 41 79 12 3 )
Number of Classes: 4
Levels:
0 1 2 3
> p=predict(classifier_rad, newdata = test_data, type='class')
> Table=table(test$Anaemia,p)
> Table
  p
  0 1 2 3
0 0 3 0 0
1 0 31 2 0
2 0 19 1 1
3 0 1 0 3
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.5737705

```

Interpretation:

From the above SVM (Support Vector Machine) analysis we can see that the kernel Radial used for the classification. here Cost=1 denotes the SVM model’s penalty for incorrect classification. The Coef.0=0 indicating that the intercept term is set to 0. The number of training data points that were chosen as support vectors, or the number of support vectors, was 135. Following is a distribution of how support vectors are distributed throughout the classes: 41 support vectors in class 0, 79 support vectors in class 1, 12 support vectors in class 2 and 3 support vectors in class 3.

Comparison of SVM models:

Table 5.3 SVM model’s accuracy comparison

kernel	linear	polynomial	Sigmoid	Radial
Accuracy	62.29%	60.66%	55.74%	57.38%
No. Support vectors	105	141	110	135

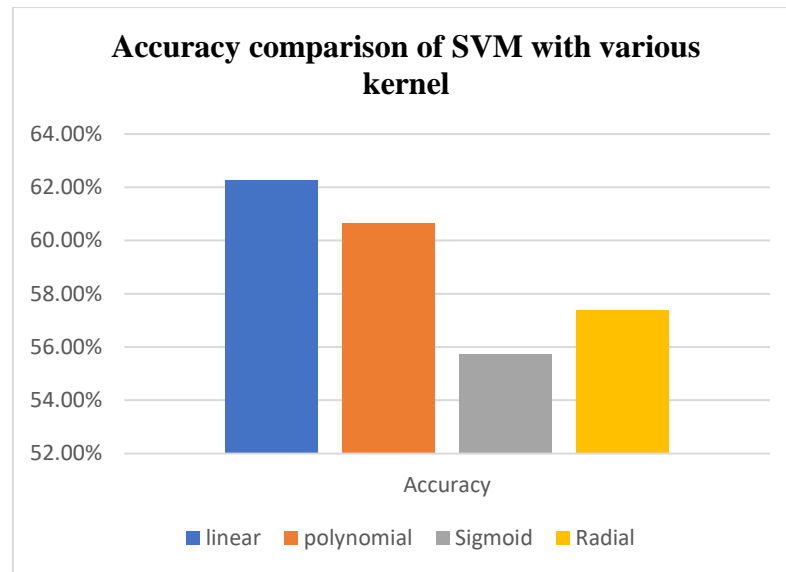


Fig. 5.5 Comparison of SVM model with various kernels

Interpretation:

In terms of accuracy the linear kernel has the highest accuracy of 62.29%, followed closely by the polynomial kernel with an accuracy of 60.66%. The sigmoid and radial kernels have lower accuracies of 57.38% and 55.74% respectively. We can also consider the total support vectors. With 105, the linear kernel has the fewest support vectors, followed by the polynomial kernel with 110. There are more support vectors in the sigmoid and radial kernels, 135 and 141, respectively. Support vectors are the data points in a support vector machine (SVM) or other similar classifier that are closest to the decision border. The formation of the decision boundary is mainly affected by these data points.

A complex decision boundary may be indicated by a large number of support vectors, which could result in overfitting. However, fewer support vectors typically imply a clearer decision boundary that is more adaptable to incoming data. It enhances computing performance and lowers the chance of overfitting. The linear kernel seems to be the most suitable model. Compared to the other kernels, it is the most accurate and has the fewest support vectors, which suggests better generalisation and perhaps faster inference. Looking towards more accuracy in the next section K-nearest Neighbour algorithm was developed on the same data.

5.3.4 K-Nearest Neighbour Algorithm:

Before developing K nearest neighbour algorithm on the data first we have to select the value of k. 10 fold cross validation were used for optimal value of K.. The output as follows:

```
> trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
> mc2=train(Anaemia~., data=data, method='knn',trControl=trctrl)
> mc2
```

k-Nearest Neighbors

203 samples

39 predictor

Summary of sample sizes: 184, 182, 183, 183, 182, 182, ...

Tuning parameters:

k	RMSE	Rsquared	MAE
5	0.6931384	0.07962057	0.5688861
7	0.6478117	0.14406121	0.5396773
9	0.6490390	0.14567025	0.5540523

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was k = 7.

Interpretation:

According to the cross-validation results, the k-NN model with k = 7 exhibits the best performance out of all the evaluated values for k. Compared to the models with k = 5 and k = 9, it has the lowest RMSE, indicating that it makes more accurate predictions. According to the R-squared values, the target variable's variation is only partially explained by the model. Additionally, the MAE values indicate that, average of absolute errors which is also small for k=7.

So, from the above cross validation result the we can develop the KNN model with k=7.

The results are as follows:

K Nearest Neighbour algorithm with k=7:

```
> classifier_knn <- knn(train = train,test = test,cl = train$Anaemia,k = 7)
> classifier_knn
[1] 1 1 3 1 1 1 1 1 1 1 2 1 1 3 1 1 1 1 1 1 2 2 1 1 1 1 1 2 1 1 2 1 2 1 1 1 1 1 2
1 1 1
[48] 1 1 1 1 1 2 3 2 1 2 1 2 1 1
Levels: 0 1 2 3
> summary(classifier_knn)
0 1 2 3
0 47 11 3
> cm <- table(test$Anaemia, classifier_knn)
```

```

> cm
  classifier_knn
  0 1 2 3
1 0 31 5 0
2 0 14 4 1
3 0 2 2 2
> # Calculate Sample error
> misClassError=mean(classifier_knn != test$Anaemia)
> print(paste('Accuracy =', 1-misClassError))
[1] 'Accuracy = 0.60655737704918'

```

Here we can see that the accuracy of KNN algorithm with k=7 is 60.6565%.

Interpretation:

In pursuit of higher accuracy and improved predictive power, here initially experimented with individual machine learning models, including decision trees, SVM, and K-nearest neighbors (KNN), achieving accuracy levels of 65%, 60%, and 50%, respectively. However, recognizing the need for further enhancement, there is need to take a strategic step forward by building ensemble models. Ensemble models combine the strengths of multiple base models to create a more robust and accurate predictive framework. This approach aims to leverage the diversity of individual models to mitigate their weaknesses and boost overall performance. By employing ensemble techniques such as Random Forest, AdaBoost, or Bagged Decision tree, I intend to harness the collective intelligence of these models to achieve higher accuracy and unlock their full potential in tackling the predictive challenges at hand. So, in the next section three ensemble algorithms have been developed to predict status of anaemia in accordance with 39 predictor variables.

5.4 Ensemble techniques:

Ensemble learning is an advanced machine learning technique that combines the predictions of numerous base models in order to increase overall predicted accuracy and robustness. The basic premise behind ensemble learning is that by combining the opinions of numerous models, the collective outcome is often more accurate and dependable than any single model.

5.4.1 Bagged Decision tree:

A Bagged Decision Tree is an ensemble learning technique that combines numerous decision trees to increase prediction precision and prevent overfitting. It is

also known as a Bootstrap Aggregating Decision Tree or simply Bagging with Decision Trees. Bagging is a broad ensemble approach that may be used to a variety of base models; when applied to decision trees, it transforms into a Bagged Decision Tree.

```
> # Convert target variable to factor
> data$Anaemia=as.factor(data$Anaemia)
> train$Anaemia=as.factor(train$Anaemia)
> # Create bagged classification tree model
> bagged.tree <- bagging(Anaemia ~ ., data = train, nbagg = 25, coob = TRUE,
+                       control = rpart.control(maxdepth = 2, minsplit = 1))
> bagged.tree
```

Bagging classification trees with 25 bootstrap replications

```
Call: bagging.data.frame(formula = Anaemia ~ ., data = train, nbagg = 25,
  coob = TRUE, control = rpart.control(maxdepth = 2, minsplit = 1))
```

Out-of-bag estimate of misclassification error: 0.3803

```
> # Create bagged classification tree model
> bagged.tree <- bagging(Anaemia ~ ., data = train, nbagg = 50, coob = TRUE,
+                       control = rpart.control(maxdepth = 2, minsplit = 1))
> bagged.tree
```

Bagging classification trees with 50 bootstrap replications

```
Call: bagging.data.frame(formula = Anaemia ~ ., data = train, nbagg = 50,
  coob = TRUE, control = rpart.control(maxdepth = 2, minsplit = 1))
```

Out-of-bag estimate of misclassification error: 0.3592

```
> # Create bagged classification tree model
> bagged.tree <- bagging(Anaemia ~ ., data = train, nbagg = 100, coob = TRUE,
+                       control = rpart.control(maxdepth = 2, minsplit = 1))
> bagged.tree
```

Bagging classification trees with 100 bootstrap replications

```
Call: bagging.data.frame(formula = Anaemia ~ ., data = train, nbagg = 100,
  coob = TRUE, control = rpart.control(maxdepth = 2, minsplit = 1))
```

Out-of-bag estimate of misclassification error: 0.3521

```
> #calculate variable importance
> VI=data.frame(var=names(train[,-1]), imp=varImp(bagged.tree))
```

Variable Importance:

Sr. No	Variable Name	Importance
1	Suffers from Diabetes	13.56
2	Feeling weak	11.82
3	Acidity problem	11.80
4	addiction	4.73
5	Regular visit to doctor	4.04
6	Suffers from long term disease	3.74
7	Number of family members	3.72
8	Household wealth status	2.80
9	Education	2.15
10	Daily eat fresh vegetables, fruits, milk	1.47
11	Use of iron supplementation	1.34
12	Toilet facility	1.12
13	Community women education	1.02

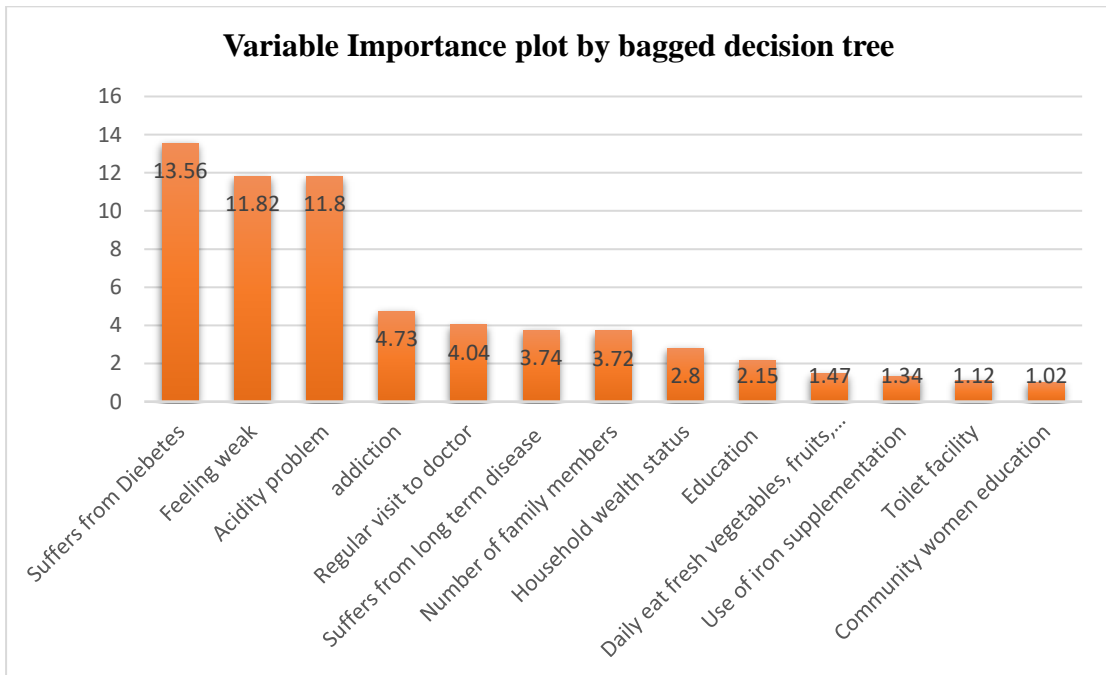


Fig. 5.6 Variable Importance plot by bagged decision tree

Interpretation:

These importance scores reflect the relative influence of each predictor variable on the status of anaemia made by a model. At the top of the list, ‘Suffers from Diabetes’ stands out as the most influential variable with a relatively high importance score of 13.56. This suggests that the presence or absence of diabetes plays a significant role in

determining the model's predictions. Following closely are 'Feeling Weak' and 'Acidity Problem,' each with importance scores of 11.82 and 11.80, respectively. These variables seem to have a substantial impact on the status of anaemia, indicating their relevance to the context under study.

Moving down the list, 'Addiction' and 'Regular Visits to the Doctor' have moderate importance scores of 4.73 and 4.04, respectively, indicating their noteworthy influence on the status of anaemia. 'Suffers from Long-Term Disease,' 'Number of Family Members,' 'Household Wealth Status,' and 'Education' also make a contribution with importance scores ranging from 3.72 to 2.15.

Further down the list, we find 'Daily Consumption of Fresh Vegetables, Fruits, and Milk,' 'Use of Iron Supplementation,' 'Toilet Facility,' and 'Community Women Education,' each with importance scores of 1.47 to 1.02. While these variables have a comparatively lower influence on the model's outcomes, they still play a role in shaping the predictions.

These importance scores provide valuable insights into the predictor variables that are mostly influential in prediction of target variable. The variables at the top of the list are the key drivers of the model's decisions, while those at the bottom, while less impactful, still contribute to the overall understanding and predictive power of the model. The interpretation of these variables is context-dependent and can help guide further analysis and decision-making in the domain to which they pertain.

Model Interpretation:

To develop the bagged classification trees data is spitted into train and test. The response variable Anaemia is a categorical variable that is turned to a factor variable using the `as.factor()` function. With varying numbers of bootstrap replications (`nbagg` parameter), bagged classification tree models are produced using the `bagging()` function. The `coob = TRUE` tells the models to estimate performance using out-of-bag (OOB) samples. The control parameters for the underlying decision tree model are set using the control parameter with the syntax `rpart.control(maxdepth = 2, minsplit = 1)`. In this case, the tree can only grow to a maximum depth of 2, and a split can only be made after at least one observation. The output provides the OOB estimate of misclassification error for each bagged tree model after training.

The rate of incorrect estimations provided by the model on OOB samples is shown by the misclassification error. The model's performance on new data is approximately represented by the OOB estimate.

OOB misclassification error rates for the bagged tree models with 25, 50, and 100 bootstrap replications are about 0.3803, 0.3592, and 0.3521, respectively. The OOB estimate provides a prediction of the models' potential performance on unobserved data or simply test data. Better predictive performance is suggested by lower misclassification error rates. The models often get better at minimising misclassification errors as the number of bootstrap replications rises. Therefore, the model with 100 bagged trees was selected for further analysis.

Achieving an accuracy of 64.79% with a Bagged Decision Tree is a good start, but there's always room for improvement. If you're aiming for higher accuracy, it's a great idea to explore other ensemble techniques. In the next section Random forest algorithm was implemented on the same data.

5.4.2 Random Forest Algorithm:

Random Forest is an extension of Bagging and can give rise to better performance. It introduces additional randomness by selecting a random subset of features at each node split, which might increase the variety of the individual trees. As in the previous section bagged decision tree gives comparatively better results for 100 trees, the RF model was developed on 100 trees.

```
> rf_model <- randomForest(Anaemia ~ ., data = train, ntree = 100)
> # Make predictions on the test set
> predictions <- predict(rf_model, newdata = test)
> rf_model
```

Call:

```
randomForest(formula = Anaemia ~ ., data = train, ntree = 100)
```

Type of random forest: classification

Number of trees: 100

No. of variables tried at each split: 6

OOB estimate of error rate: 35.21%

Confusion matrix:

```
0 1 2 3 class.error
0 0 2 0 0 1.0000000
1 0 67 15 0 0.1829268
2 0 30 13 1 0.7045455
3 0 1 1 12 0.1428571
```

```
> Table=table(test$Anaemia,predictions)
```

> Table

predictions

```
0 1 2 3
0 0 3 1 0
1 0 31 5 1
2 0 9 9 0
3 0 0 0 2
```

> Accuracy=sum(diag(Table))/sum(Table);Accuracy

[1] 0.6885246

Random forest algorithm gives highest accuracy which is 68.85%.

Interpretation:

The random forest model was constructed with 100 decision trees, and at each split in each tree, a random subset of 6 variables was considered, which adds an element of randomness and diversity to the model. The out-of-bag (OOB) estimate of the error rate, which measures the model's performance on unseen data points, was found to be 35.21%. This indicates that, on average, the model misclassifies about 35.21% of the data points that were not part of the training process.

Variable Importance:

Table 5.4 Top Important variable by random forest

sr. no.	variable name	importance
1	No. days of blood flow	5.76
2	No. years lives in residential area	5.53
3	BMI	5.05
4	age	4.96
5	weight	4.9
6	family income	4.51
7	HIV status	3.97
8	age at menstrual cycle begins	3.52
9	eating habits	3.5
10	height	3.35
11	number of family members	3.012
12	average rest in day	2.56
13	No. pads per day	2.37
14	cooking fuel	2.35
15	pain in menstrual period	2.35
16	drinking water source	1.57
17	regular visit to doctor	1.57

18	exposure of domestic violence	1.52
19	regularity of menstrual cycle	1.43
20	daily eat fresh vegetables, fruits, milk	1.4
21	acidity problem	1.24
22	household wealth status	1.2
23	mass media exposure	1.18
24	sufferes from stress	1.1044
25	feeling weak	1.1
26	education	0.92
27	menstrual cycle	0.9
28	food type	0.87
29	occupation	0.77
30	daily tea intake	0.66
31	Region	0.62
32	alcohol consumption	0.6
33	suffers from any long term disease	0.48
34	community women education	0.47
35	toilet facility	0.25
36	type of addiction	0.048
37	addiction	0.007

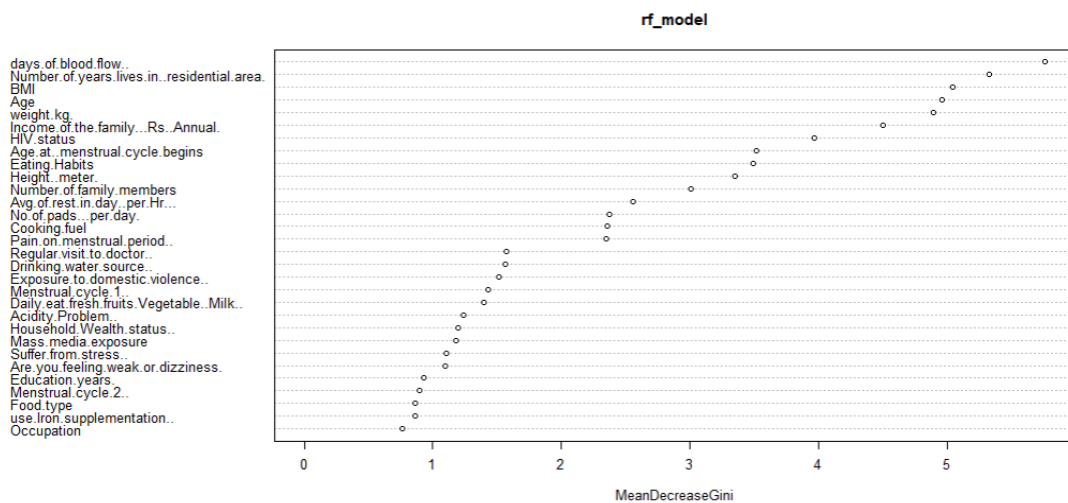


Fig. 5.7 Important variable by RF classifier

With a comparatively high importance score of 5.76, ‘Number of Days of Blood Flow’ is the most influential predictor variable and appears at the top of the list. This implies that the model’s predictions are significantly influenced by the length of blood flow during menstrual. ‘Number of Years Living in Residential Area’ and ‘BMI,’ with importance scores of 5.53 and 5.05, respectively, come in close succession. These variables appear to be relevant to the context being studied, as they appear to have a

significant influence on the model's decisions. The importance scores of the predictors 'Age,' 'Weight,' and 'Family Income' range from 4.9 to 4.96, making them significant as well. The anaemia predictions seem to be significantly shaped by these parameters.

Down the list, factors such as 'HIV Status,' 'Age at Menstrual Cycle Begins,' 'Eating Habits,' 'Height,' and 'Number of Family Members' have significance scores above three. Even though they have lower significance scores, the variables 'Alcohol Consumption,' 'Suffers from Any Long-Term Disease,' 'Community Women Education,' 'Toilet Facility,' 'Type of Addiction,' and 'Addiction' nonetheless add to the model's comprehension and ability to predict status of anaemia.

Interpretation:

While considering confusion matrix, random forest model demonstrated varying levels of performance across different classes, achieving near-perfect accuracy for some classes and struggling with others. The overall accuracy of 68.85% suggests that the model performs reasonably well but may benefit from further fine-tuning and optimization, especially for class '2' where it had the highest misclassification rate.

Even if a Random Forest model's 68% accuracy rate is a good result, more improvements must be taken into account for a better prediction performance. Developing an AdaBoost algorithm might be a strategic step to increase accuracy. Strong overall model creation and enhanced performance of weak learners are two of AdaBoost's many notable ensemble techniques. Therefore, in the next section ADA Boost algorithm was developed.

5.4.3 ADA boost algorithm:

Adaptive Boosting, or AdaBoost, is a one of the most powerful ensemble learning technique that emphasises strengthening the performance of base or weak learners by assigning greater weight to incorrectly classified data points. The R cod and output is as follows:

```
# Train an AdaBoost.M1 model
ada_model <- boosting(Anaemia ~ ., data = train, boos = TRUE, mfinal = 100)
ada_model
$call
boosting(formula = Anaemia ~ ., data = train, boos = TRUE, mfinal = 100)
attr(,"vardep.summary")
0 1 2 3
```

```

2 82 44 14
attr(,"class")
[1] "boosting"
$confusion

```

```

          Observed Class
Predicted Class 0 1 2 3
          0 0 1 0 0
          1 0 32 9 0
          2 1 6 10 0
          3 0 1 0 1

```

```

$error
[1] 0.295082

```

Interpretation:

The boosting model, with a total of 100 boosting iterations (mfinal = 100), has been applied to the 'Anaemia' prediction. The error rate was calculated to be 0.295082, which represents the overall misclassification rate of the Ada Boost model. In other words, about 29.51% of the predictions made by the model do not match the true class labels. Lower error rates indicate better model performance.

Ada boost algorithm shows 70.4918% accuracy.

Interpretation: As compared with previously developed models here the Ada Boost algorithm shows greatest accuracy. So, to predict the anaemia Ada Boost may be best model. To study more deeply of prediction point of view the variable importance should be estimated. The results are as follows:

```
> ada_model$importance
```

Acidity.Problem..	Age
0.3651887	7.8319268
Age.at..menstrual.cycle.begins	Alcohol.Consumption..
6.7946824	0.1668714
Any.Addiction..	Are.you.feeling.weak.or.dizziness.
0.0000000	1.7484945
Avg.of.rest.in.day..per.Hr...	BMI
1.8919778	8.9909611
Community.women.education	Cooking.fuel
0.1468177	3.3691950

Daily..Tea.intake	Daily.eat.fresh.fruits.Vegetable..Milk..
2.8258207	0.6257283
days.of.blood.flow..	Drinking.water.source..
4.1583585	1.0728130
Eating.Habits	Education.years.
2.3140093	0.8745133
Exposure.to.domestic.violence..	Food.type
2.4795182	0.8510047
Height..meter.	HIV.status
6.2336096	4.6160201
Household.Wealth.status..	Income.of.the.family...Rs..Annual.
0.1684940	5.1498211
Mass.media.exposure	Menstrual.cycle.1..
1.3370671	0.9367338
Menstrual.cycle.2..	No.of.pads...per.day.
0.7654389	4.1318144
Number.of.family.members	Number.of.years.lives.in..residential.area.
3.8938902	7.4425857
Occupation	Pain.on.menstrual.period..
2.6156551	1.1998966
Region	Regular.visit.to.doctor..
1.9430207	1.4465572
Suffer.from.any.long.term.disease	Suffer.from.stss..
0.6167066	0.9986046
Suffers.from.Diabetes	Toilet.facility..
0.0000000	0.0000000
Type.of.Addiction	use.Iron.supplementation..
0.0000000	0.9595109
weight.kg.	
9.0366919	

Variable Importance Table:

Table 5.5 Top significant variable by Ada Boot for unmarried WRA.

Variable	Variable importance
weight	9.0366
BMI	8.9909
age	7.8319
number of years lives in residential area	7.4425
age at menstrual cycle begins	6.7946
height	6.2336
family income	5.1498
HIV status	4.616
Days of blood flow	4.1583
No. of pads per day	4.1318
number of family members	3.8938
Cooking fuel	3.3691
Daily tea intake	2.8258
occupation	2.6156
Exposure to domestic violence	2.4795
eating habits	2.314

Interpretation:

With a comparatively high relevance score of 9.0366, 'Weight' stands out as the most influential predictor variable at the top of the list. This implies that a women's weight has a big impact on status of anaemia. 'BMI' (body mass index) and 'Age,' with relevance values of 8.9909 and 7.8319, respectively, are next in line. These predictor variables highlight their significance in the context under discussion by demonstrating their significant influence on status of anaemia. The significant relevance scores of 'Age at Menstrual Cycle Begins' and 'Number of Years Living in Residential Area' also suggest their impact on the status of anaemia. With scores higher than 5, 'Height' and 'Family Income' are also deemed to be quite significant with respect to anaemia. With relevance values ranging from 4.1583 to 4.616, 'HIV Status,' 'Days of Blood Flow,' 'No. of Pads per Day,' and 'Number of Family Members'

continue to contribute as you move down the list. Even though they are ranked significantly lower, the factors ‘Cooking Fuel,’ ‘Daily Tea Intake,’ ‘Occupation,’ ‘Exposure to Domestic Violence,’ and ‘Eating Habits’ are nonetheless significant.

The variables at the top of the list are the key drivers of the anaemia in unmarried WRA, while those at the bottom, though less influential, still contribute to the model’s overall understanding and predictive power.

5.5 Overall comparison of machine learning algorithms:

Table 5.6 Comparison of machine learning algorithms for Unmarried women.

Sr. No.	Machine Learning Algorithm	Accuracy
1	Decision tree	50.90%
2	Decision tree (after cross validation)	65.45%
3	support vector machine(linear)	62.29%
4	support vector machine(Polynomial)	60.66%
5	support vector machine(Sigmoid)	55.74%
6	support vector machine(Radial)	57.38%
7	K- nearest Neighbour(with k=7)	60.65%
8	Bagged Decision tree(with nbag=100)	64.80%
9	Random Forest Algorithm (with 100 trees)	68.85%
10	Ada Boost (mfinal=100)	70.49%

Above result can be well examined by graphically.

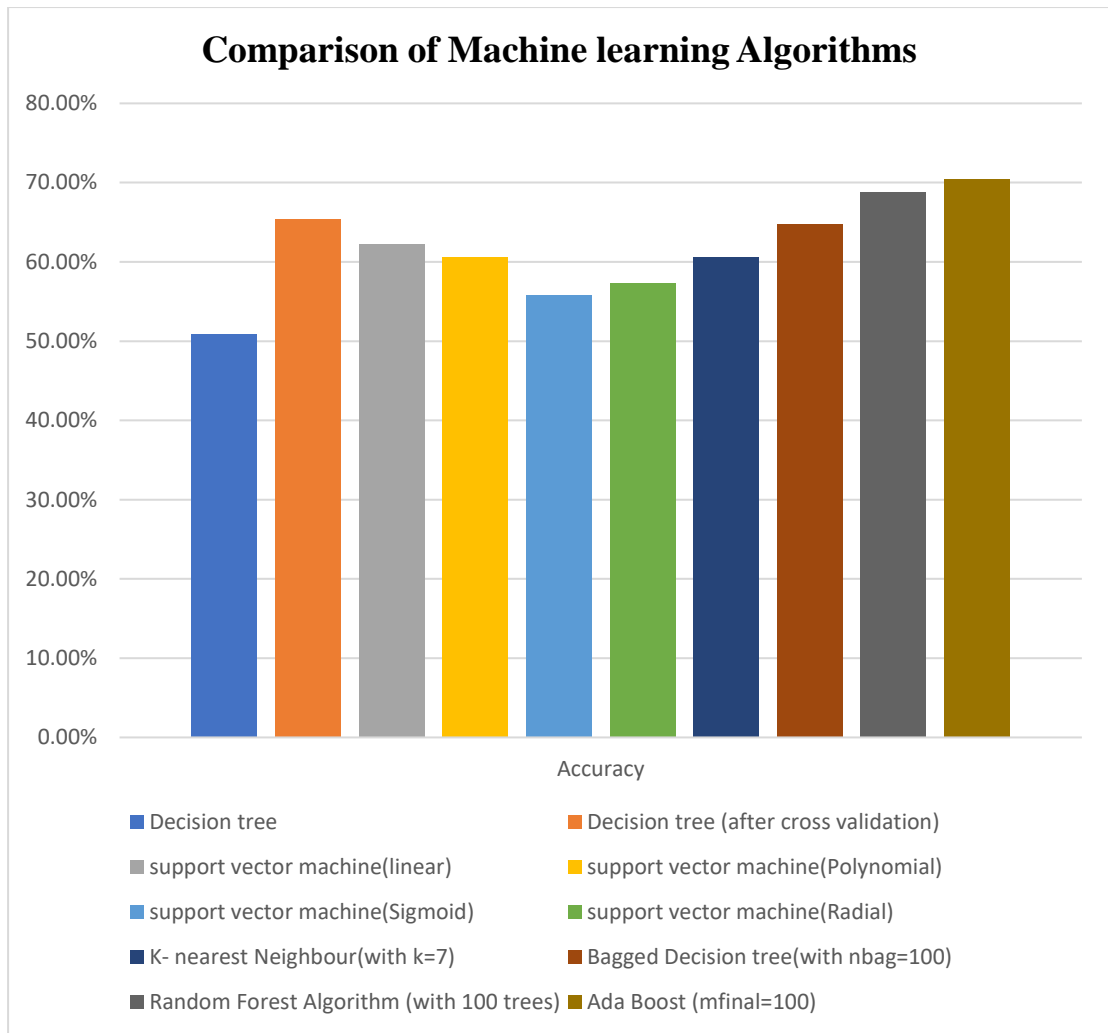


Fig. 5.8 Comparison of Machine learning Algorithms for Unmarried WRA

From the above table we can see that the first-generation decision tree model had a 50.90% accuracy rate. The decision tree model profited from the cross-validation procedure, as seen by the accuracy's increase to 65.45% after cross-validation was completed. According to the SVM method, the linear kernel produced the best accuracy compared to other kernels, suggesting that the data might be linearly separable. Based on the accuracy results, it can be seen that the Ada Boost algorithm (with mfinal=100) produced the examined models' maximum accuracy. Since the data was unbalanced there was problem in the prediction of class "no anaemia". In the unmarried WRA data the not anaemic girls are in smaller count this problem was arise. But in influential factor point of view we can find influential factors by using best fitted machine learning model. At least we can make statement about key contributors of presence of anaemia in unmarries WRA by using this model.

Overall Interpretation:

Since the Ada Boost algorithm shows greatest accuracy further conclusions were made on the basis of Ada Boost algorithm results. According to Ada Boost algorithm, ‘Weight,’ ‘BMI,’ ‘Age,’ ‘Number of Years Living in Residential Area,’ and ‘Age at Menstrual Cycle Begins’ are among the most influential factors. These factors plays significant role in shaping the model’s predictions and are crucial in understanding the outcomes related to anaemia in unmarried WRA. Given their high importance scores, it is paramount to pay special attention to these factors. Moreover, ‘Height,’ ‘Family Income,’ ‘HIV Status,’ ‘Days of Blood Flow,’ and ‘No. of Pads per Day’ also hold substantial importance, indicating that they significantly influenced the anaemia in unmarried WRA.

Monitoring and potentially modifying these factors could have a notable impact on the anaemia in unmarried WRA. It’s important to consider interventions or strategies related to these factors to potentially reduce the anaemia in unmarried WRA.

While all the listed variables are important, focusing on the top contributors can be a priority when designing interventions or strategies to reduce the anaemia in reproductive aged girls. These insights provide valuable guidance for data-driven decision-making and actions to improve outcomes or predictions related to anaemia in unmarried WRA. Therefore, in the next section most important variables according to best model was examined.

5.6 Relationship of anaemia with significant contributors in prediction.

Table 5.7 Average weight of unmarried WRA according to Anaemia status

Anaemia Status	No-anaemia	Mild-Anaemia	Moderate-anaemia	Severe-Anaemia
Average of weight(kg)	50.5	47.175	46.69354839	45.625

The above table data presents the average weights in kilograms for Unmarried WRA categorized by their ‘Anaemia Status,’ which is divided into four groups: ‘No anaemia,’ ‘Mild Anaemia,’ ‘Moderate Anaemia,’ and ‘Severe Anaemia.’ The data reveals distinct trends in weight across these groups.

First, for individuals with ‘No anaemia,’ the average weight is approximately 50.5 kilograms. This is the category with the greatest average weight, indicating that people without anaemia often weigh a little bit more than people with anaemia.

Next, in the ‘Mild Anaemia’ category, the average weight is slightly lower at approximately 47.175 kilograms. Moving on to the ‘Moderate Anaemia’ group, the average weight is approximately 46.69 kilograms. Finally, the ‘Severe Anaemia’ group

has the lowest average weight of approximately 45.625 kilograms. This group is associated with the most severe form of anaemia, and the notably lower average weight underscores the impact of severe anaemia on an individual's weight.

In summary, the data illustrates a clear association between the severity of anaemia and the average weight of individuals. As the severity of anaemia increases from mild to moderate and severe, the average weight tends to decrease. These findings could have important implications for healthcare and nutrition interventions, as they suggest that weight monitoring and nutritional support may be particularly crucial for Unmarried WRA with more severe forms of anaemia to improve their overall health and well-being. The next important factor associated with anaemia is BMI. Therefore, in the next section average BMI and anaemia status was studied.

Table 5.8 Average BMI of unmarried WRA according to Anaemia status

Anaemia Status	No-anaemia	Mild-Anaemia	Moderate-anaemia	Severe-Anaemia
Average of BMI	27.19660086	27.1943161	26.89326249	23.96684595

First, in the 'No anaemia' category, the average BMI is approximately 27.20. Moving on to the 'Mild Anaemia' group, the average BMI is nearly identical, at approximately 27.19. This implies that individuals with mild anaemia have similar average BMI values to those without anaemia, with very little difference between the two groups. In the 'Moderate Anaemia' category, the average BMI is slightly lower at approximately 26.89. This indicates that individuals with moderate anaemia, while still maintaining a reasonable average BMI, tend to have a somewhat lower BMI compared to individuals without anaemia. Finally, in the 'Severe Anaemia' group, the average BMI is notably lower, at approximately 23.97. This group, representing individuals with the most severe form of anaemia, exhibits the lowest average BMI, suggesting a significant drop in body mass index in WRA with severe anaemia.

In conclusion, the data reveals that an intriguing relationship between the severity of anaemia and BMI. Individuals with more severe forms of anaemia tend to have significantly lower average BMI values. This observation suggests that severe anaemia can be associated with a decreased BMI, potentially indicating a connection between the health status of WRA with severe anaemia and their nutritional well-being. Monitoring BMI in individuals with anaemia, especially severe cases, is crucial for healthcare professionals to tailor appropriate interventions and support to improve their

overall health. Next important factor according to Ada Boost is age. So, in the next section age was examined according to status of anaemia.

Table 5.9 Average age of unmarried WRA according to Anaemia status

Anaemia Status	No anaemia	Mild Anaemia	Moderate anaemia	Severe Anaemia
Average of Age	23.33333333	18.35	17.93548387	24

First, for individuals categorized as having ‘No anaemia,’ the average age is approximately 23.33 years. Moving on to the ‘Mild Anaemia’ group, the average age is notably lower at 18.35 years. In the ‘Moderate Anaemia’ category, the average age is approximately 17.94 years. Lastly, the ‘Severe Anaemia’ group has an average age of 24 years. This group, which is associated with the most severe form of anaemia, shows a slightly higher average age, indicating that individuals with severe anaemia are somewhat older than those with moderate anaemia. From the above table we can’t make strong conclusion about trend of relationship between anaemia status and age. Therefore, move toward the nature of the distribution of both variables.

```
> table(data$Anaemia,data$Age)
 15 16 17 18 19 20 21 22 23 24 25 28 29 30 34 38 41
 0 0 0 2 0 1 0 0 1 0 1 0 0 0 0 0 0 1
 1 2 21 42 19 10 11 3 2 4 1 2 0 1 1 0 1 0
 2 1 12 18 14 10 3 3 0 0 0 0 0 0 0 1 0 0
 2 0 0 0 4 0 0 0 0 2 2 2 6 0 0 0 0 0
```

Table 5.10 Frequency distribution of age of unmarried WRA according to Anaemia status

		Age of WRA																
		15	16	17	18	19	20	21	22	23	24	25	28	29	30	34	38	41
Anaemia Status	No Anaemia	0	0	2	0	1	0	0	1	0	1	0	0	0	0	0	0	1
	Mild	2	2	4	1	1	1	3	2	4	1	2	0	1	1	0	1	0

	Moderate	1	1	1	1	1	3	3	0	0	0	0	0	0	1	0	0
	Severe	0	0	0	4	0	0	0	0	2	2	2	6	0	0	0	0

The table gives a thorough summary of the prevalence of anaemia in women who are of reproductive age (WRA) in various age groups, from 15 to 41 years old. The rows correspond to the different levels of anaemia severity: ‘No Anaemia,’ ‘Mild,’ ‘Moderate,’ and ‘Severe.’ The columns show the individual ages of the women.

There were only 6 WRA found to be non- anaemic. The ‘Mild’ category demonstrates a higher prevalence of anaemia, especially in the earlier age groups. The highest occurrences are observed at ages 16-20 years, suggesting that mild anaemia is more common in younger women ‘ Moderate anaemia’ found in age 16-19 years and severe anaemia occurred in 23-28 years age group.

Next important factor according to Ada Boost algorithm is ‘number of years lives in residential area’. So in the next section this variable was observes in accordance with status of anaemia.

Table 5. 11 Average of Number of years lives in residential area.

Anaemia Status	No anaemia	Mild Anaemia	Moderate anaemia	Severe Anaemia
Mean of Number of years lives	17	13.76041667	16.11290323	17.125

First, for individuals categorized as having ‘No anaemia’, the mean of no. of years alive is 17 years. Moving on to the ‘Mild Anaemia’ group, the average no. of years lived is slightly lower at approximately 13.76 years. This indicates that individuals with mild anaemia have, on average, a shorter residential tenure compared to those without anaemia. In the ‘Moderate Anaemia’ category, the average number of years lived in the residential area is approximately 16.11 years. This group has a slightly shorter average residential tenure compared to those without anaemia but longer than individuals with mild anaemia. This indicates that moderate anaemia is associated with intermediate residential stability. Lastly, the ‘Severe Anaemia’ group has an average of 17.13 years lived in the residential area, which is similar to the ‘No anaemia’ group.

```
> table(data$Anaemia, data$`Number of years lives in residential area.`)
```

```
0.25 1 2 3 4 5 6 7 8 9 10 11 12 13 15 16 17 18 19 20 21 22 23 25 26 30 41
0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1
1 1 1 1 5 3 2 1 4 0 0 2 2 1 1 1 2 26 32 10 6 4 2 0 1 1 1 1 0
```

2 0 0 0 0 0 2 0 0 1 0 2 0 1 0 6 13 21 12 2 2 0 0 0 0 0 0 0
 3 0 0 0 0 0 0 0 2 2 0 0 0 0 0 0 4 0 2 0 2 0 2 0 0 0 2 0

Average age at menstrual cycle begins of unmarried WRA according to Anaemia status:

The presented data provides a nuanced exploration of the age at which individuals experience the onset of their menstrual cycles, considering various levels of anaemia severity.

Table 5.12 Average age at menstrual cycle begins with anaemia.

	No anaemia	Mild Anaemia	Moderate anaemia	Severe Anaemia
Variance of Age at menstrual cycle begins	8.266666667	1.825210084	1.842411423	1.983333333
Arithmetic mean of Age at menstrual cycle begins	14.33333333	13.7	13.83870968	14.125

Variance in the age at menarche offers insights into the dispersion or diversity of this critical developmental milestone within each anaemia category. Notably, individuals without anaemia exhibit the highest variance, signifying a wider range of ages for menarche in this group. Conversely, both mild and moderate anaemia categories display lower variances, indicating a more consistent age range for menarche among individuals with these anaemia severities. The variance for severe anaemia is also relatively low, suggesting uniformity in the age range at which menarche occurs in this group.

Examining the average age at menarche within each anaemia category provides further context. Individuals without anaemia and those with severe anaemia share a similar average age of approximately 14.33 and 14.13 years, respectively. In contrast, individuals with mild anaemia have a slightly lower average age of 13.7 years, potentially suggesting an earlier onset of menstrual cycles on average in this group. Similarly, individuals with moderate anaemia exhibit an average age of 13.84 years, aligning closely with the mild anaemia group.

Variance in the age at menarche offers insights into the dispersion or diversity of this critical developmental milestone within each anaemia category. Notably, individuals without anaemia exhibit the highest variance, signifying a wider range of ages for menarche in this group. These findings underscore the intricate relationship between anaemia severity and the timing of menarche. While the average age provides

a central tendency for each group, the variance elucidates the degree of variability in the age at which individuals in these groups experience this crucial developmental event. Further exploration into the factors influencing these patterns could unveil important connections between anaemia and reproductive health, contributing to a more comprehensive understanding of the complex interplay between health outcomes and the onset of menstrual cycles.

Average height of unmarried WRA according to Anaemia status:

Table 5.13 Average height of unmarried WRA according to Anaemia.

	No anaemia	Mild Anaemia	Moderate anaemia	Severe Anaemia
Average Height in cm	156.8	151.9	152	158

From the above table it was discovered that the average height for no Anaemia is 156.8 cm. Average heights for mild anaemia and moderate anaemia was found to be 151.9 and 152 cm respectively which is nearly same. The average height of Severe anaemia same 158 cm. In summary, the data suggests that individuals with mild and moderate anaemia may, on average, have a reduced height compared to those without anaemia. However, the trend shifts for individuals with severe anaemia, who exhibit a somewhat taller average height.

According to Ada boost annual income was also key factor. Financial condition may reflect the status of anaemia. To study this factor following table was draw.

Table 5.14 Income of family and anaemia.

	No anaemia	Mild Anaemia	Moderate anaemia	Severe Anaemia
Mean Income of the family (Rs.) Annual	140000	144633.3333	138548.3871	51250

From the above table some conclusions were made about the factor family income of the respective WRA. According to average, ‘No Anaemia’ category, the average family income stands at Rs. 140,000 annually. For individuals with ‘Mild Anaemia,’ the average family income slightly increases to Rs. 144,633.33 per annum. The ‘Moderate Anaemia’ group reports an average family income of Rs. 138,548.39 annually. In the ‘Severe Anaemia’ category, the average family income notably decreases to Rs. 51,250 per year. From these figures we can say that, family income slightly inversely proportional to the anaemia severity, Anaemia severity increases as the annual income decreases.

According to Ada Boost algorithm HIV status is also a significant factor associated with anaemia. The distribution of anaemia according to HIV status is as follows:

Table 5.15 HIV status and anaemia.

	HIV Negative	HIV Positive
No Aneamia	6	0
Mild Anaemia	119	1
Moderate Anaemia	61	1
Severe Anaemia	6	10

The table suggests that there is a relationship between HIV status and the presence/severity of anaemia. Individuals with HIV seem to have a higher prevalence of anaemia, in the severe category. This could imply that HIV-positive individuals are more likely to experience severe anaemia compared to HIV-negative individuals.

This information might be useful for healthcare professionals and policymakers to better understand the health conditions of individuals with HIV and to tailor interventions accordingly. Further statistical analysis, such as chi-square tests, could provide more insights into the strength and significance of the observed associations.

```
> table(data$Anaemia, data$HIV.status)
 0 1
0 6 0
1 118 1
2 61 1
3 6 10
> chisq.test(data$Anaemia, data$HIV.status, correct=TRUE)
Pearson's Chi-squared test
data: data$Anaemia and data$HIV.status
X-squared = 100.06, df = 3, p-value < 2.2e-16
```

Interpretation:

Null Hypothesis (H₀): There is no association between Anaemia and HIV status.

Alternative Hypothesis (H₁): There is association between Anaemia and HIV status.

Test Conclusion: Since the p-value is very low, there is strong evidence against the null hypothesis. Therefore, it was concluded that there is a significant association between status of anaemia and HIV status in unmarried WRA.

The next significant variable is no. of days of blood flow which was numeric variable therefore mean of no. of days of blood flow was examined with anaemia status.

Table 5.16 Number of days of blood flow with anaemia.

	No anaemia	Mild Anaemia	Moderate anaemia	Severe Anaemia
Average No. days of blood flow	4.333333333	4.733333333	4.258064516	2.375

There seems to be a trend where the average days of blood flow decrease as the severity of anaemia increases. WRA with severe anaemia have the lowest average days of blood flow, suggesting that severe anaemia might be associated with a shorter duration of blood flow.

No of pads (per day) also found to be significant contribution while predicting anaemia.

Table 5.17 No of pads (per day) with anaemia.

	No anaemia	Mild Anaemia	Moderate anaemia	Severe Anaemia
Mean of No. of pads (per day)	2.166666667	2.408333333	2.419354839	3.25

There is a trend where the average number of pads used per day increases with the severity of anaemia. WRA with severe anaemia, on average, use a higher number of pads per day compared to those with milder forms of anaemia or no anaemia.

Number of family members shows significant association according to the best model.

Table 5.18 Mean family size according to anaemia categories.

	No anaemia	Mild Anaemia	Moderate anaemia	Severe Anaemia
Average of total family members	7.5	6.025	5.661290323	5.25

Whenever the severity of anaemia is increased, there appears to be a pattern in which the average number of family members drops. Those with WRA who do not suffer from anaemia have the largest average number of family members available to them, whereas those who suffer from severe anaemia have the lowest average number. This result was exactly contradictory to pilot study result. It may be because this data is of unmarried girls and pilot study data was entire married, unmarried and pregnant WRA. In the next the Type of Cooking fuel was examined. There are various types of cooking fuel examined. Following table shows the distribution of anaemia according to various fuel type.

```
> t=table(data$Anaemia, data$Cooking.fuel)
```

```
> t
```

```

0 2 3 4 5 6 7 8 9 13 16 17 18 23 25 26
0 4 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0
1 59 1 0 1 1 0 2 15 0 1 3 1 22 13 0 0
2 30 0 3 0 0 3 0 8 1 0 0 1 8 7 0 1
3 4 0 0 0 0 6 0 0 0 0 0 0 0 4 2 0

```

```
> chisq.test(t, correct = TRUE)
```

Pearson's Chi-squared test

data: t

X-squared = 97.59, df = 45, p-value = 9.393e-06

From the R output we can say that ,

P-value<0.05, Therefore we reject the null hypothesis.

Therefore, we can say that there is association between type of fuel and anaemia type.

Daily Tea intake:

```
> t2=table(data$Anaemia, data$Daily..Tea.intake)
```

```
> t2
```

```

0 1
0 4 2
1 29 90
2 16 46
3 2 14

```

Upon initial inspection, it appears that individuals who take tea are more prevalent across all levels of anaemia compared to those who do not take tea. This observation suggests a potential association between tea consumption and anaemia severity.

To rigorously assess the statistical significance of this association, a chi-squared test

```
> chisq.test(t2, correct=TRUE)
```

Pearson's Chi-squared test

data: t2

X-squared = 6.9113, df = 3, p-value = 0.07478

Interpretation:

The p-value of 0.07478 is greater than the typical significance level of 0.05. Therefore, at the 0.05 significance level, we do not have enough evidence to reject the null hypothesis. This means that, based on the data and the chi-squared test, we do not have sufficient evidence to conclude that there is a significant association between the two variables.

CHAPTER 6

COMPARING THE PERFORMANCE OF MACHINE LEARNING ALGORITHMS ON MARRIED NON-PREGNANT WRA

6.1 Introduction: Although the questionnaire was same for married and unmarried WRA but there was difference in both the categories. Therefore, there was need to separate the married and unmarried WRA. In this section we are going to study the anaemia in married non-pregnant WRA. There were 210 married non-pregnant WRA selected for this study. In this section some additional variables than unmarried WRA were included such as pregnancy and marital factors related variables like miscarriage history, number children ever born, use of contraceptive, husband's occupation, husband's age, etc. Same methodology was used on this data set to find the significant contributors of anaemia in married non-pregnant WRA.

The original data contains 210 samples, after data pre-processing there were 203 samples. The next step is to check the distribution of Anaemia classes. Following R code shows the number of instances in each classes.

6.2 Prevalence of anaemia among Non-Pregnant WRA:

```
> table(data$Anaemia)
 0  1  2  3
61 82 42 18
```

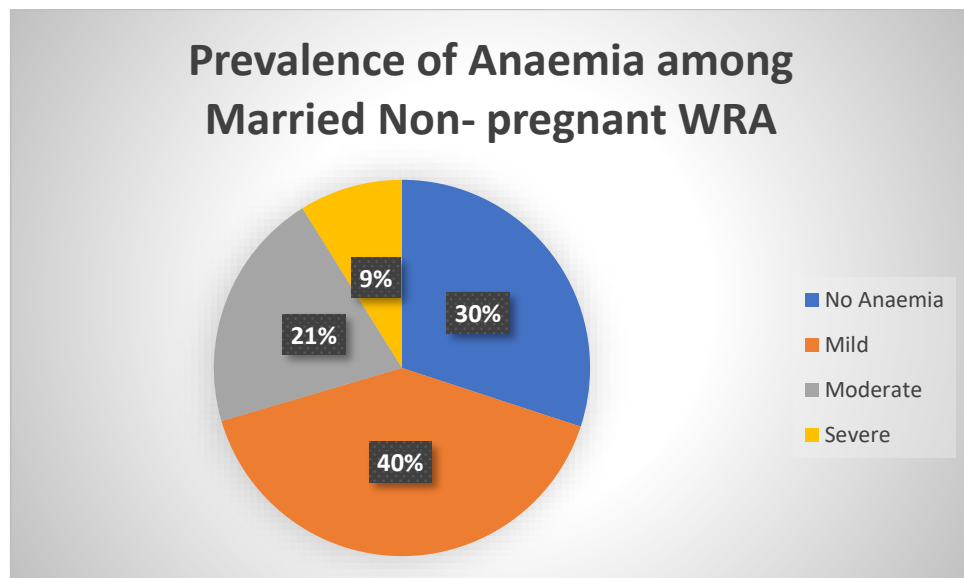


Fig. 6.1 Prevalence of Anaemia among Married Non-pregnant WRA.

The distribution of anaemia prevalence rates within the population is worrying when looking at the specific categories. The data suggests that a 30% are not impacted by anaemia, indicating that this grouping is generally in good condition. But as we

progress across the spectrum, the prevalence steadily rises, with 40% dealing with mild anaemia, 21% with moderate anaemia, and 9% with severe anaemia. This distribution indicates possible health issues by highlighting the existence of anaemia in varied degrees within the population.

The greater frequency of mild anaemia may point to a generalised, albeit manageable, health issue, whereas the presence of moderate and severe anaemia in a significant section of the population may necessitate immediate attention and treatment. Policymakers and healthcare providers must make sure that.

Here we can easily see the class 2 (moderate anaemia) and class 3 (severe) anaemia has minimum number of instances than class 0 (No anaemia) and class 1 (mild anaemia). In examining health data, it is promising to note that moderate anaemia and severe anaemia exhibit lower frequencies. This observation holds substantial significance as it suggests a potential positive trend in the prevalence of these conditions within a given population or dataset. Lower frequencies of moderate and severe anaemia imply a potential improvement in overall health outcomes, indicating fewer individuals grappling with these more severe forms of anaemia. Such a trend could indicate effective healthcare interventions, advancements in medical treatments, or improvements in nutritional awareness and access. This promising revelation signifies progress in combating these conditions, showcasing a positive trajectory towards better health and well-being for the affected individuals and communities. Identifying reduced frequencies of moderate and severe anaemia is a testament to the effectiveness of various interventions and underscores the importance of continued efforts to sustain and further improve these encouraging trends in healthcare. For this there is need to assess the factors affecting on the presence of anaemia in non-pregnant WRA. Therefore in the next section various supervised machine learning algorithms were developed.

6.3 Decision Tree:

To develop the various machine learning algorithms data was divided into train and test data the splitting criteria was 80-20. Train data contains 162 samples and test data contains 41 sample points. Therefore, the Machine learning algorithms developed on train data and tested on test data.

```
> set.seed(123)
> train_indices <- sample(1:nrow(data), 0.8 * nrow(data))
> train_data <- data[train_indices, ]
```

```

> test_data <- data[-train_indices, ]
> train=data.frame( train_data)
> test=data.frame(test_data)
> data$Anaemia=as.factor(data$Anaemia)
> train$Anaemia=as.factor(train$Anaemia)
> test$Anaemia=as.factor(test$Anaemia)
> length(train$Anaemia)
[1] 162
> length(test$Anaemia)
[1] 41

```

First the simple decision tree algorithm with 10 fold cross validation was developed on whole data. The R output given below:

```

> # 10-fold cross validation decision tree
> train$Anaemia <- as.factor(train$Anaemia)
> trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
> mc1=train(Anaemia~., data=data, method='rpart',trControl=trctrl)
> mc1

```

CART

203 samples

52 predictor

4 classes: '0', '1', '2', '3'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 183, 184, 183, 182, 182, 183, ...

Resampling results across tuning parameters:

cp	Accuracy	Kappa
0.03719008	0.3791331	0.08946299
0.04793388	0.3949225	0.11115273
0.12396694	0.3987286	0.02304890

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was cp = 0.1239669.

Interpretation:

The summary of the trained model 'mc1' is shown in the output. CART (Classification and Regression Trees) is the type of model used. There are four classes, denoting

various degrees or divisions of the target variable: '0,' '1,' '2,' and '3.' From 10-fold cross-validation, the model's performance is assessed. The data was divided into 10 folds, as shown in the 'Resampling' section, and the summary lists the sample sizes for each fold. The assessment outcomes for various 'cp' parameter values, which regulate the complexity of the tree, are displayed in the 'Resampling results across tuning parameters'. The accuracy and Kappa coefficient are given for each value of 'cp'. While the Kappa coefficient assesses the degree of agreement between predictions and actual values, accuracy evaluates the general reliability of model predictions. The ideal model is chosen as the one with the maximum accuracy. The chosen model's 'cp' value is 0.1239669 in the end. So to develop the Decision tree by using cp parameter 0.1239669.

The R output is as follows:

```
> m1=rpart(Anaemia~.,data= train, method='class',cp = 0.1239669)
```

```
> summary(m1)
```

Call:

```
rpart(formula = Anaemia ~ ., data = train, method = 'class',
      cp = 0.1239669)
```

n= 162

	CP	nsplit	rel error	xerror	xstd
1	0.1485149	0	1.0000000	1.0000000	0.06105858
2	0.1239669	1	0.8514851	1.039604	0.06018014

Variable importance

Are.you.feeling.weak.or.dizziness.	78	husband.s.age.at.marriage	6
husband.s.age	5	Age	4
Income.of.the.family...Rs..Annual.	4	weight.kg.	4

```
> rpart.plot(m,extra=104)
```

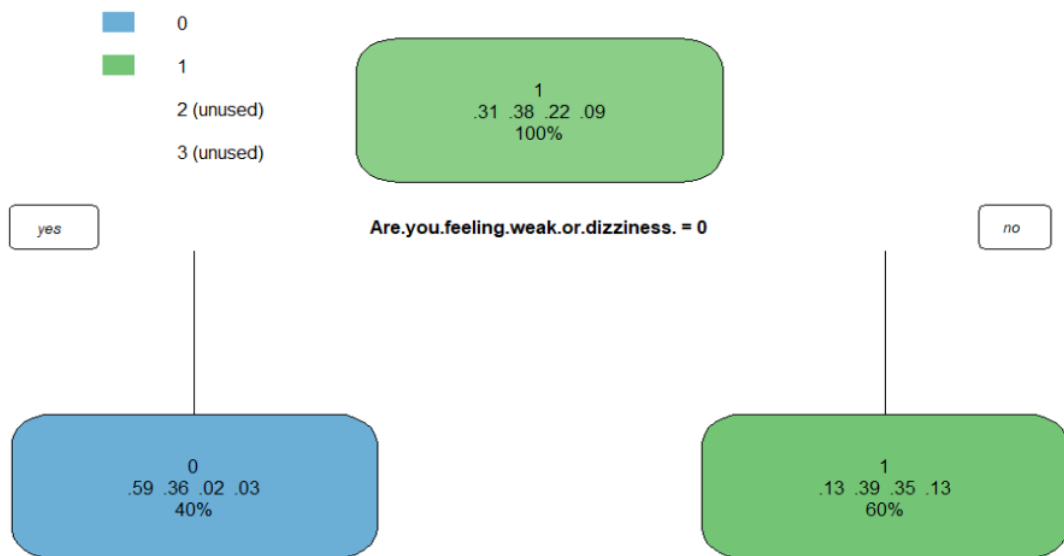


Fig. 6.2 Decision tree 1 for Non-pregnant WRA

From the above decision tree plot we can't say about the significant factors associated with the Anaemia since it is only contains feeling weak or dizziness. But from variable importance table we can say that Feeling Weak, age of Husband at the time of marriage, husband's age, family income, weight are found to be most significant factors associated with the stage of Anaemia according to decision tree. To make further conclusion there was need of model assessment. Confusion matrix was drawn to find model accuracy, precision and recall.

```
> P=predict(m1, newdata = test, type='class')
> Table=table(test$Anaemia,P)
> Table
  P
  0 1 2 3
0 7 3 0 0
1 7 14 0 0
2 0 7 0 0
3 0 3 0 0
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.5121951
```


The decision tree with $cp= 0.1239669$ shows 51.21% accuracy for new data. But when observing the confusion matrix, we can see that the decision tree model only able to predict the class 0 and 1.

```
> # Confusion matrix
> confusion_matrix <- matrix(c(7, 3, 0, 0,
+                             7, 14, 0, 0,
+                             0, 7, 0, 0,
+                             0, 0, 3, 0),
+                             nrow = 4, byrow = TRUE)
> # Function to calculate precision, recall, and F1 score for each class
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FN <- rowSums(cm) - TP
+   FP <- colSums(cm) - TP
+
+   precision <- TP / (TP + FP)
+   recall <- TP / (TP + FN)
+   f1_score <- 2 * precision * recall / (precision + recall)
+
+   metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall =
recall, F1_Score = f1_score)
+   return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
  Class Precision Recall F1_Score
1    0 0.5000000 0.7000000 0.5833333
2    1 0.5833333 0.6666667 0.6222222
3    2 0.0000000 0.0000000      NaN
4    3      NaN 0.0000000      NaN
```

The precision, recall, and F1-score figures show that the model works reasonably well for classes 0 and 1. For classes 2 and 3, the model does not perform well, either as a result of the lack of true positives or the extremely unbalanced class

distributions. When this occurs, the model's accuracy, recall, and F1-scores are undefined (NaN), demonstrating that it is unable to accurately predict these classes.

Other evaluation criteria, such as ROC AUC, area under the precision-recall curve (AUC-PR), and class balance, must be considered since the data was unbalanced.

```
> # ROC curves
> # Create a multiclass ROC object
> # Predict probabilities for all classes
> predicted_probabilities <- predict(m1, newdata = test_data, type = 'vector')
> # Create a multiclass ROC object
> roc_obj <- multiclass.roc(test_data$Anaemia, predicted_probabilities)
> # Calculate and print ROC AUC for each class
> roc_auc <- roc_obj$auc
> cat('ROC AUC for each class:\n')
ROC AUC for each class:
> print(roc_auc)
Multi-class area under the curve: 0.7028
```

In conclusion, a multi-class AUC of 0.7028 indicates that the fitted model has some ability to distinguish between the different classes. To further enhance the model's performance, particularly for minority classes, strategies like resampling (either oversampling or under sampling), altering class weights, or utilising various algorithms may be required. SVMs can handle unbalanced data, and we can increase their performance by adjusting class weights or applying strategies like SMOTE (Synthetic Minority Over-sampling Technique). In the next section the Support Vector Machine (SVM) with various kernels were developed and accuracy results were examined for comparison.

6.4 Support Vector Machine with cost sensitive model:

It was needful to check all the kernels of SVM algorithm on this dataset. The SVM model was developed by using linear, radial, polynomial and sigmoid kernel. Comparative study was done to select the kernel which gives maximum accuracy. Following R output shows the SVM algorithm for various kernels.

```
> # Load necessary libraries
> library(e1071)
> # Define the cost value (C) to be used for all kernels
> cost_value <- 1
```

```

> # Define a vector of kernel names
> kernels <- c('linear', 'radial', 'polynomial', 'sigmoid')
> # Initialize a vector to store the accuracy values
> accuracies <- numeric(length(kernels))
> # Loop through different kernels and train SVM models
> for (i in 1:length(kernels)) {
+   kernel <- kernels[i]
+
+   # Train the SVM classifier with the specified kernel and cost value
+   svm_model <- svm(
+     formula = Anaemia ~ .,
+     data = train,
+     type = 'C-classification',
+     kernel = kernel,
+     cost = cost_value
+   )
+
+   # Make predictions on the test data
+   P <- predict(svm_model, newdata = test)
+
+   # Calculate accuracy
+   Table <- table(test$Anaemia, P)
+   Accuracy <- sum(diag(Table)) / sum(Table)
+
+   # Store accuracy in the vector
+   accuracies[i] <- Accuracy
+
+   # Print accuracy for the current kernel
+   cat('Kernel:', kernel, 'Accuracy:', round(Accuracy, 2), '\n')
+ }
Kernel: linear Accuracy: 0.55
Kernel: radial Accuracy: 0.4
Kernel: polynomial Accuracy: 0.36
Kernel: sigmoid Accuracy: 0.36

```

```

> # Create a bar plot of accuracy for different kernels
> barplot(accuracies, names.arg = kernels, ylim = c(0, 1), col = 'skyblue',
+       main = 'Accuracy of SVM Kernels', ylab = 'Accuracy')
> # Add accuracy values as text labels above each bar
> text(1:length(kernels), accuracies, labels = round(accuracies, 2), pos = 3)

```

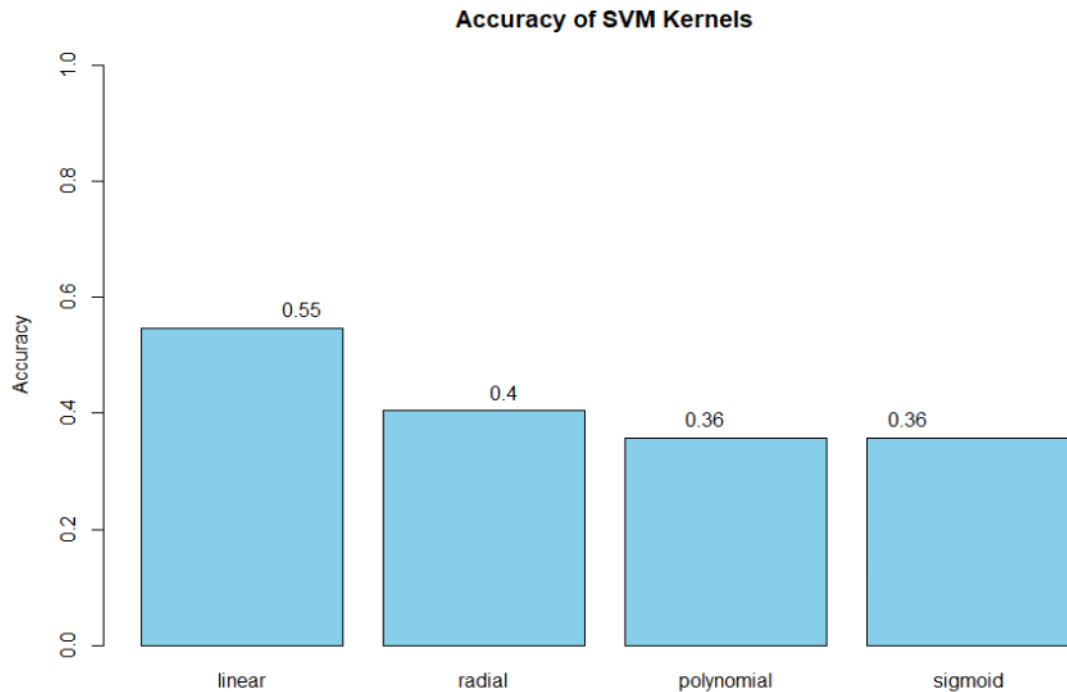


Fig. 6.3 Accuracy with various kernels of SVM

Result and Discussion:

The simplest kind of kernel function is the linear kernel. Without any modifications, it represents the data in the original feature space. The linear kernel assumes that the data can be separated linearly with 54.76% accuracy.

The polynomial kernel function used a polynomial function to translate the data into a higher-dimensional space. This make it possible to capture intricate patterns and nonlinear relations with accuracy of 36%. Another nonlinear kernel function that translates the data into a higher-dimensional space is the sigmoid kernel. It possesses sigmoid function characteristics, which are frequently utilised in logistic regression. It achieved 36% accuracy, which was a bit less than the polynomial kernel. There were Here we can't move forward with sigmoid kernel as the response variable is not binary.

The RBF (Radial Basis Function) kernel, also referred to as the Gaussian kernel, turns the data into an infinitely large space. It has the ability to detect intricate and

irregular patterns in the data. Among the aforementioned kernels, it performs the 40% accuracy. The SVM's radial kernel (RBF) is a flexible kernel function that can be applied in a variety of circumstances. The following are some situations where the radial kernel is frequently used:

- i) Non-linearly separable data,
- ii) Multi-class classification
- iii) Unbalanced datasets
- iv) Feature extraction

In conclusion, among the listed kernels, the linear kernel had the highest accuracy (55%), followed by the radial kernel (40%). The accuracy of the polynomial and sigmoid kernel was found to be lowest (36%).

According to the comparative analysis, it was discovered the linear kernel had an accuracy of about 55%, which was higher than the radial kernel's accuracy of about 40%. Both the sigmoid and polynomial kernels attained accuracy levels of about 36%. These findings suggest that the linear kernel outperforms the other kernels for this particular problem and dataset in terms of accuracy. But still we can't move forward with this SVM algorithm since accuracy was just 55%. Therefore, the further analysis was done by using 'linear' kernel. There is scope of cost selection. In an SVM model, it is crucial for finding an appropriate balance between obtaining a low training error and a low testing error. The cost parameter, frequently abbreviated as 'C,' is an important hyperparameter in Support Vector Machines (SVMs). So, the SVM models of liner kernel with various cost parameters were developed. The results are as follows:

```
> library(e1071)
># Define a range of cost values to try
> cost_values <- c(seq(1,10))
> # Initialize a list to store the SVM models
> svm_models <- list()
> # Loop through different cost values and train SVM models
> for (cost_value in cost_values) {
+ # Train the SVM classifier with cost-sensitive learning
+ classifier_cost_sensitive <- svm(
+ formula = Anaemia ~ .,
+ data = train,
+ type = 'C-classification',
```

```

+ kernel = 'linear',
+ cost = cost_value
+ )
+
+ # Store the trained model in the list
+ svm_models[[as.character(cost_value)]] <- classifier_cost_sensitive
+ }
> # Make predictions on the test data for each model
> predictions <- lapply(svm_models, function(model) predict(model, newdata = test))
> # Calculate accuracy for each model
> accuracies <- lapply(predictions, function(P) {
+ Table_cost_sensitive <- table(test$Anaemia, P)
+ sum(diag(Table_cost_sensitive)) / sum(Table_cost_sensitive)
+ })
> # Display the accuracies for each cost value
> for (i in 1:length(cost_values)) {
+ cat('Cost Value:', cost_values[i], 'Accuracy:', accuracies[[i]], '\n')
+ }
Cost Value: 1 Accuracy: 0.547619
Cost Value: 2 Accuracy: 0.5238095
Cost Value: 3 Accuracy: 0.4761905
Cost Value: 4 Accuracy: 0.5
Cost Value: 5 Accuracy: 0.5
Cost Value: 6 Accuracy: 0.5
Cost Value: 7 Accuracy: 0.5238095
Cost Value: 8 Accuracy: 0.5238095
Cost Value: 9 Accuracy: 0.5238095
Cost Value: 10 Accuracy: 0.5238095

```

Results and Discussion for SVM model:

A number of Support Vector Machine (SVM) models that were trained by using various cost values were represented in the above R output. A cost value range of 1 to 10 was investigated in this study. For each cost value, an SVM algorithm was developed during the training loop and then stored in a list. The accuracy of these models is subsequently assessed on test dataset. From accuracy values the results show that how the cost value

selection affects the model's performance. Notably, the accuracy scores vary along with the pricing value. For instance, a cost value of 1 produced the test data's best accuracy, which was 0.5476. Further SVM analysis with cost 1 was as follows,

```
> # single SNM model with cost =1
```

```
> classifier_cost_sensitive <- svm(
```

```
+ formula = Anaemia ~ .,
```

```
+ data = train,
```

```
+ type = 'C-classification',
```

```
+ kernel = 'linear',
```

```
+ cost = 1)
```

```
> classifier_cost_sensitive
```

Call:

```
svm(formula = Anaemia ~ ., data = train, type = 'C-classification', kernel = 'linear',  
     cost = 1)
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: linear

cost: 1

No. of Support Vectors: 129

```
> summary(classifier_cost_sensitive)
```

Call:

```
svm(formula = Anaemia ~ ., data = train, type = 'C-classification', kernel = 'linear',  
     cost = 1)
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: linear

cost: 1

Number of Support Vectors: 129

(49 12 38 30)

Number of Classes: 4

Levels:

0 1 2 3

```
> pcc=predict(classifier_cost_sensitive, newdata=test)
```

```
> tc=table(pcc,test$Anaemia)
```

```

> tc
pcc 0 1 2 3
  0 7 5 0 0
  1 4 5 3 0
  2 1 4 6 0
  3 0 0 2 5
> accuracy=sum(diag(tc))/sum(tc)
> accuracy
[1] 0.547619
> # Confusion matrix
> confusion_matrix <- matrix(c(7, 5, 0, 0,
+                               4, 5, 3, 0,
+                               1, 4, 6, 0,
+                               0, 0, 2, 5),
+                               nrow = 4, byrow = TRUE)
> # Function to calculate precision, recall, and F1 score for each class
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FP <- rowSums(cm) - TP
+   FN <- colSums(cm) - TP
+
+   precision <- TP / (TP + FP)
+   recall <- TP / (TP + FN)
+   f1_score <- 2 * precision * recall / (precision + recall)
+
+   metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall =
recall, F1_Score = f1_score)
+   return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
  Class Precision Recall F1_Score
1    0 0.5833333 0.5833333 0.5833333

```



```

2 1 0.4166667 0.3571429 0.3846154
3 2 0.5454545 0.5454545 0.5454545
4 3 0.7142857 1.0000000 0.8333333

```

Results and discussion:

The simplest kind of kernel function is the linear kernel. Without any modifications, it represents the data in the original feature space. The linear kernel assumes that the data can be separated linearly with 54.76% accuracy. The resultant performance metrics shows precision, recall, and F1 score. For each class, recall examines the capacity to correctly identify instances of the class, while precision and recall are balanced to get the F1 score. Precision represents the accuracy of positive predictions. With a perfect recall and a moderately high F1 score, Class 3 sticks out as being exceptionally well predicted. Class 1 has poorer precision and recall lead to a lower F1 score, which suggests that its prediction may face some difficulties while doing class prediction. Classes 0 and 2 demonstrate balanced precision and recall with moderate F1 values, indicating acceptable model performance for these classifications. But There was still need of searching a algorithm which has maximum accuracy since SVM model with linear kernel gives only 55% accuracy. But this 55% accuracy not enough. So, move towards better machine learning algorithm. In the next section K-nearest neighbour algorithm was developed for prediction of anaemia in non-pregnant WRA.

6.5 K-Nearest Neighbour algorithm:

```

> # k-fold cross validation KNN
> trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
> mc1=train(Anaemia~., data=data, method='knn',trControl=trctrl)
> mc1

k-Nearest Neighbors
203 samples
52 predictor
4 classes: '0', '1', '2', '3'
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 183, 184, 183, 182, 182, 183, ...
Resampling results across tuning parameters:
  k Accuracy Kappa

```

```
5 0.4275097 0.1597688
```

```
7 0.4027728 0.1007303
```

```
9 0.4074880 0.1047216
```

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was $k = 5$.

```
> #Fitting KNN Model to training dataset
```

```
> classifier_knn <- knn(train = train,test = test,cl = train$Anaemia,k = 5)
```

```
> classifier_knn
```

```
[1] 1 2 1 2 1 1 1 1 3 1 1 0 1 0 1 1 3 3 1 1 2 0 1 1 1 2 0 1 0 0 1 2 2 0 1 0 0 1 1 0 3
```

```
Levels: 0 1 2 3
```

```
> summary(classifier_knn)
```

```
0 1 2 3
```

```
10 21 6 4
```

```
> cm <- table(classifier_knn, test$Anaemia,)
```

```
> cm
```

```
classifier_knn
```

```
0 1 2 3
```

```
0 3 5 1 1
```

```
1 5 12 3 1
```

```
2 2 2 2 0
```

```
3 0 2 1 1
```

```
> # Confusion matrix
```

```
> confusion_matrix <- matrix(c(3, 5, 1, 1,
```

```
+           5, 12, 3, 1,
```

```
+           2, 2, 2, 0,
```

```
+           0, 2, 1, 1),
```

```
+           nrow = 4, byrow = TRUE)
```

```
> # Function to calculate precision, recall, and F1 score for each class
```

```
> calculate_metrics <- function(cm) {
```

```
+ TP <- diag(cm)
```

```
+ FP <- rowSums(cm) - TP
```

```
+ FN <- colSums(cm) - TP
```

```
+ 
```

```
+ precision <- TP / (TP + FP)
```

```

+ recall <- TP / (TP + FN)
+ f1_score <- 2 * precision * recall / (precision + recall)
+
+ metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall = rec
all, F1_Score = f1_score)
+ return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
  Class Precision  Recall F1_Score
1    0 0.3000000 0.3000000 0.3000000
2    1 0.5714286 0.5714286 0.5714286
3    2 0.3333333 0.2857143 0.3076923
4    3 0.2500000 0.3333333 0.2857143
> misClassError=mean(classifier_knn != test$Anaemia)
> print(paste('Accuracy =', 1-misClassError))
[1] 'Accuracy = 0.439024390243902'

```

Results and discussion:

The evaluation of a k-nearest-neighbours (KNN) algorithm shows that with 'k' values of 5, 7, and 9, the model was trained and evaluated using 10-fold cross-validation. The 'final model with chosen 'k' was 5, which produced an accuracy of about 43.9% on a test dataset. The model's performance is shown in the confusion matrix for each class, with variations in precision, recall, and F1 score between classes. Notably, class 1 had the highest precision, recall, and F1 score, which suggests that its predictive abilities were higher. The final product offers a thorough analysis of the KNN model's performance in multiclass classification, showing both its advantages and disadvantages for various class types.

From all the three algorithms it was found that there was need of ensemble techniques since accuracies are not much acceptable to predict the anaemia among WRA. Ensemble techniques were developed in the next section.

6.6 Ensemble techniques to boost the model performance.

All previously developed models show lower performance as accuracy doesn't exceed to 55%. It is the indication that we should move forward for ensemble technique

to enhance the performance. In the following subsection bagged decision tree with 100 trees were developed.

6.6.1 Bagged decision tree

As explained in chapter 5, the first ensemble technique was used here is bagged decision tree. The bagged decision tree with 100 trees were developed following is the R output:

```
> bagged.tree <- bagging(Anaemia ~ ., data = train, nbagg = 100, coob = TRUE, control = rpart.control(maxdepth = 2, minsplit = 1))
```

```
> bagged.tree
```

Bagging classification trees with 100 bootstrap replications

Call: bagging.data.frame(formula = Anaemia ~ ., data = train, nbagg = 100, coob = TRUE, control = rpart.control(maxdepth = 2, minsplit = 1))

Out-of-bag estimate of misclassification error: 0.5988

Results and Discussion:

100 bootstrap samples were used to train the model. The misclassification error's 'out-of-bag' (OOB) estimate was determined. The OOB values tells that model misclassified on an average 59.88% of the observations. The value of this measure gives an estimate of the model's potential performance on hypothetical data and indicates that the model's prediction accuracy may be constrained. This algorithm was not much helpful to predict the Anaemia among married Non-pregnant WRA. By considering accuracy we can't move forward with this bagged decision tree model so another ensembling techniques should be used. Random forest algorithm was developed in the next section their results are as follows:

6.6.2 Random Forest Algorithm:

```
> rf_model = randomForest(Anaemia ~ ., data = train, ntree = 100)
```

```
> rf_model
```

Call:

```
randomForest(formula = Anaemia ~ ., data = train, ntree = 100)
```

Type of random forest: classification

Number of trees: 100

No. of variables tried at each split: 7

OOB estimate of error rate: 45.68%

Confusion matrix:

```
0 1 2 3 class.error
```

```

0 29 19 3 0 0.43137255
1 18 34 8 1 0.44262295
2 7 15 11 2 0.68571429
3 0 0 1 14 0.06666667

```

```
> varImp(rf_model)
```

```

Overall
HIV.status          0.6321990
Are.you.feeling.weak.or.dizziness.  7.1883984
Age                 4.1282307
Education.years.   1.7023459
Occupation          0.7616753
Income.of.the.family...Rs..Annual.  3.9850836
weight.kg.         5.5395817
Height..meter.     5.3704947
BMI                 6.3819006
Eating.Habits      2.4672410
Food.type          0.8379850
Daily..Tea.intake.. 0.9300423
Acidity.Problem..  1.0802187
Age.at.the.marriage 4.8088733
husband.s.age      4.3539488
husband.s.age.at.marriage 4.9234612
Husband.s.Occupation 2.4444309
husband.s.education..in.years.  2.0442128
Alcohol.Consumption.. 0.9190398
Any.Addiction..    0.8994226
Type.of.Addiction  0.9273560
Suffer.from.any.long.term.disease.. 0.6241904
Suffer.from.stress.. 1.1044719
use.Iron.supplementation.. 0.8565040
Suffers.from.Diabetes 0.2891170
Household.Wealth.status.. 0.4882331
Number.of.family.members 3.5766064

```

Toilet.facility..	0.3362332
Drinking.water.source..	1.7712440
Cooking.fuel	1.9800934
Exposure.to.domestic.violence..	0.9351991
Avg.of.rest.in.day..per.Hr...	2.0897751
Regular.visit.to.doctor..	0.5014168
Daily.eat.fresh.fruits.Vegetable..Milk..	0.5639153
Menstrual.cycle.1..	1.6347267
Menstrual.cycle.2..	0.4636023
No.of.pads...per.day.	2.2680646
days.of.blood.flow..	3.3250177
Pain.on.menstrual.period..	1.9465526
Age.at..menstrual.cycle.begins	2.7966764
Total.number.of.children.ever.born	1.7249380
Premature.Delivery..	0.6394819
Miscarage.History..	0.8491331
Age.at.first.birth.of.child..	4.1392771
Age.of.last..children..month...	6.3122273
No..of.births.in.last.5.years	1.5793963
Use.of.Contraceptive	0.9478507
Method.of.Contraceptive	0.8319930
Region	1.0675892
Number.of.years.lives.in..residential.area.	3.7837561
Mass.media.exposure	0.7783506
Community.women.education	1.0863729

Variable Importance Plot:

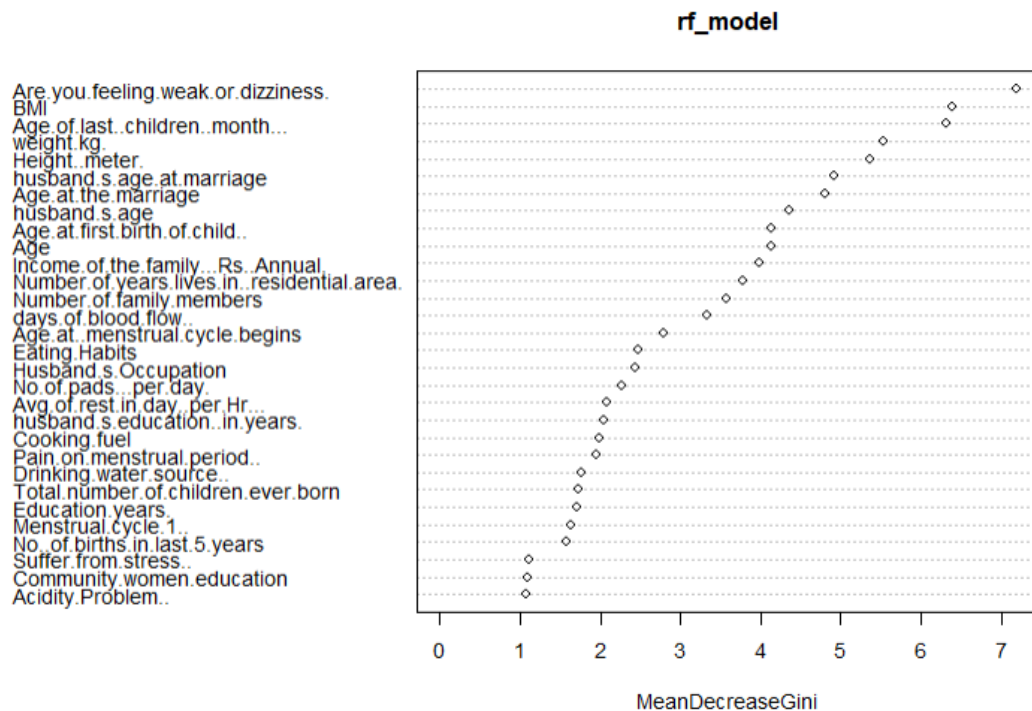


Fig. 6.4 Variable importance by RF classifier

Results and discussion :

The classification model created using random forests appears to be an advanced ensemble model made up of 100 decision trees. A number of features evidently plays an important roles in formulating predictions when the variable importance was examined. Notably, ‘Are you feeling weak or dizzy’, ‘BMI’, ‘Age of the last child (months)’, ‘Weight (kg)’, ‘Height (in metres), ‘ Husband’s age at the marriage’, Age at the marriage,’ ‘Husband’s age,’ ‘age at first birth of child’, ‘age’, ‘ family income’ etc were among the most important factors associated with status of anaemia. The model’s ability to correctly classify instances depends mostly on these attributes. Nevertheless, the model’s out-of-bag error rate is roughly 45.68%, indicating that it may not perform very well on unobserved data despite its complexity and these important qualities. It’s true that when used with the test dataset, the accuracy was found to be approximately 54%. It may be necessary to further refine the model, build new features, or experiment with new methods to increase predicted accuracy. Now the task is to find out one more ensembling technique for prediction of anaemia which may be better than random forest. So, the ADA boost algorithm was developed in the next section.

6.6.3 ADA Boost Ensemble technique:

```
> # Train an AdaBoost.M1 model
> ada_model <- boosting(Anaemia ~ ., data = train, boos = TRUE, mfinal = 100)
$call
boosting(formula = Anaemia ~ ., data = train, boos = TRUE, mfinal = 100)
$terms
Anaemia ~ HIV.status + Are.you.feeling.weak.or.dizziness. + Age +
  Education.years. + Occupation + Income.of.the.family...Rs..Annual. +
  weight.kg. + Height..meter. + BMI + Eating.Habits + Food.type +
  Daily..Tea.intake.. + Acidity.Problem.. + Age.at.the.marriage +
  husband.s.age + husband.s.age.at.marriage + Husband.s.Occupation +
  husband.s.education..in.years. + Alcohol.Consumption.. +
  Any.Addiction.. + Type.of.Addiction + Suffer.from.any.long.term.disease.. +
  Suffer.from.stress.. + use.Iron.supplementation.. + Suffers.from.Diabetes +
  Household.Wealth.status.. + Number.of.family.members + Toilet.facility.. +
  Drinking.water.source.. + Cooking.fuel + Exposure.to.domestic.violence.. +
  Avg.of.rest.in.day..per.Hr... + Regular.visit.to.doctor.. +
  Daily.eat.fresh.fruits.Vegetable..Milk.. + Menstrual.cycle.1.. +
  Menstrual.cycle.2.. + No.of.pads...per.day. + days.of.blood.flow.. +
  Pain.on.menstrual.period.. + Age.at..menstrual.cycle.begins +
  Total.number.of.children.ever.born + Premature.Delivery.. +
  Miscarage.History.. + Age.at.first.birth.of.child.. + Age.of.last..children..month... +
  No..of.births.in.last.5.years + Use.of.Contraceptive + Method.of.Contraceptive +
  Region + Number.of.years.lives.in..residential.area. + Mass.media.exposure +
  Community.women.education
```

\$importance

Acidity.Problem..	Age
0.19667957	4.50916123
Age.at..menstrual.cycle.begins	Age.at.first.birth.of.child..
1.86721460	3.53562668
Age.at.the.marriage	Age.of.last..children..month...
5.50718486	6.47535763
Alcohol.Consumption..	Any.Addiction..

1.11517139	0.48322267
Are.you.feeling.weak.or.dizziness.	Avg.of.rest.in.day..per.Hr...
10.91022111	2.85359650
BMI	Community.women.education
6.67450545	0.80669993
Cooking.fuel	Daily..Tea.intake..
2.50613625	0.33437516
Daily.eat.fresh.fruits.Vegetable..Milk..	days.of.blood.flow..
0.20608688	1.93568449
Drinking.water.source..	Eating.Habits
0.78564293	1.11119548
Education.years.	Exposure.to.domestic.violence..
0.31443835	0.31466839
Food.type	Height..meter.
0.07575005	9.84235944
HIV.status	Household.Wealth.status..
1.10132610	0.08936922
husband.s.age	husband.s.age.at.marriage
3.29099155	2.87027898
husband.s.education..in.years.	Husband.s.Occupation
0.79303928	0.93559842
Income.of.the.family...Rs..Annual.	Mass.media.exposure
2.49476788	0.78546761
Menstrual.cycle.1..	Menstrual.cycle.2..
1.04587293	0.03275416
Method.of.Contraceptive	Miscarage.History..
0.49677521	0.35908069
No..of.births.in.last.5.years	No.of.pads...per.day.
1.53294706	0.88030187
Number.of.family.members	Number.of.years.lives.in..residential.area.
1.56617456	5.13108724
Occupation	Pain.on.menstrual.period..
0.00000000	1.37962751
Premature.Delivery..	Region

0.26937700	0.66535174
Regular.visit.to.doctor..	Suffer.from.any.long.term.disease..
0.29698196	0.26767678
Suffer.from.stress..	Suffers.from.Diabetes
1.18924269	0.00000000
Toilet.facility..	Total.number.of.children.ever.born
0.17447644	0.81761453
Type.of.Addiction	use.Iron.supplementation..
0.57507372	1.58431675
Use.of.Contraceptive	weight.kg.
0.72567567	6.28777337

Interpretation:

The factor ‘feeling weak or dizziness’ stands out prominently with the highest importance score of 10.91, indicating that this aspect significantly influences the model’s predictions. This suggests that individuals experiencing feelings of weakness or dizziness play a crucial role in the model’s ability to make accurate predictions of anaemia.

Following closely is the factor ‘Height’ with an importance score of 9.84. This underscores the significance of height in the model, suggesting that taller or shorter individuals may exhibit patterns that strongly contribute to the prediction of anaemia in Non-pregnant WRA

Other notable variables include ‘age of last children born’ (6.47), ‘BMI’ (6.67), and ‘weight’ (6.28). These variables highlight the importance of physiological factors, such as age-related metrics, body mass index, and weight, in influencing the status of anaemia.

The variable ‘age at marriage’ (5.51) and ‘Number of years lives in residential area’ (5.1311) also hold substantial importance, suggesting that the age at which individuals get married and the duration of residency in a particular area significantly impact the model’s outcomes.

Variables like ‘age’ (4.5), ‘age at first birth of child’ (3.54), and ‘husband’s age’ (3.29) are moderately important, indicating their relevance but with less influence compared to the aforementioned factors.

On the other hand, variables like ‘husband’s age at marriage’ (2.87), ‘family income’ (2.4947), and ‘cooking fuel’ (2.51) have lower importance scores, implying a relatively lesser impact on the model’s predictive capabilities.

In conclusion it was discovered that physiological factors such as feelings of weakness or dizziness, height, BMI, and weight are highly influential in the model’s predictions. Additionally, demographic factors like residency, marital factors like age, age at marriage, and household level factors like income of family, cooking fuel used of also play significant roles. Understanding these influential variables provides valuable insights into the factors that contribute most to the model’s decision-making process in the context of the data under consideration.

```
> p=predict(ada_model, newdata = test, type='class')
> p
$class
[1] '3' '2' '1' '1' '1' '1' '0' '0' '2' '1' '0' '0' '0' '1' '1' '1' '1' '1' '0' '0' '0' '2' '1'
'1' '1'
[26] '1' '1' '1' '2' '1' '1' '1' '2' '2' '0' '1' '2' '2' '1' '1' '2'
$confusion
      Observed Class
Predicted Class 0 1 2 3
      0 4 5 0 0
      1 6 12 4 0
      2 0 4 3 2
      3 0 0 0 1
$error
[1] 0.5121951
> # Confusion matrix
> confusion_matrix <- matrix(c(4, 5, 0, 0,6, 12, 4, 0, 0, 4, 3, 2,0, 0, 0, 1), nrow = 4,
byrow = TRUE)
> # Function to calculate precision, recall, and F1 score for each class
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FP <- rowSums(cm) - TP
+   FN <- colSums(cm) - TP
+ }
```

```

+   precision <- TP / (TP + FP)
+   recall <- TP / (TP + FN)
+   f1_score <- 2 * precision * recall / (precision + recall)
+
+   metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall =
recall, F1_Score = f1_score)
+   return(metrics)
+   }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
  Class Precision  Recall F1_Score
1    0 0.4444444 0.4000000 0.4210526
2    1 0.5454545 0.5714286 0.5581395
3    2 0.3333333 0.4285714 0.3750000
4    3 1.0000000 0.3333333 0.5000000

```

Results and discussion:

According to the precision, recall and F1 score following conclusions were made, For Class 0, the precision is 44.44%, suggesting that out of all instances predicted as Class 0, only 44.44% are true positives. The recall for Class 0 is 40.00%, indicating that the model correctly identifies 40.00% of all actual Class 0 instances. The F1 score, a harmonic mean of precision and recall, is 42.11%, reflecting a balanced performance in capturing true positives while minimizing false positives.

Moving to Class 1, the model demonstrates better precision at 54.55%, meaning over half of the instances predicted as Class 1 are true positives. The recall for Class 1 is 57.14%, indicating that the model captures 57.14% of all actual instances of Class 1. The F1 score for Class 1 is 55.81%, suggesting a balanced performance between precision and recall.

For Class 2, the model exhibits a precision of 33.33%, signifying that one-third of the instances predicted as Class 2 are true positives. The recall for Class 2 is 42.86%, indicating that the model identifies 42.86% of all actual Class 2 instances. The F1 score for Class 2 is 37.50%, reflecting a moderate balance between precision and recall.

In contrast, Class 3 shows perfect precision at 100.00%, implying that all instances predicted as Class 3 are true positives. However, the recall for Class 3 is 33.33%,

indicating that the model only captures 33.33% of all actual instances of Class 3. The F1 score for Class 3 is 50.00%, suggesting a trade-off between precision and recall. In summary, the model exhibits varying degrees of performance across the four classes, with strengths in some classes (e.g., Class 1) and potential areas for improvement in others (e.g., Class 0 and Class 2). These insights can guide further model refinement and optimization efforts, potentially addressing specific challenges associated with each class to enhance overall predictive accuracy.

From the above ADA algorithm, it was found that error=0.5121951. The fraction of examples that were incorrectly classified is used to calculate the model's error rate, which is represented by error. The error rate in this instance was found to be 0.5121951, meaning that the model incorrectly classified approximately 51% of the cases. So the accuracy was 49%.

As from the accuracy, ADA boost and random forest gives same accuracy the variable importance table was compared and it was discovered that age, BMI, number of days of blood flow during periods, age at which menstrual cycle begins, height of the WRA, eating habits, Cooking fuel, Exposure to domestic violence, HIV status and average rest in day were common in both the algorithms. At the conclusion we can say that these factors are mostly affects the status of anaemia in WRA.

6.7 Comparison of Machine learning algorithms by Accuracy:

Table 6.1 Comparison of Machine learning algorithms by Accuracy.

Sr. No.	Machine Learning Algorithm	Accuracy
1	Decision tree	51.22%
2	support vector machine(linear)	55.00%
3	support vector machine (Polynomial)	36.00%
4	support vector machine (Sigmoid)	36.00%
5	support vector machine (Radial)	40.00%
6	K- nearest Neighbour (with k=7)	44.00%
7	Bagged Decision tree(with nbag=100)	40.12%
8	Random Forest Algorithm (with 100 trees)	54.00%
9	Ada Boost (mfinal=100)	49.00%

The table presents the accuracy results for machine learning algorithms applied to a given dataset. The DT model yielded an accuracy of 51.22%, indicating moderate predictive performance. The linear support vector machine achieved a slightly higher accuracy of 55.00%, suggesting a modest improvement in predictive power. However, the polynomial and sigmoid kernel-based support vector machines, as well as the radial

kernel variant, displayed lower accuracies ranging from 36.00% to 40.00%, suggesting challenges in capturing the underlying patterns in the data.

The K-nearest neighbour algorithm with $k=7$ achieved an accuracy of 44.00%, indicating a moderate level of predictive capability. The bagged decision tree model, employing 100 trees, yielded an accuracy of 40.12%, which appears lower than some other models. The random forest algorithm, comprising 100 trees, showed a 54.00% accuracy, suggesting a decent predictive performance. Lastly, the Ada Boost algorithm with $m_{final}=100$ resulted in an accuracy of 49.00%.

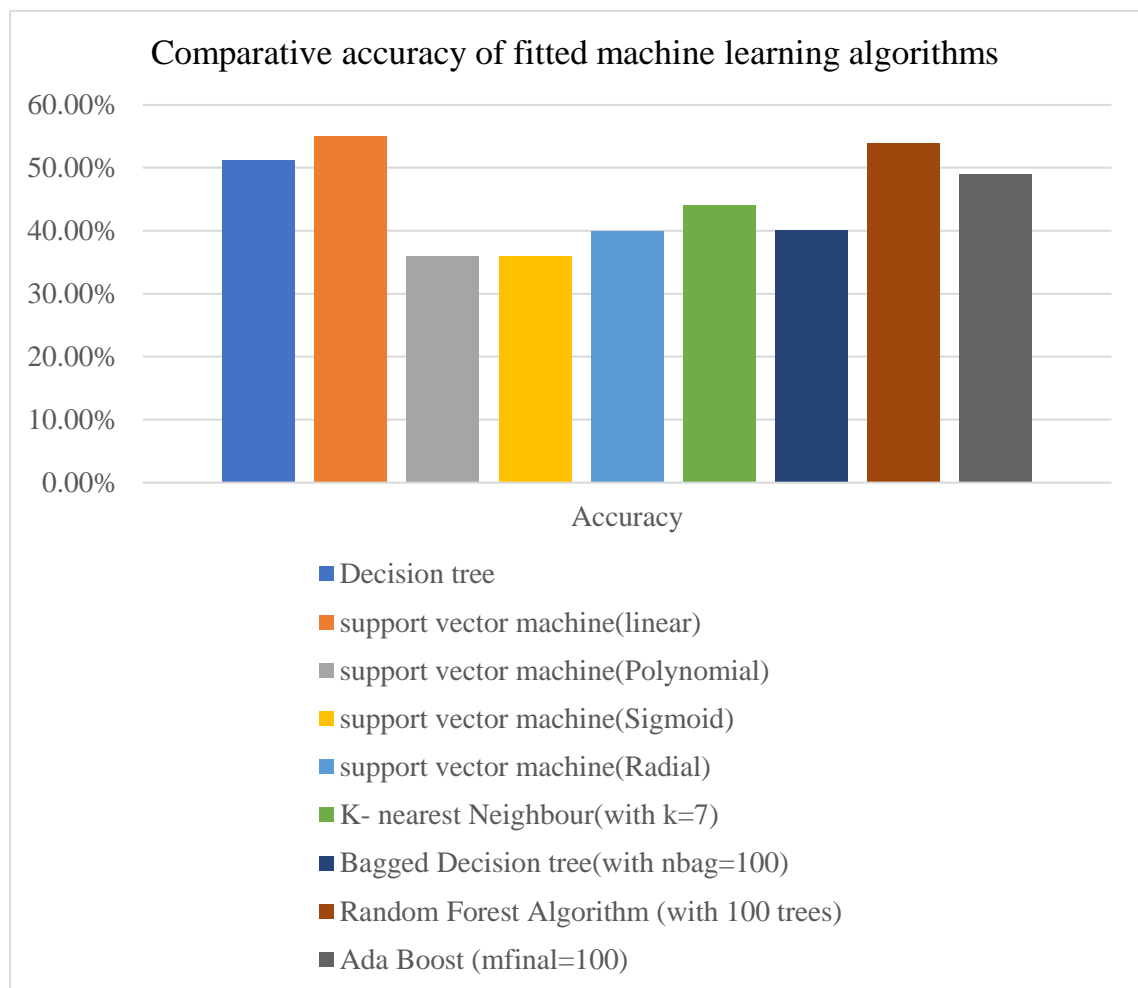


Fig. 6.5 Comparative accuracy plot of fitted machine learning algorithms

Results and discussion:

From all the developed machine learning algorithms it can be conclude that no machine learning algorithm gives good performance to predict anaemia among Non-pregnant married WRA. Therefore, to make better predictions there was need of some advanced ensemble technique. In the next section stacking ensemble model was developed to enhance predictive capacity.

6.8 Stacking Ensemble algorithm.

In an effort to boost predictive accuracy, a stacking ensemble approach was employed by combining the outputs of previously developed machine learning algorithms. This ensemble includes decision tree, linear support vector machine, K-nearest neighbour with $k=7$, bagged decision tree with 100 bags, random forest with 100 trees, and Ada Boost with $m_{final}=100$.

6.8.1 Base models:

On the first stage 6 base machine learning algorithms were developed. The new data set was created by using these machine learning algorithms. The new data set contains 6 predictor variables and one response variable that is anaemia status. The 6 predictors are nothing but the predations of 6 machine learning algorithm. The results are as follows:

```
> set.seed(123)
> train_indices <- sample(1:nrow(non), 0.8 * nrow(non))
> train_data <- non[train_indices, ]
> test_data <- non[-train_indices, ]
> train=data.frame( train_data)
> test=data.frame(test_data)
> non$Anaemia=as.factor(non$Anaemia)
> train$Anaemia=as.factor(train$Anaemia)
> test$Anaemia=as.factor(test$Anaemia)
> trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
> mcd=train(Anaemia~., data=train, method='rpart',trControl=trctrl)
> mcd
CART
162 samples
52 predictor
4 classes: '0', '1', '2', '3'
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 147, 145, 145, 145, 146, 146, ...
Resampling results:
cp      Accuracy  Kappa
0.04290429 0.3974510 0.14747817
0.06435644 0.4401471 0.21171816
```

0.14851485 0.3888971 0.03793195

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was $cp = 0.06435644$.

```
> mcs1=train(Anaemia~., data=train, method='svmLinear',trControl=trctrl)
```

```
> mcs1
```

Support Vector Machines with Linear Kernel

162 samples

52 predictor

4 classes: '0', '1', '2', '3'

Summary of sample sizes: 147, 146, 146, 144, 145, 146, ...

Resampling results:

Accuracy Kappa

0.500049 0.2980569

Tuning parameter 'C' was held constant at a value of 1

```
> mck=train(Anaemia~., data=train, method='knn',trControl=trctrl)
```

```
> mck
```

k-Nearest Neighbors

162 samples

52 predictor

4 classes: '0', '1', '2', '3'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 146, 144, 146, 147, 146, 146, ...

Resampling results across tuning parameters:

k Accuracy Kappa

5 0.3191095 0.01158928

7 0.3829085 0.09309576

9 0.3529984 0.04245834

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was $k = 7$.

```
> mrf=train(Anaemia~., data=train, method='rf',trControl=trctrl)
```

```
> mrf
```

Random Forest

162 samples

52 predictor

4 classes: '0', '1', '2', '3'

Summary of sample sizes: 145, 146, 147, 144, 146, 146, ...

Resampling results across tuning parameters:

mtry	Accuracy	Kappa
2	0.5552124	0.3354578
27	0.5668382	0.3855162
52	0.5782761	0.3985030

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 52.

```
> mbag=train(Anaemia~., data=train, method='treebag',trControl=trctrl)
```

```
> mbag
```

Bagged CART

162 samples

52 predictor

4 classes: '0', '1', '2', '3'

Summary of sample sizes: 145, 145, 146, 145, 146, 147, ...

Resampling results:

Accuracy	Kappa
0.5691667	0.3896811

```
> mboost=train(Anaemia~., data=train, method='gbm',trControl=trctrl)
```

```
> mboost
```

Stochastic Gradient Boosting

162 samples

52 predictor

4 classes: '0', '1', '2', '3'

Summary of sample sizes: 145, 144, 145, 146, 146, 147, ...

Resampling results across tuning parameters:

interaction.depth	n.trees	Accuracy	Kappa
1	50	0.5288971	0.3212541
1	100	0.5163971	0.3094130
1	150	0.5275163	0.3361379
2	50	0.5387745	0.3470733
2	100	0.5709069	0.3907764

2	150	0.5691013	0.3864208
3	50	0.5498775	0.3608709
3	100	0.5352451	0.3386648
3	150	0.5292320	0.3282518

Tuning parameter ‘shrinkage’ was held constant at a value of 0.1

Tuning parameter ‘n.minobsinnode’ was held constant at a value of 10

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were n.trees = 100, interaction.depth = 2, shrinkage = 0.1

and n.minobsinnode = 10.

6.8.2 Meta Model:

As the RF shows relatively acceptable accuracy than other models here RF was used as a meta model. There is no restriction on meta model. We can use any machine learning algorithm as a meta model. Here Random forest model used as a meta model.

```
> Pcd=as.integer(predict(mcd, newdata = test))
> Pcs1=as.integer(predict(mcs1, newdata = test))
> Pk=as.integer(predict(mck, newdata = test))
> Prf=as.integer(predict(mrf, newdata = test))
> Pbag=as.integer(predict(mbag, newdata = test))
> Pboost=as.integer(predict(mboost, newdata = test))
> stack_data=data.frame(test$Anaemia,Pcd,Pcs1,Pk,Prf,Pbag,Pboost);stack_data
```

	test.Anaemia	Pcd	Pcs1	Pk	Prf	Pbag	Pboost
1	3	2	4	2	4	4	4
2	2	2	2	2	3	3	3
3	1	2	3	3	2	3	2
4	1	1	3	2	1	2	2
5	1	2	4	2	3	3	2
6	0	2	1	2	2	2	2
7	1	1	2	2	1	1	2
8	1	1	1	1	1	1	1
9	1	2	2	3	2	2	3
10	1	2	2	2	2	2	2
11	0	1	1	2	1	2	1
12	0	1	1	1	1	1	1

```

13    0 1 2 3 2 2 1
14    1 2 2 2 2 2 2
15    0 1 2 2 2 2 2
16    2 2 2 2 2 2 2
17    1 2 2 2 2 3 2
18    2 2 2 2 2 2 2
19    0 1 2 2 1 1 1
20    1 1 2 2 1 1 2
21    1 1 1 1 1 1 2
22    1 2 3 2 3 3 3
23    1 2 2 3 2 3 3
24    1 2 2 2 2 3 2
25    1 2 2 2 2 2 2
26    0 1 2 2 2 2 2
27    1 1 2 3 2 2 2
28    1 2 2 2 2 2 2
29    1 2 3 1 3 3 3
30    2 2 2 1 2 3 2
31    2 2 2 2 2 2 2
32    0 1 2 3 2 2 2
33    2 2 3 3 3 3 3
34    3 2 1 1 3 3 3
35    1 1 1 2 1 1 1
36    0 2 2 1 2 2 2
37    1 2 3 1 2 3 2
38    2 2 3 2 3 3 3
39    1 2 2 2 2 2 2
40    0 2 1 1 2 2 1
41    3 2 1 1 3 3 3

```

```
> dim(stack_data)
```

```
[1] 41 7
```

```
> str(stack_data)
```

```
'data.frame': 41 obs. of 7 variables:
```

```
$ test.Anaemia: Factor w/ 4 levels '0','1','2','3': 4 3 2 2 2 1 2 2 2 2 ...
```

```

$ Pcd      : int 2 2 2 1 2 2 1 1 2 2 ...
$ Pcs1     : int 4 2 3 3 4 1 2 1 2 2 ...
$ Pk       : int 2 2 3 2 2 2 2 1 3 2 ...
$ Prf      : int 4 3 2 1 3 2 1 1 2 2 ...
$ Pbag     : int 4 3 3 2 3 2 1 1 2 2 ...
$ Pboost   : int 4 3 2 2 2 2 2 1 3 2 ...
> train_indices2 <- sample(1:nrow(stack_data), 0.8 * nrow(stack_data))
> train_data2 <- stack_data[train_indices2, ]
> test_data2 <- stack_data[-train_indices2, ]
> train=data.frame( train_data2)
> test=data.frame(test_data2)
> rfmeta_model =randomForest(train$test.Anaemia ~ ., data = train, ntree = 100)
> rfmeta_model
Call:
randomForest(formula = train$test.Anaemia ~ ., data = train, ntree = 100)

Type of random forest: classification
Number of trees: 100
No. of variables tried at each split: 2
OOB estimate of error rate: 31.25%

Confusion matrix:
 0  1  2  3 class.error
0 5  4  0  0  0.4444444
1 2 15  1  0  0.1666667
2 0  3  0  0  1.0000000
3 0  0  0  2  0.0000000

```

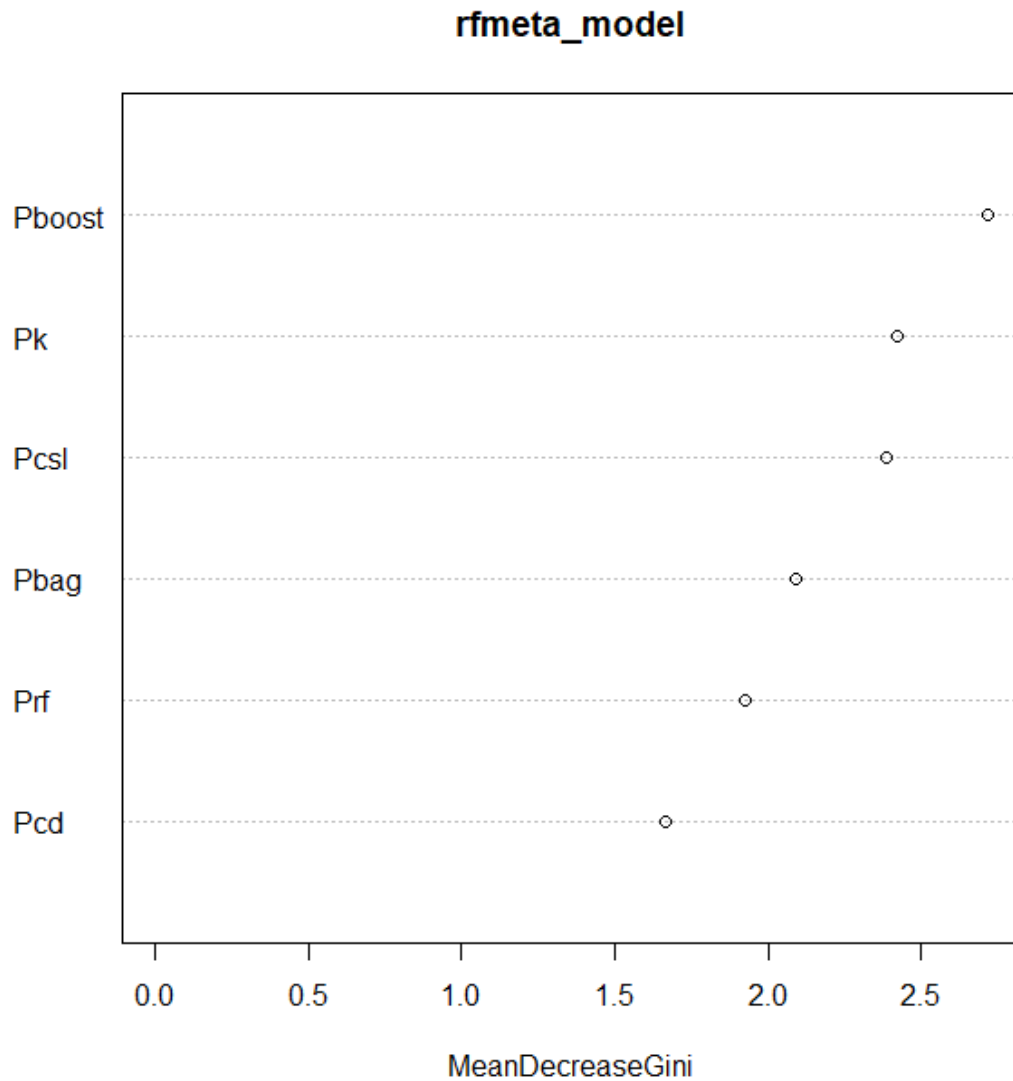


Fig. 6.6 Meta model Variable importance plot

As shown in the above plot the most significant factor was Pboost. So, to find the significant factors associated with anaemia status among the non-pregnant WRA the gradient boosting algorithm was used. As gradient boosting algorithm was previously developed so in the following section only variable importance were extracted. Output is as follows:

```
> varImp(mboost)
gbm variable importance
only 20 most important variables shown (out of 52)
```

	Overall
Are.you.feeling.weak.or.dizziness.	100.00
BMI	80.66

Number.of.years.lives.in.residential.area.	68.32
Age.at.the.marriage	64.20
Age.of.last.children.month...	53.30
Age	46.98
Height.meter.	45.90
weight.kg.	44.78
husbands.age.at.marriage	34.15
husband.s.age	32.68
Age.at.first.birth.of.child..	27.59
days.of.blood.flow..	23.51
Income.of.the.family...Rs..Annual.	23.13
Cooking.fuel	18.54
Husband.s.Occupation	17.80
Region	16.83
Age.at..menstrual.cycle.begins	16.66
Number.of.family.members	16.43
husband.s.education..in.years.	14.54
Alcohol.Consumption..	14.33

The Gradient Boosting Machine (GBM) model's variable importance analysis shows how important different predictors are in predicting the anaemia in non-pregnant WRA. Out of 52 predictors above table shows 20 most significant factors associated with status of anaemia among NPW. According to variable importance table here are some conclusions regarding influential factors of anaemia in non-pregnant WRA. With a score of 100.00, the attribute 'Are you feeling weak or dizzy?' appears to be the most influential factor in the model and is ranked highest on the relevance list. This implies that this specific characteristic has a significant role in the prediction of anaemia in non-pregnant WRA, meaning that those who report feeling weak or lightheaded may be more likely to have anaemia based on the dataset.

Factors including BMI (body mass index), number of years living in a residential area, age at marriage, and age of last child in months have importance score 80.66, 68.32, 64.20, 53.30 respectively. The model also shows that these variables are highly significant in predicting anaemia in non-pregnant WRA. With the importance scores 46.98,45.90, 44.78 factors like age, height and weight shows significant relation with anaemia status. Factors like husband's age, age at the first birth of child, number

of days of blood flow during periods, income of family shows significant correlation with anaemia with important scores 34.15, 32.68, 27.59, 23.51, 23.13 respectively. In the last but factors like cooking fuel, Husband’s occupation, Region, age at menstrual cycle begins, number of family members, husband’s education in years, and alcohol consumption shows importance score from 18.54 to 14.33.

It is noteworthy that these relevance scores are particular to the training data of this model and are relative. Greater significance values indicate more robust prediction ability in the context of this dataset. Based on the given dataset and model, these results suggest that certain physiological characteristics (like reported symptoms, BMI) and demographic data (like age-related factors, residential details) may have a considerable impact on the risk of anaemia incidence. To fully comprehend the intricate interactions between these factors and provide an accurate prediction of anaemia, more investigation and research may be required.

After examining the significant factors associated with stage of anaemia it is crucial to find pattern of these factors. In the following section various factors were studied deeply.

6.9 Relationship between significant factors and anaemia status:

Are you feeling weak or dizziness? Was found to be higher significant score than other so it was examined at the top of the list. The distribution of WRA according to their anaemia status and whether or not they experienced weakness or feeling dizzy is shown in the table. This is a contingency table that shows how the four categories of anaemia severity No anaemia, Mild, Moderate, and Sever relate to the variable ‘Feeling weak or dizzy’ which was categorical variable. It has two categories 0 stands for ‘no’ and 1 stands for ‘yes’.

Table 6.2 Feeling weak or dizziness over anaemia severity.

		Anaemia status				Total
		No anaemia	Mild	Moderate	Severe	
Feeling weak or dizziness	0	47	30	1	2	80
	1	18	52	41	16	127
	Total	65	82	42	18	207

The association between an individual’s anaemia status and their subjective experience of weakness or dizziness is presented in the above table. It was clear that the majority of WRA (65 out of 207) who did not have anaemia did not experience any weakness or dizziness. On the other hand, among WRA with different levels of

anaemia, weakness or dizziness seems to be more common as anaemia gets worse. In particular, nearly all (16 out of 18) of the people with severe anaemia reported feeling weak or dizziness. Similarly, for moderate anaemia there were total 42 moderate anaemia cases out of which 41 WRA have feeling weak and dizziness. In case of mild anaemia this ratio was slightly decrease but still there is also impact of feeling weak and dizziness on presence of anaemia. Therefore, at the last it was found that if the WRA feeling weak or dizziness then there is high probability of getting anaemic. This pattern points to a significant correlation between the prevalence of weakness or dizziness and the degree of anaemia.

The next factor was BMI. BMI was calculated from height and weight of WRA. Which was continuous variable. Therefore, to check the trend of BMI among various categories of anaemia average BMI was used. Following table shows the average BMI of non-pregnant WRA among various categories of anaemia.

Table 6.3 Average BMI among Anaemia.

	No anaemia	Mild	Moderate	Severe
Average of BMI	32.23677976	31.991624	31.015942	27.289871

The average Body Mass Index (BMI) for the four-anaemia status i.e No anaemia, mild, moderate, and severe was shown in the table. It was evident that the group with the greatest average BMI those without anaemia, at roughly 32.24 was followed by the group with moderate anaemia, at about 31.99. An average BMI of approximately 31.02 was slightly lower in moderate anaemia. Notably, the average BMI of those suffering from severe anaemia was the lowest, at approximately 27.29. According to this research, there may be a negative correlation between anaemia severity and BMI, meaning that the average BMI tends to fall as anaemia gets worse.

The next significant factor identified by gradient boosting was ‘Number of years lives in residential area’. This variable contains the number of years WRA lives in particular area. So, the variable is numeric. To study the pattern of relationship between anaemia status and this variable the average of Number of years lives in residential area was used.

Table 6. 4 Average of Number of years lives in residential area across anaemia.

	No anaemia	Mild	Moderate	Severe
On an average of No. of years lives in area.	11.76119403	10.672289	10.247619	8

From the above table some conclusions were made, as the anaemia severity increases from no anaemia to severe the average of number of years lives in residential area were decreases. WRA without anaemia have the highest average number of years lived in the residential area at approximately 11.76 years. Those with mild anaemia have a slightly lower average of around 10.67 years, followed by WRA with moderate anaemia at about 10.25 years. Notably, WRA with severe anaemia have found to be the lowest average number of years lived in the residential area, with an average of 8 years.

The next significant factor associated with anaemia status among non-pregnant WRA was age of respective WRA. The age variable was considered as age of WRA in years. So, it was numerical variable. To observe trend average was used. Following table shows the relationship between anaemia among WRA and age of WRA.

Table 6.5 Anaemia and average of Age at the marriage.

	No anaemia	Mild	Moderate	Severe
Mean of Age at the marriage	18.80597015	18.108434	18.52381	21.444444

There was no any decreasing or increasing trend but from the above table we can say that the severe anaemia was found at average age 21.44 years. So, we can say that older age WRA have high chance of severe anaemic.

Age of last children was also significant to the status of anaemia. Since the age of children was observed in months, the age of children also a numeric variable. To analyse trend or pattern of relationship the average was used.

Table 6. 6 Average of Age of last children (month) Vs Anaemia.

	No Anaemia	Mild	Moderate	Severe
A.M. of Age of last children (month)	17.375	9.861333333 3	12.5263157 9	24.8571428 6

Table shows the Average of Age of last children (month) with anaemia severity. Here no any pattern of relationship was found but WRA those have age of last child was approximately 25 months have chance of severe anaemic.

Age of WRA was found to be influential factor while predicting anaemia in non-pregnant WRA. Age variable is in years so it is numeric. To examine the relationship, trend the average of age WRA was used in the following table.

Table 6.7 Anaemia with average age of WRA.

	No Anaemia	Mild	Moderate	Severe
Average of Age	33.14925373	31.39759036	32.26190476	34

It was found that average age of WRA in No anaemia category was 33 years, for mild anaemia was 31 years, for moderate anaemia was 32 years and that of severe anaemia was 34 years. From the results unable to recognise any pattern between age of WRA and anaemia status.

Height of WRA in meter found to be one of the key factors associated with status of anaemia among NPW. To examine relationship, height was converted into meter to centimetre for more specification. The results are as follows:

Table 6.8 Anaemia with average height of WRA.

	No Anaemia	Mild	Moderate	Severe
Average of Height (cm)	153.3689552	151.74337	150.99143	157.72444

It was observed that the reproductive aged women who were not anaemic have average height on an average 153cm. The WRA who were found to be mild and moderate anaemia have average height approximately 152 and 151 centimetres respectively. The women those have severe anaemia found to be 158 cm average height. From these figures some conclusions can be made like short heighted women have high chance of mild and moderate anaemia, whereas heighted WRA have high chance of severe anaemia.

The weight is the main factor of health examination and in the research weight factor also shows significant importance related to the anaemia status for predicting anaemia in non-pregnant WRA. Weight of WRA measured in kg. Following table shows the average weight of WRA among various categories of anaemia.

Table 6.9 Anaemia with average weight of WRA.

	No Anaemia	Mild	Moderate	Severe
Average of weight(kg)	57.31940299	55.30120482	53	52.33333333

Results shows that the average weight of non-pregnant WRA who have no anaemia was approximately 57 kg. The WRA with mild and moderate anaemia have average weight 55 and 53 kg respectively. The WRA who were suffering from severe anaemia have average weight was 52 kg. The data presented here made it abundantly evident that there was a negative correlation between the severity of anaemia and the weight of the WRA. As we can see here as if weight decreases the anaemia severity increases from mild to severe. From this we can suggest that to avoid anaemia in non-pregnant WRA we have to aware WRA about the weight factor.

Not only individual but husband's related factors also impacts on the status of anaemia. Here the factor 'husband's age at marriage' found to be significant according

to gradient boost algorithm. husband age at marriage was numeric variable as the measured in years. To check the association pattern between anaemia and ‘husband’s age at marriage’ mean of age of husband at marriage time was used. Therefore, in the following table average husband’s age at marriage was observed for different stages of anaemia severity.

Table 6.10 Average of husband’s age at marriage with anaemia status.

	No Anaemia	Mild	Moderate	Severe
Average of husband’s age at marriage	25.59090909	24.21686 747	26.28571 429	26.77777 778

The WRA with no anaemia have average husband’s age at marriage was found to be 25 years. The WRA suffering from mild and moderate anaemia have average husband’s age at marriage was approximately 24 and 26 years respectively. Severely anaemic WRA whose average husband’s age at marriage found to be approximately 27 years. There was no such increasing or decreasing pattern but from severe category anaemia we can say that the WRA who marry at elder age of husbands have higher chance of moderate and severe anaemia.

Not only individual but also marital factor like husband’s age found to be significant. Husband’s age was numeric variable to investigate relationship pattern the average husband’s age was observed according to anaemia levels. Following table shows the average husband’s age at various levels of anaemia status.

Table 6. 11 Average age of husband (years) of WRA with anaemia.

	No Anaemia	Mild	Moderate	Severe
Average of husband’s age (years)	39.76119403	37.3493975 9	39.2380952 4	37.3333333 3

The non-pregnant WRA with no anaemia have husband’s age approximately 40 years. The WRA whose anaemia status ‘mild’ have husband’s average age was 37 years and that of ‘moderate’ anaemia have husband’s age nearly 39 years. The WRA with ‘severe’ anaemia have husband’s average age was 37 years. According to average there was no any pattern found. But still the husband’s age was important while predicting anaemia among non-pregnant WRA.

The children related factors were also impacting the health of women. Here in gradient boost algorithm the WRA’s age when she gives her first birth of child shows significant importance against anaemia status. This variable was numeric as age was measured in years. Following table displays the on an average age of WRA at first birth of child against several stages of anaemia.

Table 6.12 WRA's age at first birth of child versus anaemia.

	No Anaemia	Mild	Moderate	Severe
Average of Age at first birth of child	19.4328358 2	18.3614457 8	18.904761 9	18.7777777 8

There was no any trend or pattern among average age at first birth and anaemia severity but still we can say that some conclusion against the severity trend. The WRA having age at first birth of child was 19 years then she has no anaemia. It seems that there is a trend towards anaemia severity and WRA's age at first birth of child. WRA having their first child at slightly younger ages have possibly severe anaemia.

Having heavy or prolonged menstrual bleeding increases the risk of anaemia. A woman loses more blood and iron as a result of having a heavy menstrual flow. Which leads to iron deficiency anaemia. As the 'number of days of blood flow' found to be influential factor according to the best model which was obtained previously. So, there was important to analyse pattern. The variable 'number of days of blood flow' was numeric which takes value 1,2,3, up to 16. Actually there were only 2 cases were the blood flows up to 15-16 days otherwise in remaining cases blood flows up to 1,2,..8 days. So it was numeric variable. Excluding these 2 cases following table shows the mean of no. of days of blood flow during menstrual cycle.

Table 6.13 Average of number of days of blood flow according to anaemia classes.

	No Anaemia	Mild	Moderate	Severe
Average of days of blood flow	4	3.914285714	4.2	3.4

During their menstrual cycle, WRA without anaemia typically experience menstrual blood flow for four days. The average menstrual blood flow duration for those with mild anaemia was 3.91 days. The average menstrual blood flow length for WRA with moderate anaemia was approximately 4.2 days. The average menstrual blood flow duration for WRA with severe anaemia was approximately 3.4 days. Based on this statistics, it seems possible that the average number of days of menstrual blood flow varies slightly depending on the degree of anaemia.

Women's economic status can have a substantial impact on their health in a number of ways. The state of the economy has a big impact on women's health. Their general well-being can be negatively impacted and their susceptibility to health issues including malnutrition, inadequate prenatal care, and higher rates of preventable diseases can be increased by limited financial resources, which can also make it more difficult for them to get needed hygiene products, wholesome food, and high-quality

healthcare. According to machine learning model the family's annual income has significant effect on the state of anaemia. The Annual family income was measured in rupees so its numeric variable. The relationship was examined by using average annual income with respect to various categories of anaemia.

Table 6.14 Average family income with anaemia.

	No Anaemia	Mild	Moderate	Severe
Average family Income in year	194432.8358	161746.988	112619.0476	136111.1111

According to data values displays in the table some conclusions were made about the relationship pattern. Here the WRA without anaemia have average annual family income nearly 2 lacks. Whereas WRA experiencing Milds and moderate anaemia have average annual family income 1.7 and 1.12 lacks respectively. Those with severe anaemia have average annual family income approximately 1.36 lacks. From these figures it was clearly highlights that the income and anaemia severity have slightly inverse relationship. As income decreases the anaemia severity increases. There is a discernible relationship between women's anaemia severity and declining income levels. Inadequate financial means frequently impede the ability to obtain iron-rich foods, healthcare facilities, and a balanced diet, which greatly increases the incidence and severity of anaemia in lower-class communities.

A woman's health may be impacted by her husband's occupation through a number of different indirect channels. A woman's general quality of life, mental and physical health, and access to healthcare can all be negatively impacted by a number of factors, including her financial situation, stress at work, and lifestyle decisions affected by her husband's profession. Furthermore, some jobs may expose workers to environmental risks or have unpredictable work schedules, which may have an indirect impact on a woman's health. Hence identified best ML model gives next influential factor related to status of anaemia was 'Husband's occupation'. Husband's occupation was categorical variable it has four categories such as 0, 1, 2, and 3. Where 0 stands for farmer, 1 stands for labour, 2 stands for job and 3 stands for business. Since this factor is categorical with four different categories, the relationship was assessed by using simple frequency table. Following contingency table shows the frequency of anaemia category against the various types of 'husband's occupation'.

Table 6.15 Contingency table of anaemia and husband's occupation.

Husband's Occupation	Row Labels	Anaemia Status				Grand Total
		No Anaemia	Mild	Moderate	Severe	
	0	11	17	11	0	39
	1	10	14	8	4	36
	2	29	40	15	10	94
	3	17	12	8	4	41
	Grand Total	67	83	42	18	210

The above table illustrates the possible correlation between the occupation of a woman's husband and the presence of anaemia in the WRA. Among 39 WRA whose husbands were farmers, 11 have no anaemia, 17 have “mild anaemia”, 11 have “moderate anaemia”, and none have “severe anaemia”. For 36 WRA whose husbands working as labours, 10 have “no anaemia”, 14 have “mild anaemia”, 8 have “moderate anaemia”, and 4 have “severe anaemia”. Among 94 women whose husbands have a job, 29 have “no anaemia”, 40 have “mild anaemia”, 15 have “moderate anaemia”, and 10 have “severe anaemia”. There were total 41 women with husbands involved in business show that 17 have “no anaemia”, 12 have “mild anaemia”, 8 have “moderate anaemia”, and 4 have “severe anaemia”.

If we see the anaemia severity there were total 67 WRA without anaemia out of which 11 WRA have their husband's occupation was farmer, 10 WRA with husband's occupation labour, 29 WRA with their husband's occupation was job and 17 WRA with their husband's occupation was business. These numbers show that the distribution of husbands' occupations is not uniform among women who do not have anaemia. It was high in firstly job and secondly business category. It can be seems that the WRA whose husbands are employee or businessmen have chances of no anaemia.

There were 83 cases of mild anaemia overall, of which 17 WRA had their husbands working as farmers, 14 WRA had their husbands working as labourers, 40 WRA had their husbands working as jobs, and 12 WRA had their husbands working as businessmen. It seems that a relatively higher proportion of WRA whose husbands had a job experienced mild anaemia compared to the total number of cases within this group. This implies an association between the husband's occupation (specifically a job) and a higher likelihood of the wives experiencing mild anaemia.

For moderate anaemia category there were total 42 WRA, of which 11 WRA had their husbands working as farmers, 8 WRA had their husbands working as labourers, 15

WRA had their husbands working as jobs, and 8 WRA had their husbands working as businessmen. And for severe anaemia there were only 18 cases of severe anaemia overall, of which 0 WRA had their husbands working as farmers, 4 WRA had their husbands working as labourers, 10 WRA had their husbands working as jobs, and 4 WRA had their husbands working as businessmen. According to figures it was seen that high proportion of severe anaemic in the category job. That is the WRA whose husband's occupation was job have high proportion of severe anaemic than other.

The differences between rural and urban areas have a big impact on the health of women. Women frequently experience difficulties in rural locations because of restricted access to medical facilities, which can lead to a shortage of healthcare professionals and specialised services. Women's health can also be impacted by the socioeconomic gaps, poor infrastructure, and limited educational possibilities that are common in many rural areas. On the other hand, women may have better health outcomes in urban areas because they have greater access to jobs, educational opportunities, and healthcare. Urban environments can, however, also expose women to lifestyle-related health problems, such as elevated stress levels, sedentary habits, and heightened exposure to environmental contaminants, which can lead to a distinct range of health difficulties. Within the next section, an examination of the anaemia status and the region of residency was held. Rural and urban areas are the two groups that make up the categorical variable known as region. The following table presents a contingency analysis that illustrates the distribution of anaemia severity according to geography.

Table 6. 16 Contingency table of Region and anaemia.

		Anaemia Status				Grand Total
		No Anaemia	Mild	Moderate	Severe	
Region	Rural	62	69	37	12	180
	Urban	5	14	5	6	30
Grand Total		67	83	42	18	210

The above 2×4 contingency table shows the distribution of anaemia categories according to the region. Here there are total 180WRA from rural region out of which 62 have “no anaemia”, 69 have “mild anaemia”, 37 have “moderate anaemia” and 12 have “severe anaemia”. It was good indication that in rural region anaemia severity was not serious. In case of Urban area there WRA total 30 WRA out of which only 5 WRA have no anaemia remaining 25 shows varying degree of anaemia. That is 14 WRA have “mild anaemia”, 5 WRA have “moderate anaemia” and 6 WRA have “severe anaemia”.

The final findings indicates that anaemia appears to be a more prevalent concern among WRA in urban areas compared to those in rural settings. This tendency may be related to a number of urban-related issues, such as eating patterns that may be deficient in important minerals, such iron, which raises the risk of anaemia. Furthermore, urban living frequently entails higher stress levels, sedentary behaviours, and maybe restricted access to wholesome, fresh food all of which can have an adverse effect on general health, including the prevalence of anaemia. This finding emphasises the necessity for focused treatments and healthcare programmes designed with urban populations in mind in order to address and lessen the greater rate of prevalence of anaemia among women living in metropolitan regions.

Menarche, the age at which a woman starts her menstrual cycle, has a major influence on her health in a number of ways. An increased chance of developing various health disorders later in life, including obesity, type 2 diabetes, heart disease, and some cancers like breast and ovarian cancer, has been linked to an early menarche. An earlier menarcheal age may also bring with it psychological and emotional difficulties, such as an increased risk of mental health problems including anxiety and depression. On the other hand, a late menarche may also signal underlying medical issues that need to be addressed and may potentially have an adverse effect on fertility and bone health. In the next table Age at menstrual cycle begins examined in accordance with different anaemia levels of WRA.

Table 6. 17 Average of Age at menstrual cycle begins with anaemia.

	No Anaemia	Mild	Moderate	Severe
Average of Age at menstrual cycle begins	13.97014925	13.9759036 1	13.6190476 2	13.66666666 7

For no anaemia and mild anaemia, the average age at menstrual cycle begins was approximately 14 years. And those for moderate and severe anaemia have nearly 13.61 years. Here unable to make conclusion about the pattern of relationship since there was no much variation in the anaemia categories.

Women's health can be impacted by the number of family members in a number of ways. Larger families may be more susceptible to resource strain, particularly in lower-income environments. This could result in restricted access to proper education, healthcare, and nutrition, all of which can have a negative impact on a woman's health. On the other hand, benefits like better healthcare, more individualised attention, and easier access to resources may come with smaller family numbers. Less stress from

carrying for family members and other responsibilities may be experienced by women in smaller households, which may be beneficial to their mental health. Additionally, smaller families might be able to better plan and space out their pregnancies, which would benefit the health of both mothers and their offspring. To study this relationship average of Number of family members were examined in four categories of anaemia severity in non-pregnant WRA.

Table 6.18 Average of family size with anaemia.

	No Anaemia	Mild	Moderate	Severe
Average of Number of family members	5.149253731	4.7710843 37	5.1666666 67	5.2222222 22

The above table doesn't show any indication as discussed in the above paragraph. Average number of family members approximately same over all classes of anaemia. Though 'number of family members' factor was significant in ML algorithm but it has low importance score than other variables.

Previous studies have shown that a woman's health can be positively impacted by her husband's educational attainment. Higher educated males tend to have more understanding about health-related issues, and better socioeconomic circumstances, all of which improve the women's overall health results. Since the 'Husband's education' found to be significant factor according to ML model, 'Husband's education' thoroughly examined according to various levels of anaemia. 'Husband's education' was categorical variable which has 5 categories such as 0,1,2,3,4. Where the categories are as follows:

Primary=(0-5)=0

Secondary=(6-10)=1

Higher secondary =(11-12)=2

Graduate=(13-15)=3

Post graduate= 4

Following table shows the frequency distribution of 'Husband's education' with different stages of anaemia.

Table 6. 19 Frequency table of husband's education and anaemia status.

		Anaemia Status				
		"No Anaemia"	"Mild"	"Moderate"	"Severe"	Grand Total
	0	0	4	4	0	8
	1	24	25	15	12	76

husband's education (in years)	2	28	27	11	4	70
	3	8	20	7	0	36
	4	7	4	5	2	18
	Grand Total	67	80	42	18	210

There were total 8 WRA whose husband's education was primary. Out of that 8 WRA 4 were mild and 4 were moderate anaemic. 76 WRA having husband's education level was secondary out of them 24 were no anaemic, 25 were mild, 15 were moderate and 4 were severe anaemic. Among 70 WRA whose husband's education level was higher secondary, 28 were no anaemia, 27 were mild, 11 were moderate and 4 were severe anaemic. Total 36 WRA of which husbands were graduate of which 8 were no anaemic, 20 were mild, 7 where moderate. There are 18 WRA with husbands are post graduate, out of them 7 are non-anaemic, 4 are mils, 5 are moderate and 2 are severe anaemic. For clarification the percentage of anaemic WRA with different levels of education of husband was observed.

Table 6.20 Percentage of anaemic WRA with different levels of husband's education.

	Anaemia Status			
	No Anaemia	Mild	Moderate	Severe
Primary	0	50%	50%	0
secondary	31.57%	32.89%	19.73%	15.78%
higher secondary	40%	38.57%	15.71%	5.77%
graduate	22.22%	55.56%	19.44%	0
post graduate	38.89%	22.22%	27.78%	11.11%

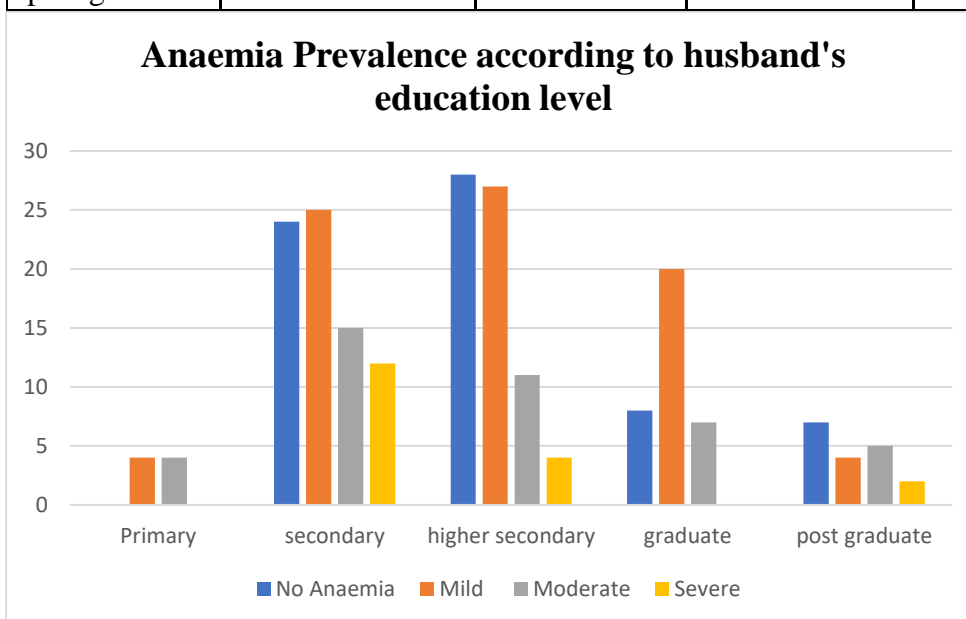


Fig. 6.7 Plot of anaemic WRA with different levels of husband's education.

From the above figure combined conclusion about the anaemia prevalence according to husband's education level can be made. The WRA whose husband's education was primary and secondary have generally been found to be anaemic since the percentage of mild, moderate and severe anaemia was high in these groups. Whereas in higher secondary, graduate and post graduate groups anaemia prevalence decreases. Therefore, we can say that as education level increases the anaemia severity decreases.

Drinking alcohol puts your physical and mental health at serious danger. Which will affect the health. Here the 'Alcohol consumption' variable was found to have a significant impact on anaemia status in the gradient boost algorithm. The 'Alcohol consumption' was a categorical variable which takes values 0 and 1. Where 0 is for no alcohol consumption and 1 is for alcohol consumption. The following table shows the distribution of anaemia according to alcohol consumption.

Table 6.21 Alcohol consumption status of WRA with Anaemia severity.

		Anaemia Status			
		No Anaemia	Mild	Moderate	Severe
Alcohol Consumption	0	34%	36%	20%	10%
	1	23%	51%	21%	4%

From the above data table, it appears that WRA who reported alcohol consumption 1 had higher percentages of mild anaemia compared to those who reported no alcohol consumption 0. Conversely, those who did not consume alcohol showed higher percentages of moderate and severe anaemia.

CHAPTER 7
COMPARING THE PERFORMANCE OF MACHINE LEARNING
ALGORITHMS ON MARRIED PREGNANT WRA

7.1 Introduction

In the previous sections the machine learning algorithms were developed on the Non- pregnant married WRA. According to WHO the Anaemia cut off is different for pregnant and non-pregnant women. Here there are total 268 pregnant WRA with 53 predictors most are same as that of married non-pregnant WRA but one more pregnancy related variable like gestational month was added in this data. after the data pre-processing there were total 258 pregnant women. The further analysis was done on these 258 pregnant WRA. The prevalence of anaemia was determined on these 258 pregnant women the results are as discussed in next section.

7.2 Prevalence of anaemia among pregnant WRA.

Table 7.1 Anaemia distribution of pregnant WRA.

Row Labels	Count of Anaemia
Mild	64
Moderate	60
No Anaemia	110
Severe	24
Grand Total	258

According to the results, it was evident that 110 out of the 258 women do not have anaemia. However, Anaemia still affects an enormous percentage of women, with 64 having mild anaemia, 60 having moderate anaemia, and 24 having severe anaemia.

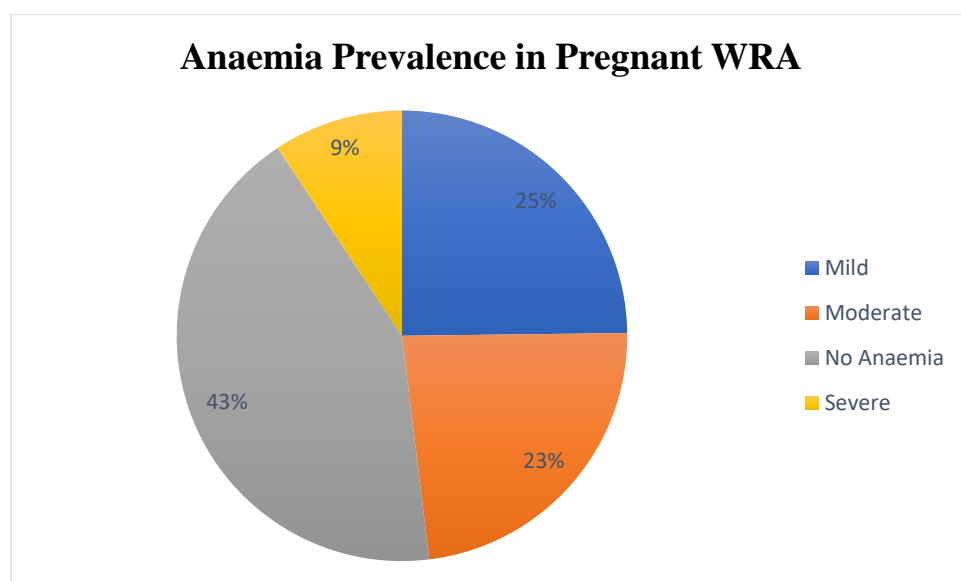


Fig. 7.1 Anaemia Prevalence in Pregnant WRA.

The results show that 57.36 % pregnant women found to be anaemic. This shows that the various degrees of anaemia's severity demand attention and possible action as a relevant health problem in this population.

To predict the status of anaemia and to identify the determinants associated with anaemia, various ML algorithms were developed. The ML algorithms trained on train data and tested on test data. To test the various machine learning models here the data was partitioned into train-test by 80-20 rule. Therefore, the train data consists of 206 samples and test data consists of 52 sample points. The various machine learning models were developed on this train data set and tested on test data set in the further section.

7.3 Decision Tree:

First the simple decision tree algorithm with 10-fold cross validation was developed on whole data. The R output given below:

```
> trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
```

```
> mc1=train(Anaemia~., data=data, method='rpart',trControl=trctrl)
```

```
> mc1
```

```
CART
```

```
258 samples
```

```
53 predictor
```

```
4 classes: '0', '1', '2', '3'
```

```
No pre-processing
```

```
Summary of sample sizes: 233, 231, 233, 232, 232, 231, ...
```

```
Results across tuning parameters:
```

cp	Accuracy	Kappa
0.04954955	0.5430883	0.27201925
0.06756757	0.5153960	0.20809376
0.13513514	0.4573504	0.07577496

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was $cp = 0.04954955$.

Interpretation:

Overall, the output suggests that the CART model trained on the given dataset did not perform very well, as indicated by the low accuracy and kappa values. The model refinement may be needed to improve the performance of the model. From the

above model cp value 0.04954955 is significant as compared to other two values. Therefore, the decision tree with cp=0.04954955 was developed and analysed the accuracy of the model.

Decision tree with cp=0.04954955:

```
> # decision tree with cross validation:
> library(rpart)
> m1=rpart(Anaemia~.,data= train, method='class',cp = 0.04954955)
> m1
n= 206
  > summary(m1)
Call:
rpart(formula = Anaemia ~ ., data = train, method = 'class',
      cp = 0.04954955)
n= 206
```

	CP	nsplit	rel error	xerror	xstd
1	0.15000000	0	1.0000000	1.0000000	0.05898275
2	0.06666667	1	0.8500000	0.8500000	0.05980010
3	0.05833333	3	0.7166667	0.8666667	0.05980010
4	0.04954955	4	0.6583333	0.8083333	0.05970136

Variable importance

HIV.status	18
Age.at.the.marriage	14
Age	7
Age.of.last.children.month...	6
days.of.blood.flow..	6
Education.years.	6
No.of.births.in.last.5.years	

5
Pain.on.menstrual.period..
5
Premature.Delivery..
5
Number.of.years.lives.in..residential.area.
5
Household.Wealth.status..
4
Total.number.of.children.ever.born
4
Occupation
2
BMI
2
Community.women.education
2
Age.at.first.birth.of.child
1
Number.of.family.members
1
Type.of.Addiction
1
Alcohol.Consumption..
1
husband.s.education..in.years.
1
Gestational.month

Variable Importance Table for decision tree:

Table 7.2 Most influential factors related to anaemia by CART.

Variable	Variable importance
HIV Status	18
age at marriage	14

age	7.8319
age of last children (in month)	6
Days of blood flow	6
Education	6
No. of births in last five years	5
Pain on menstrual period	5
Premature Delivery	5
No. of years lives in residential area	5
Houshold wealth status	4
Total Number of children ever born	4
Occupation	2
BMI	2

```
> P=predict(m1, newdata = test, type='class')
```

```
> Table=table(test$Anaemia,P)
```

```
> Table
```

```
P
```

```
0 1 2 3
```

```
0 15 4 5 0
```

```
1 10 5 2 0
```

```
2 2 2 3 2
```

```
3 0 0 0 2
```

```
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
```

```
[1] 0.4807692
```

From the Confusion matrix the decision tree model shows only 48% accuracy which is not significant.

```
> rpart.plot(m,extra=104)
```

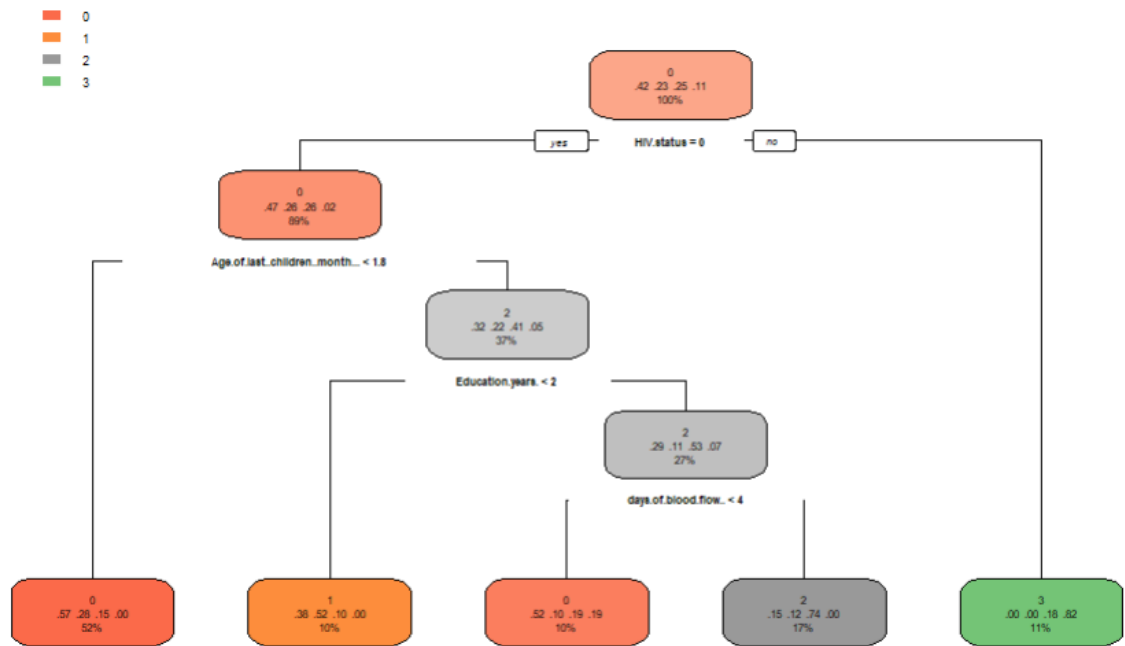



Fig. 7.2 Decision Tree plot 1 for Pregnant WRA

From the variable importance table as well as the decision tree plot here it was discovered that, according to decision tree algorithm factors like HIV status, age at marriage, age of last children in month, women’s education, no. of days of blood flow etc were significantly associated with status of anaemia. But Decision tree model not much predictive since its accuracy was 48% only.

Due to insufficient accuracy we can’t move forward with this decision tree model. So for the further prediction here some other models were developed.

7.4 Support vector machine with various kernels.

As seen in the previous two chapters SVM gives significant accuracy than decision tree. But while building SVM there is need to choose best or appropriate tuning parameters so that we can achieve maximum accuracy. In this section the 10-fold cross validation technique were employed first to select optimum CP value then by using that optimum CP value model was fitted. The results are discussed in following section:

7.4.1 Support Vector machine with Linear kernel

```
> trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
> mc2=train(Anaemia~., data=data, method='svmLinear',trControl=trctrl)
> mc2
```

Support Vector Machines with Linear Kernel

258 samples

53 predictor

4 classes: '0', '1', '2', '3'

No pre-processing

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 231, 233, 232, 233, 232, 233, ...

Resampling results:

Accuracy Kappa

0.8596125 0.7935032

Tuning parameter 'C' was held constant at a value of 1

Interpretation:

During the 10-fold cross-validation, the SVM model with a linear kernel had an accuracy of approximately 85.96% on average hence it displayed significant agreement with the actual class labels. It's crucial to remember that a number of variables, including the kernel selection, hyperparameter tuning, feature selection, and the characteristics of the data itself, may have an impact on how well the SVM model performs. We can might wish to experiment with various kernels, such as the radial basis function, or try fine-tuning the hyperparameters, such as C, using methods like grid search or random search. Consider feature engineering and pre-processing procedures as well to improve the model's capacity for prediction. Here the SVM with linear kernel was used to developed the model for prediction purpose in the next section.

```
> classifier_li = svm(formula = Anaemia ~ ., data = train, type = 'C-classification', kernel = 'linear')
```

```
> classifier_li
```

```
> summary(classifier_li)
```

Call:

```
svm(formula = Anaemia ~ ., data = train, type = 'C-classification',  
     kernel = 'linear')
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: linear

cost: 1

No. of Support Vectors: 124

(38 40 6 40)

Number of Classes: 4

Levels:

```

0 1 2 3
> P=predict(classifier_li, newdata = test)
> Table=table(test$Anaemia,P)
> Table
  P
  0 1 2 3
0 18 4 2 0
1  3 14 0 0
2  0 0 9 0
3  0 0 0 2
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.8269231

```

Here the SVM with linear kernel shows 82.69 % accuracy.

The support vector machine with linear kernel has 124 support vectors and cost is constant. According to accuracy model seems to be significant but we can't move forward with only accuracy. It is crucial to consider additional measures, particularly for each class, like precision, recall, and F1 score in order to assess the model's performance in greater detail.

Confusion matrix

```

> confusion_matrix <- matrix(c(18, 4, 2, 0,
+                               3, 14, 0, 0,
+                               0, 0, 9, 0,
+                               0, 0, 0, 2),
+                               nrow = 4, byrow = TRUE)
>
> # Function to calculate precision, recall, and F1-score for each class
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FN <- rowSums(cm) - TP
+   FP <- colSums(cm) - TP
+
+   precision <- TP / (TP + FP)
+   recall <- TP / (TP + FN)
+   f1_score <- 2 * precision * recall / (precision + recall)

```

```

+
+   return(data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall = recall,
+ F1_Score = f1_score))
+ }
>
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
Class Precision  Recall F1_Score
1  0 0.8571429 0.7500000  0.8
2  1 0.7777778 0.8235294  0.8
3  2 0.8181818 1.0000000  0.9
4  3 1.0000000 1.0000000  1.0

```

Classes 2 and 3 show remarkable precision and recall, with class 3's F1 score being 1.0. In comparison to classes 2 and 3, classes 0 and 1 have slightly less precision and recall. The F1 scores for all classes are comparatively high, demonstrating a classifier that performs fairly evenly across all classes. In conclusion, the model performs perfectly for the majority of courses, with classes 2 and 3 notably outstanding. To fully evaluate the model's performance and make wise decisions.

7.4.2 Support vector machine with Radial kernel:

For non-linear classification tasks, the Radial Basis Function (RBF) kernel commonly referred it is also known as Gaussian kernel, which is a well-liked SVM kernel. Finding a nonlinear decision boundary in the original feature space is made possible by the RBF kernel's transformation of the data into a higher-dimensional space. First the 10 fold cross validation of SVM with radial kernel was developed the results are as follows:

```

trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
> mc3=train(Anaemia~., data=data, method='svmRadial',trControl=trctrl)
> mc3

```

Support Vector Machines with Radial Basis Function Kernel

258 samples

53 predictor

4 classes: '0', '1', '2', '3'

No pre-processing

Summary of sample sizes: 233, 231, 233, 232, 232, 231, ...

Results across tuning parameters:

C	Accuracy	Kappa
0.25	0.5040000	0.1753733
0.50	0.5469345	0.2611997
1.00	0.7438519	0.6138148

Tuning parameter 'sigma' was held constant at a value of 0.01147407

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were sigma = 0.01147407 and C = 1.

The single SVM model with radial kernel was developed for sigma = 0.01147407 and C = 1 and it gives following output:

```
> classifier_rad = svm(formula = Anaemia ~ ., data = train, type = 'C-classification', kernel = 'radial', sigma = 0.01147407, C = 1.)
```

```
> classifier_rad
```

Call:

```
svm(formula = Anaemia ~ ., data = train, type = 'C-classification', kernel = 'radial', sigma = 0.01147407, C = 1)
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost:

Number of Support Vectors: 178

```
> summary(classifier_rad)
```

Call:

```
svm(formula = Anaemia ~ ., data = train, type = 'C-classification', kernel = 'radial', sigma = 0.01147407, C = 1)
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: radial

cost: 1

No. of Support Vectors: 178

(70 49 13 46)

Number of Classes: 4

Levels:

```

0 1 2 3
> Pr=predict(classifier_rad, newdata = test)
> Table=table(test$Anaemia,Pr)
> Table
  Pr
  0 1 2 3
0 22 0 2 0
1  5 12 0 0
2  1 0 7 1
3  0 0 0 2
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.8269231
> # Confusion matrix
> confusion_matrix <- matrix(c(22, 0, 2, 0,
+                               5, 12, 0, 0,
+                               1, 0, 7, 1,
+                               0, 0, 0, 2),
+                               nrow = 4, byrow = TRUE)
> # Function to calculate precision, recall, and F1 score for each class
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FN <- rowSums(cm) - TP
+   FP <- colSums(cm) - TP
+   precision <- TP / (TP + FP)
+   recall <- TP / (TP + FN)
+   f1_score <- 2 * precision * recall / (precision + recall)
+   metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall = recall, F1_Score = f1_score)
+   return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
  Class Precision Recall F1_Score

```

```

1 0 0.7857143 0.9166667 0.8461538
2 1 1.0000000 0.7058824 0.8275862
3 2 0.7777778 0.7777778 0.7777778
4 3 0.6666667 1.0000000 0.8000000

```

Overall Model Performance: The model performs well in some aspects and shows room for improvement in others. It's particularly strong in Class 1 with perfect recall, while Class 0 exhibits good precision. Class 2 showcases balanced precision and recall, resulting in consistent performance. However, in Class 3, the model's recall is lower, impacting the overall F1-score. To enhance the model's performance, further tuning and consideration of class-specific behaviours may be beneficial.

7.4.3 Support Vector Machine with Polynomial kernel:

```
> mc=train(Anaemia~., data=data, method='svmpoly',trControl=trctrl)
```

Error: Model svmpoly is not in caret's built-in library

```
> mc=train(Anaemia~., data=data, method='svmPoly',trControl=trctrl)
```

```
> mc
```

Support Vector Machines with Polynomial Kernel

258 samples

53 predictor

4 classes: '0', '1', '2', '3'

Summary of sample sizes: 233, 233, 231, 232, 232, 233, ...

Tuning parameters:

	degree	scale	C	Accuracy	Kappa
1	0.001	0.25	0.4267123	0.0000000	
1	0.001	0.50	0.4267123	0.0000000	
1	0.001	1.00	0.4267123	0.0000000	
1	0.010	0.25	0.5042963	0.1827286	
1	0.010	0.50	0.5229573	0.2239830	
1	0.010	1.00	0.5818917	0.3535831	
1	0.100	0.25	0.6242564	0.4417550	
1	0.100	0.50	0.6433333	0.4744612	
1	0.100	1.00	0.6739829	0.5174543	
2	0.001	0.25	0.4267123	0.0000000	
2	0.001	0.50	0.4267123	0.0000000	
2	0.001	1.00	0.5042963	0.1796912	

2	0.010	0.25	0.5384957	0.2482017
2	0.010	0.50	0.6553333	0.4704466
2	0.010	1.00	0.7016752	0.5552535
2	0.100	0.25	0.9461083	0.9219417
2	0.100	0.50	0.9461083	0.9219417
2	0.100	1.00	0.9461083	0.9219417
3	0.001	0.25	0.4267123	0.0000000
3	0.001	0.50	0.4930427	0.1523423
3	0.001	1.00	0.5118462	0.1968592
3	0.010	0.25	0.6624330	0.4757386
3	0.010	0.50	0.8017322	0.7038260
3	0.010	1.00	0.8831852	0.8289365
3	0.100	0.25	0.9461083	0.9219417
3	0.100	0.50	0.9461083	0.9219417
3	0.100	1.00	0.9461083	0.9219417

Accuracy measure was used here to select the optimal model by using the largest value.

The final values used for the model were degree = 2, scale = 0.1 and C = 0.25.

```
> classifier_poly = svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel = 'polynomial', degree=2, scale=0.1, C=0.25)
```

```
> classifier_poly
```

Call:

```
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel = 'polynomial',
```

```
  degree = 2, C = 0.25, scale = 0.1)
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: polynomial

cost: 1

degree: 2

coef.0: 0

Number of Support Vectors: 173

```
> summary(classifier_poly)
```

Call:


```
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel = 'poly  
nomial',
```

```
  degree = 2, C = 0.25, scale = 0.1)
```

Parameters:

SVM-Type: C-classification

SVM-Kernel: polynomial

cost: 1

degree: 2

coef.0: 0

Support Vectors: 173

(69 49 10 45)

Number of Classes: 4

Levels:

0 1 2 3

```
> p=predict(classifier_poly, newdata = test, type='class')
```

```
> Table=table(test$Anaemia,p)
```

```
> Table
```

p

0 1 2 3

0 24 0 0 0

1 9 7 1 0

2 3 0 6 0

3 0 0 0 2

```
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
```

```
[1] 0.75
```

```
> # Confusion matrix
```

```
> confusion_matrix <- matrix(c(24, 0, 0, 0,
```

```
+           9, 7, 1, 0,
```

```
+           3, 0, 6, 2,
```

```
+           0, 0, 0, 2),
```

```
+           nrow = 4, byrow = TRUE)
```

```
> # Function to calculate precision, recall, and F1 score for each class
```

```
> calculate_metrics <- function(cm) {
```

```
+ TP <- diag(cm)
```

```

+ FN <- rowSums(cm) - TP
+ FP <- colSums(cm) - TP
+
+ precision <- TP / (TP + FP)
+ recall <- TP / (TP + FN)
+ f1_score <- 2 * precision * recall / (precision + recall)
+
+ metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall = recall, F1_Score = f1_score)
+ return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
> print(metrics)
  Class Precision Recall F1_Score
1    0 0.6666667 1.0000000 0.8000000
2    1 1.0000000 0.4117647 0.5833333
3    2 0.8571429 0.5454545 0.6666667
4    3 0.5000000 1.0000000 0.6666667

```

Overall Interpretation:

The SVM model with polynomial kernel gives accuracy of 75% on the test data. The class-wise performance metrics reveal that the model performs well for certain classes (high precision or recall) but not as well for others. Class 1 has the highest recall (ability to detect positive cases), while Class 1 has the highest precision (ability to avoid false positives). Class 2 and Class 3 exhibit a trade-off between precision and recall. The F1 Score provides a balanced measure considering both precision and recall. It's important to consider these metrics together to understand the model's performance across different classes.

7.4.5 Support Vector Machine with Sigmoid kernel:

For better performance the support vector machine with sigmoid kernel was developed, the results are as follows:

```

> classifier_sig = svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel = 'sigmoid')

```

```

> classifier_sig
Call:
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel = 'sigmoid')
Parameters:
  SVM-Type: C-classification
  SVM-Kernel: sigmoid
    cost: 1
    coef.0: 0
Number of Support Vectors: 179
> summary(classifier_sig)
Call:
svm(formula = Anaemia ~ ., data = train_data, type = 'C-classification', kernel = 'sigmoid')
Parameters:
  SVM-Type: C-classification
  SVM-Kernel: sigmoid
    cost: 1
    coef.0: 0
Support Vectors: 179

( 69 48 15 47 )
Number of Classes: 4
Levels:
0 1 2 3
> p=predict(classifier_sig, newdata = test_data, type='class')
> Table=table(test$Anaemia,p)
> Table
  p
  0 1 2 3
0 21 3 0 0
1 11 4 2 0
2 3 1 3 2
3 0 0 0 2

```

```
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
```

```
[1] 0.5769231
```

Interpretation:

The SVM model with a sigmoid kernel achieved an accuracy of around 57.69% on the test data. The confusion matrix indicates how well the model's predictions align with the actual classes. Class 0 seems to be predicted relatively well, but there are challenges in predicting the other classes. Class 1, for example, has a higher number of predicted instances than actual instances, which might indicate some misclassification. The relatively lower accuracy suggests that the model might not be performing as well as desired, and further analysis or model tuning could be considered to improve its performance. As the accuracy was found to be low so there were no need of examining other model evaluation measures like precision, recall, and F1 score. Since the SVM with linear kernel gives highest accuracy among all fitted algorithms clearly shows data was linearly separable. 85% accuracy was acceptable but in previous chapters rather than DT and SVM other models were developed. So, in the next section K-nearest algorithm was run on the same data for accuracy comparison.

7.5 K-Nearest Neighbour algorithm:

The KNN algorithm is a simple and intuitive ML algorithm used for classification and regression tasks. It is a non-parametric ML algorithm, meaning it doesn't make any assumptions about the underlying data distribution and instead directly relies on the training data to make predictions. Following are the outputs of K Nearest Neighbour.

```
> trctrl=trainControl(method = 'cv', number = 10, savePredictions=TRUE)
```

```
> mc1=train(Anaemia~., data=data, method='knn',trControl=trctrl)
```

```
> mc1
```

```
k-Nearest Neighbors
```

```
258 samples
```

```
53 predictor
```

```
4 classes: '0', '1', '2', '3'
```

```
Summary of sample sizes: 232, 232, 233, 232, 231, 233, ...
```

```
Resampling-results across tuning parameters:
```

```
k Accuracy Kappa
```

```
5 0.5886610 0.4052938
```

```
7 0.5115271 0.2904955
```

```
9 0.4381083 0.1715444
```

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was $k = 5$.

```
> #Fitting KNN Model to training dataset
> classifier_knn <- knn(train = train,test = test,cl = train$Anaemia,k = 5)
> classifier_knn
[1] 0 0 0 0 1 0 2 2 2 2 2 2 0 1 0 0 3 0 0 0 0 0 0 2 1 1 3 1 2 2 1 0 0 3 0 0 2 0 0 2 2 1 2 0
2 0 0 1 2
[50] 2 2 0
Levels: 0 1 2 3
> summary(classifier_knn)
 0  1  2  3
24  8 17  3
> cm <- table(test$Anaemia, classifier_knn)
> cm
  classifier_knn
    0  1  2  3
0 15  2  7  0
1  7  6  3  1
2  2  0  6  1
3  0  0  1  1
> # Confusion matrix
> confusion_matrix <- matrix(c(15, 2, 7, 0,
+           7, 6, 3, 1,
+           2, 0, 6, 1,
+           0, 0, 1, 1),
+           nrow = 4, byrow = TRUE)
> # Function to calculate various assessment measures for each class
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FN <- rowSums(cm) - TP
+   FP <- colSums(cm) - TP
+
+   precision <- TP / (TP + FP)
+   recall <- TP / (TP + FN)
```

```

+ f1_score <- 2 * precision * recall / (precision + recall)
+
+ metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall =
recall, F1_Score = f1_score)
+ return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
Class Precision Recall F1_Score
1 0 0.6250000 0.6250000 0.6250000
2 1 0.7500000 0.3529412 0.4800000
3 2 0.3529412 0.6666667 0.4615385
4 3 0.3333333 0.5000000 0.4000000

```

Interpretation:

The KNN model’s performance was evaluated using cross-validation, and the optimal number of neighbours (k = 5) was selected based on the highest accuracy. When applied to the test dataset, the model’s performance varied across different classes. Class 1 achieved relatively good precision and recall, while Class 0 had higher recall but lower precision. Class 2 exhibited a relatively high precision but lower recall, and Class 3’s performance was moderate. The overall accuracy of the model on the test data was 53.84%. The F1-Score provides a balanced measure of precision and recall. The results suggest that the model’s performance is class-dependent and may require further analysis or improvements to achieve better results.

7.6 Ensembling techniques on pregnant women data:

In the field of machine learning, ensembling techniques have become effective methods for improving prediction performance while dealing with the weaknesses of individual models. Ensembling approaches take advantage of the variety of each model’s strengths and make up for its deficiencies by merging its forecasts with those of other models. For example, bagging uses bootstrap sampling to create varied datasets for each model’s training, lowering overfitting and boosting stability. Boosting is the technique which iteratively improves weak learners by concentrating on samples that earlier models had trouble with, resulting in a more precise and reliable ensemble. Voting methods aggregate predictions through consensus, producing trustworthy

results, while stacking adopts a creative approach by using a meta-model to combine the distinctive viewpoints of base models. These methods can combine a number of weak models into a powerful ensemble that consistently delivers improved accuracy and better generalisation, allowing for more precise and trustworthy data insights.

Here some ensembling techniques were developed to improve model's performance.

7.6.1 Bagged decision Tree:

```
> # Create bagged classification tree model
> bagged.tree <- bagging(Anaemia ~ ., data = train, nbagg = 100, coob = TRUE, control
= rpart.control(maxdepth = 2, minsplit = 1))
> bagged.tree
Bagging classification trees with 100 bootstrap replications
Call: bagging.data.frame(formula = Anaemia ~ ., data = train, nbagg = 100,
  coob = TRUE, control = rpart.control(maxdepth = 2, minsplit = 1))
Out-of-bag estimate of misclassification error: 0.4903
Error in exists(cacheKey, where = .rs.CachedDataEnv, inherits = FALSE) :
  invalid first argument
Error in if (maxRows != -1 && nrow(data) > maxRows) data <- head(data, :
  missing value where TRUE/FALSE needed
> bagged.tree$OOB
[1] TRUE
> p=predict(bagged.tree, newdata = test, type='class')
> p
[1] 0 0 0 0 0 0 0 0 3 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0
[50] 0 0 0
Levels: 0 1 2 3
> Tl=table(test$Anaemia,p)
> Tl
  p
  0 1 2 3
0 24 0 0 0
1 17 0 0 0
2  7 0 1 1
3  0 0 0 2
```

```

> Accuracy=sum(diag(TI))/sum(TI);Accuracy
[1] 0.5192308
> # Confusion matrix
> confusion_m <- matrix(c(24, 0, 0, 0,
+                        17, 0, 0, 0,
+                        7, 0, 0, 1,
+                        0, 0, 1, 2),
+                        nrow = 4, byrow = TRUE)
> # Function to calculate precision, recall, and F1 score for each class
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FN <- rowSums(cm) - TP
+   FP <- colSums(cm) - TP
+
+   precision <- TP / (TP + FP)
+   recall <- TP / (TP + FN)
+   f1_score <- 2 * precision * recall / (precision + recall)
+
+   metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall =
recall, F1_Score = f1_score)
+   return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_m)
> print(metrics)
  Class Precision Recall F1_Score
1    0 0.5000000 1.0000000 0.6666667
2    1    NaN 0.0000000    NaN
3    2 0.0000000 0.0000000    NaN
4    3 0.6666667 0.6666667 0.6666667

```

Interpretation:

After performing the bagging technique on a classification issue where the attribute ‘Anaemia’ was predicted. A maximum depth of 2 and a minimum split size of 1 classification tree were used in the created bagged model, which produced an out-of-

bag (OOB) estimate of misclassification error of about 49.03%. The OOB error is a measure of how well the model performed on training data that had not yet been seen or simply a test data.

When the bagged model applied on the test dataset, the overall accuracy of the predictions was approximately 51.92%. which was the indication that the bagged model's predictions were not satisfactory. We can examine how the model's predictions were spread among various classes by analysing the confusion matrix. The matrix reveals that class 0 had the most accurate predictions (24), while other classes had far less. This implies that the model favours class 0 in its predictions.

Despite the bagged model's ability to predict some outcomes, there was opportunity for improvement, especially in classes other than 0. It was observed that the model may need additional improvement due to relatively low accuracy and NaN (Not a Number) results for several metrics. This could be done by adjusting hyperparameters, utilising more sophisticated models, or addressing class imbalance concerns.

Therefore, here the another ensembling technique Random Forest was developed in the next section.

7.6.2 Random Forest:

```
> # Random Forest
```

```
> library(randomForest)
```

```
randomForest 4.7-1.1
```

```
Type rfNews() to see new features/changes/bug fixes.
```

```
Attaching package: 'randomForest'
```

```
The following object is masked from 'package:ggplot2':
```

```
margin
```

```
The following object is masked from 'package:dplyr':
```

```
combine
```

```
Warning message:
```

```
package 'randomForest' was built under R version 4.2.3
```

```
> rf_model =randomForest(Anaemia ~ ., data = train, ntree = 100)
```

```
> rf_model
```

```
Call:
```

```
randomForest(formula = Anaemia ~ ., data = train, ntree = 100)
```

```
Type of random forest: classification
```

```

Number of trees: 100
No. of variables tried at each split: 7
OOB estimate of error rate: 12.14%
Confusion matrix:
  0 1 2 3 class.error
0 82 3 1 0 0.04651163
1 10 33 4 0 0.29787234
2 2 3 44 2 0.13725490
3 0 0 0 22 0.00000000
> p=predict(rf_model, newdata = test, type='class')
> Table=table(test$Anaemia,p)
> Table
  p
  0 1 2 3
0 22 0 2 0
1 0 17 0 0
2 0 0 9 0
3 0 0 0 2
> Accuracy=sum(diag(Table))/sum(Table);Accuracy
[1] 0.9615385
> #Confusion matrix
> confusion_matrix <- matrix(c(22, 0, 2, 0,
+                               0, 17, 0, 0,
+                               0, 0, 9, 0,
+                               0, 0, 0, 2),
+                               nrow = 4, byrow = TRUE)
> # Function to calculate precision, recall, and F1 score for each class
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FN <- rowSums(cm) - TP
+   FP <- colSums(cm) - TP
+
+   precision <- TP / (TP + FP)
+   recall <- TP / (TP + FN)

```

```

+ f1_score <- 2 * precision * recall / (precision + recall)
+
+ metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall =
recall, F1_Score = f1_score)
+ return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
  Class Precision  Recall F1_Score
1    0 1.0000000 0.9166667 0.9565217
2    1 1.0000000 1.0000000 1.0000000
3    2 0.8181818 1.0000000 0.9000000
4    3 1.0000000 1.0000000 1.0000000

```

Interpretation:

For the ‘Anaemia’ prediction, a Random Forest classification algorithm was created using training data. In the ensemble, the model used 100 decision trees. The estimated error rate during training, known as the out-of-bag (OOB) estimate, is approximately 12%. This shows that the model worked well with the training set of data and is predicted to generalise rather well with new data. The overall prediction accuracy was around 96.15% when the trained Random Forest model was applied to the test dataset. This high accuracy indicates that the Random Forest model predicts the test data quite accurately.

The distribution of the model’s predictions across several classes is further depicted by the confusion matrix. The matrix shows that the model predicted outcomes correctly for the majority of classes. Class 1 , class 2 and Class 3 had good precision and recall levels, also Class 1 ,2 and 3 have 100% accuracy.

It was reported that the trained Random Forest model performed admirably on the test dataset, with high accuracy and evenly distributed precision and recall scores across most classes. This demonstrates the model’s robustness and ability to make precise predictions based on unobserved data.

As fitted model is robust so we can determine the factors associated with the status of anaemia. We can directly find that factors in random forest algorithm the results are as follows:

```
> varImp(rf_model)
```

	Overall
HIV.status	3.4256182
Are.you.feeling.weak.or.dizziness.	2.7982362
Age	5.6514002
Education.years.	3.3985772
Occupation	1.4785951
Income.of.the.family...Rs..Annual.	3.8713517
weight.kg.	6.9547694
Height..meter.	3.8528015
BMI	6.3294495
Eating.Habits	6.0789474
Food.type	0.6884493
Daily..Tea.intake	2.2125327
Acidity.Problem..	1.0877445
Age.at.the.marriage	8.6332547
husband.s.age	5.3230330
husband.s.age.at.marriage	5.1688552
Husband.s.Occupation	2.2132754
husband.s.education..in.years.	2.6226115
Alcohol.Consumption..	0.8454650
Any.Addiction..	0.9987863
Type.of.Addiction	1.0417566
Suffer.from.any.long.term.disease	2.0584246
Suffer.from.stress..	1.0762874
use.Iron.supplementation..	0.7143384
Suffers.from.Diabetes	1.6305352
Household.Wealth.status..	1.2157040
Number.of.family.members	5.2471379
Toilet.facility..	0.5192313
Drinking.water.source..	1.7941517
Cooking.fuel	2.4955084
Exposure.to.domestic.violence..	1.8160395
Avg.of.rest.in.day..per.Hr...	3.1579961

Regular.visit.to.doctor..	0.3249408
Daily.eat.fresh.fruits.Vegetable..Milk..	0.6517712
Menstrual.cycle.1..	0.9283361
Menstrual.cycle.2..	0.8671876
No.of.pads...per.day.	3.1725681
days.of.blood.flow..	5.3550213
Pain.on.menstrual.period..	1.9673783
Age.at..menstrual.cycle.begins	5.3790329
Gestational.month	2.0235056
Total.number.of.children.ever.born	1.7799203
Premature.Delivery..	2.5359397
Miscarage.History..	0.7832226
Age.at.first.birth.of.child	3.8463906
Age.of.last..children..month...	3.8646530
No..of.births.in.last.5.years	2.1102982
Use.of.Contraceptive	2.0881693
Method.of.Contraceptive	2.3069727
Region	0.5324025
Number.of.years.lives.in..residential.area.	4.6099380
Mass.media.exposure	0.9260372
Community.women.education	1.1977469

> varImpPlot(rf_model)

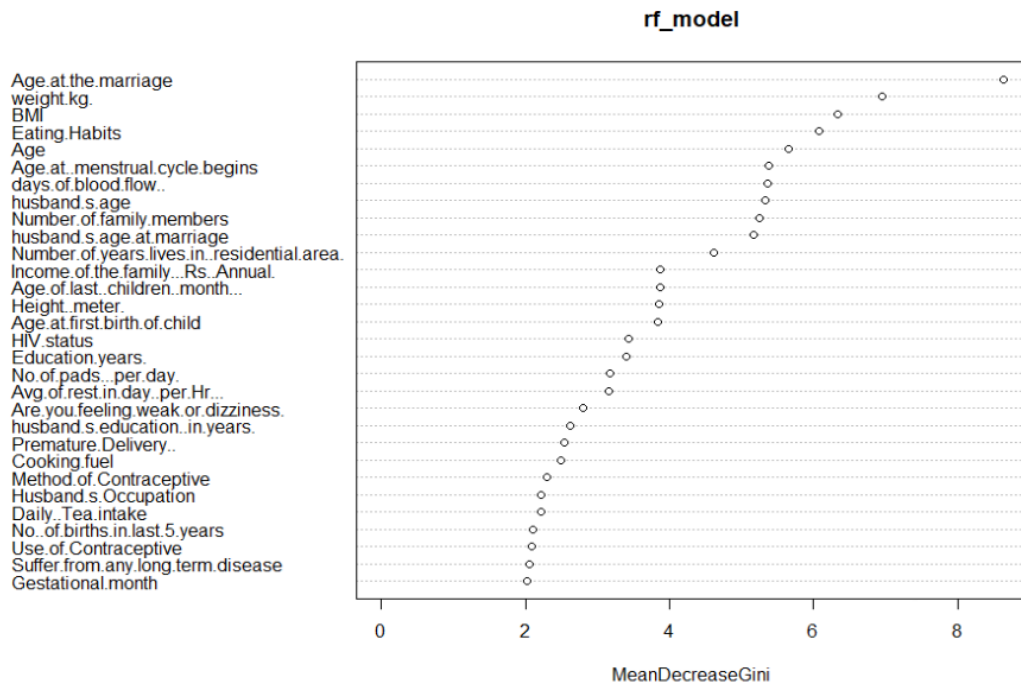


Fig. 7.3 VarImpPlot by RF for pregnant WRA

Variable Importance Table

Table 7.3 Table of variable importance by RF.

Variable	Variable importance
Age at the marriage	8.6332
weight	6.9547
BMI	6.3294
Eating habbits	6.0789
Alcohol consuption	4.5753
Age	5.6514
Age at menstrual cycle begins	5.379
Days of blood flow	5.355
Husbands age	5.3203
Number of family members	5.2471
Husband's age at marriage	5.1688
Number of years lives in residential area	4.6099
Income of the family	3.8713
Age of last children	3.8646
Height	3.8528
Age.at.first.birth.of.child	3.8463
HIV Status	3.4256
Education	3.3985
No.of.pads...per.day	3.1725
Avg.of.rest.in.day	3.1579

Interpretation:

The variable importance analysis for predicting anaemia demonstrates that a number of variables have a big impact on the result. The factor 'Age at the marriage' holds the highest importance score of 8.6332, suggesting it is a key determinant in the prediction of anaemia in pregnant WRA. This variable likely plays a pivotal role in understanding factors associated with anaemia, indicating that the age at which individuals get married significantly influences the status of anaemia. Following closely 'weight' with an importance score of 6.9547, emphasizing its substantial impact on the prediction of anaemia. Weight could be a crucial factor in determining status of anaemia in respective WRA. It indicates that weight significantly contributes to the model's understanding of the scenario under study. Other variables such as 'BMI,' 'Eating habits,' 'Alcohol consumption,' 'Age,' 'Age at menstrual cycle begins,' 'Days of blood flow,' 'Husband's age,' and 'number of family members,' 'husband's age at marriage' also hold considerable importance, with scores ranging between 5 to 6. These factors also contribute significantly while predicting anaemia in pregnant WRA. On the other hand, factors like 'Income of the family,' 'Age of last children,' 'Height,' 'HIV Status,' 'Education,' 'No. of pads per day,' and 'Average of rest in a day' have relatively lower importance scores ranging from 3 to 4. While these factors still contribute to the model's predictions, their influence might be comparatively less pronounced than the higher-scoring variables. These results highlight the complexity of the prediction problem, where lifestyle, health, and demographic factors all play a substantial role in predicting anaemia. It's significant to highlight that obtaining precise forecasts for this health-related result appears to depend on a combination of both individual features and more general contextual elements.

The Random Forest model had an impressive 96% accuracy; thus it was curious to discover another model to improve prediction abilities. A common ensemble learning strategy known as AdaBoost, or Adaptive Boosting, focuses on progressively enhancing model performance by providing greater weight to misclassified cases in each iteration. Therefore, in the next section ADA Boost algorithm was developed on the same dataset, results are as follows:

7.6.3 ADA Boosting:

AdaBoost seeks to build a reliable and accurate ensemble model by progressively training a number of weak learners and integrating their predictions. The ADA Boost algorithm with 100 iterations the results are as follows:

```

$call
boosting(formula = Anaemia ~ ., data = train, boos = TRUE, mfinal = 100)
attr(,"vardep.summary")
 0  1  2  3
86 47 51 22
attr(,"class")
[1] "boosting"
> p=predict(ada_model, newdata = test, type='class')
> p
$confusion
      Observed Class
Predicted Class 0  1  2  3
      0 18  0  0  0
      1  2 15  0  0
      2  4  2  9  0
      3  0  0  0  2
$error
[1] 0.1538462
> # Confusion matrix
> confusion_matrix <- matrix(c(18, 0, 0, 0,
+           2,15, 0, 0,
+           4, 2, 9, 0,
+           0, 0, 0, 2),
+           nrow = 4, byrow = TRUE)
> Accuracy=sum(diag(confusion_matrix))/sum(confusion_matrix)
> Accuracy
[1] 0.8461538
> # Function to calculate precision, recall, and F1 score for each class
> calculate_metrics <- function(cm) {
+   TP <- diag(cm)
+   FP <- rowSums(cm) - TP
+   FN <- colSums(cm) - TP
+
+   precision <- TP / (TP + FP)

```



```

+ recall <- TP / (TP + FN)
+ f1_score <- 2 * precision * recall / (precision + recall)
+
+ metrics <- data.frame(Class = 0:(nrow(cm) - 1), Precision = precision, Recall =
recall, F1_Score = f1_score)
+ return(metrics)
+ }
> # Calculate metrics for each class
> metrics <- calculate_metrics(confusion_matrix)
> print(metrics)
  Class Precision  Recall F1_Score
1    0 1.0000000 0.7500000 0.8571429
2    1 0.8823529 0.8823529 0.8823529
3    2 0.6000000 1.0000000 0.7500000
4    3 1.0000000 1.0000000 1.0000000

```

Interpretation:

AdaBoost is a boosting technique that emphasises iterative model enhancement by increasing the weight of cases that were incorrectly classified in each round. 100 iterations were included in the above fitted model configuration ($m_{final} = 100$). From the confusion matrix demonstrates that most of the predictions made by the model were correct. Class 0 (no anaemia) had the most accurate predictions with 18, followed by class 1 (mild anaemia) with 15, class 2 (moderate anaemia) with 9, and class 3 (severe anaemia) with two. On the test data, the model's overall error rate was approximately 15.38%. In addition, the accuracy of the model was at 84.62% when calculated by the confusion matrix.

In conclusion, the AdaBoost model has an accuracy of about 84.62%, showing a good level of prediction potential for anaemia classification. For some classes, like class 3, such as class 3, precision, recall, and F1-score values are high, however these values differ for other classes. This suggests that while the model struggles to correctly predict all classes, it does well in some areas of the prediction task. The model's performance across all classes might be improved with additional tuning and refining.

As the accuracy of the ADA Boost is satisfactory so there was a interest of determining the factors associated with status of anaemia. The results are as follows:

```
> ada_model$importance
```

Acidity.Problem..	Age
0.69723053	3.43305937
Age.at..menstrual.cycle.begins	Age.at.first.birth.of.child
5.31089045	2.93842455
Age.at.the.marriage	Age.of.last..children..month...
6.57974601	2.93829061
Alcohol.Consumption..	Any.Addiction..
0.49263450	0.30740170
Are.you.feeling.weak.or.dizziness.	Avg.of.rest.in.day..per.Hr...
2.41421194	2.44294297
BMI	Community.women.education
5.15031723	0.58070181
Cooking.fuel	Daily..Tea.intake
2.03620843	2.07054515
Daily.eat.fresh.fruits.Vegetable..Milk..	days.of.blood.flow..
0.30592787	4.93438380
Drinking.water.source..	Eating.Habits
1.02539156	4.94904286
Education.years.	Exposure.to.domestic.violence..
2.75433975	0.34514280
Food.type	Gestational.month
0.63320670	0.97507773
Height..meter.	HIV.status
2.14545996	3.96160279
Household.Wealth.status..	husband.s.age
0.81096249	3.86317705
husband.s.age.at.marriage	husband.s.education..in.years.
5.06326277	2.30651181
Husband.s.Occupation	Income.of.the.family...Rs..Annual.
1.04757951	2.82630503
Mass.media.exposure	Menstrual.cycle.1..
1.03952767	0.06577229
Menstrual.cycle.2..	Method.of.Contraceptive
0.05702829	0.98391633

Miscarage.History..	No..of.births.in.last.5.years
0.04234165	0.57885241
No.of.pads...per.day.	Number.of.family.members
1.23364116	4.26614002
Number.of.years.lives.in..residential.area.	Occupation
2.27854725	1.35701020
Pain.on.menstrual.period..	Premature.Delivery..
1.47531922	0.76924881
Region	Regular.visit.to.doctor..
0.00000000	0.00000000
Suffer.from.any.long.term.disease	Suffer.from.stress..
1.68190005	1.07977447
Suffers.from.Diabetes	Toilet.facility..
0.77827513	0.22935960
Total.number.of.children.ever.born	Type.of.Addiction
0.00000000	0.06382987
use.Iron.supplementation..	Use.of.Contraceptive
0.47729630	1.12439888
weight.kg.	
5.07784068	

Variable Importance table:

Table 7.4 Sorted important variables by Ada Boost.

Variable	Variable importance
Age at the marriage	6.5797
Age at menstrual cycle begins	5.3108
BMI	5.1503
Husband's age at marriage	5.0632
Eating habits	4.949
Days of blood flow	4.9343
Number of family members	4.2661
HIV status	3.9616
Husband's age	3.8631
age	3.433
Age at first birth of child	2.9384
Age of Last children	2.9382

Family income	2.8263
Education	2.7543
Average rest in day	2.4429
Feeling weak	2.4142
Husbands eduction	2.3065
Number of years lives in residential area	2.2785
Height	2.1454
Daily Tea intake	2.0705
Cooking fuel	2.036

The AdaBoost model's variable importance output offers insights into the relative significance of features for anaemia prediction. Notably, variables like 'Age at the marriage,' 'Age at the menstrual cycle begins,' 'Age at the first birth of child,' 'Age of last children (months),' 'Age,' 'Husband's age at marriage,' and 'Weight (kg)' have high relevance scores, indicating their significant influence on the prediction outcomes. On the other hand, factors like 'Region' and 'Regular visit to doctor' seem to be of little significance. These results show the predictive value of demographic, Household, and individual level factors in predicting anaemia while also underscoring the possibility that some contextual factors may have less of an effect on the prediction task.

7.7 Comparative Examination of developed ML algorithms:

Table 7.5 Table of accuracy for developed models.

Sr. No.	ML Algorithm	Accuracy(%)
1	Decision tree (after cross validation)	48.07%
2	SVM(linear)	82.69%
3	SVM (Polynomial)	75.00%
4	SVM (Sigmoid)	57.69%
5	SVM (Radial)	82.69%
6	K- nearest Neighbour (with k=7)	53.84%
7	Bagged Decision tree (with nbag=100)	51.92%
8	Random Forest Algorithm (with 100 trees)	96.15%
9	Ada Boost (mfinal=100)	84.61%

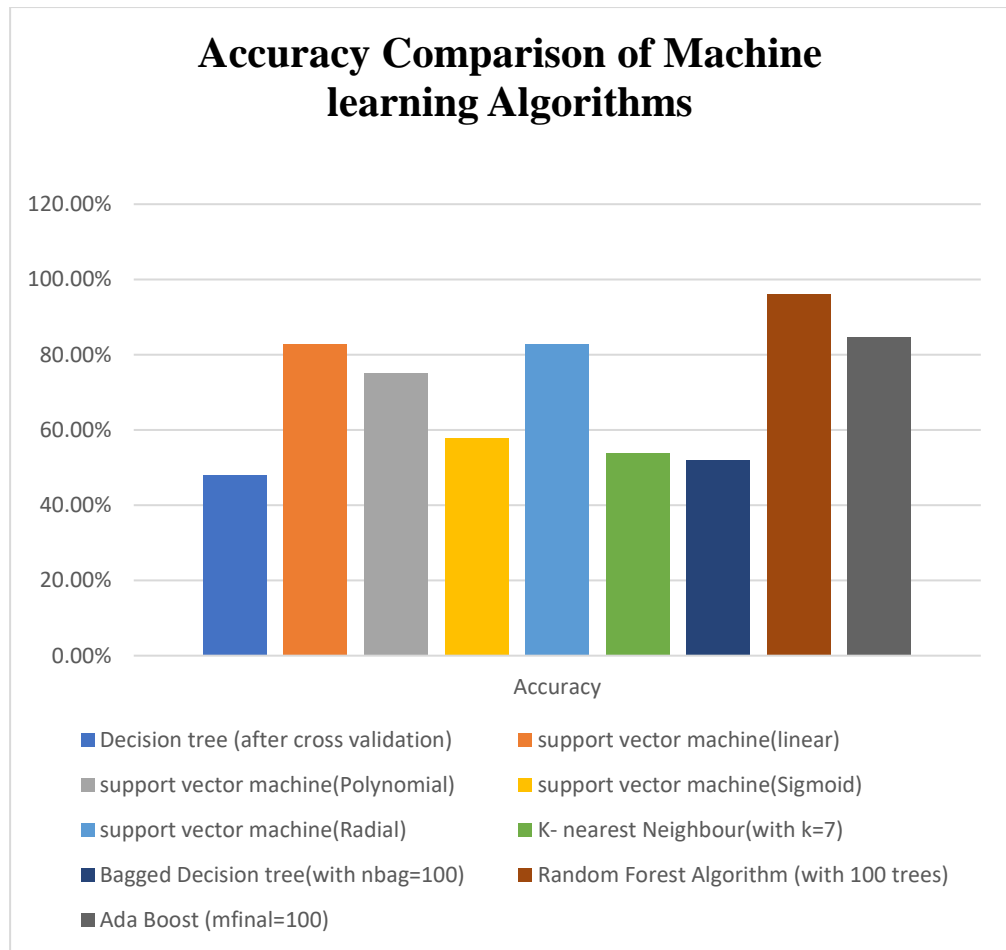


Fig. 7.4 Accuracy Comparison of Machine learning Algorithms for pregnant WRA.

The above-mentioned accuracy results for different machine learning algorithms reveal important information about how well they perform on the given problem. With an outstanding accuracy of 96.15%, the Random Forest Algorithm stands out as the most accurate algorithm. This demonstrates the effectiveness of ensemble approaches in developing robust models. Additionally, the accuracy of 82.69% is produced by both the Support Vector Machines (Linear) and the Support Vector Machines (Radial), both of which perform well. Ada Boost method considerably improves algorithm by obtaining accuracy of 84.61%, while Support Vector Machine (Polynomial) slightly lags behind with accuracy of 75.00%.

The lower accuracy of some methods, such as Bagged Decision Trees and K-Nearest Neighbours with k=7, at 51.92% and 53.84%, respectively, suggests their possible limitations for prediction of Anaemia. The Support Vector Machine (Sigmoid) technique, which also achieves 57.69% accuracy, produces results that aren't particularly good.

In conclusion, the Random Forest algorithm exhibits outstanding accuracy, whereas SVMs, particularly with linear and radial kernels, and the Ada Boost method also stand out as powerful options for prediction of Anaemia in pregnant WRA.

The RF method was utilized in this study because it yields the most accurate results when it comes to determining the main components that are related with the state of anaemia. From the Random Forest algorithm, it was discovered that the factors like 'Age at the Marriage,' 'Weight,' 'BMI,' and 'Eating Habits,' which emphasizing the significance of overall health and lifestyle choices. Additionally, variables like 'Age,' 'Age at Menstrual Cycle Begins,' and 'Days of Blood Flow' which are related to reproductive health, while 'Number of Family Members' and 'Husband's Age at Marriage' underline familial and marital contexts. Socioeconomic dimensions are represented by 'Income of the Family' and 'Education.' Factors like 'HIV Status,' 'Number of Pads per Day,' and 'Average Rest in Day' contributes status of anaemia. Overall it was discovered that the not only individual level factors but household and maternal level factors also affect the status of anaemia.

Now it is aimed to find nature of relationship of above factors with anaemia. The next section deeply explains the relationship of significant factors with anaemia in pregnant WRA.

7.8 Relationship of anaemia and influential factors.

According to the variable importance table of random forest the age at marriage of pregnant women found to be most influential. The age at marriage was numeric variable it was interested to find is for 4 categories of anaemia have equal average age at marriage or different. If we have a categorical variable with two or more categories and a continuous variable, we can use t-test or ANOVA to compare the means of the continuous variable across different groups of the categorical variable. A t-test is used for two groups, whereas ANOVA is suitable for three or more groups. Here we have 4 categories of anaemia so the ANOVA was used and the R output is as follows:

```
> # anova to find relationship between status of anaemia and age at marriage.  
> # Perform ANOVA to obtain the model  
> model <- aov(`Age at the marriage` ~ Anaemia, data = data_pregnant);model
```

Call:

```
aov(formula = `Age at the marriage` ~ Anaemia, data = data_pregnant)
```

Terms:

Anaemia Residuals

Sum of Squares 376.6526 1819.8436

Deg. of Freedom 3 254

Residual standard error: 2.676703

Estimated effects may be unbalanced

> summary(model)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Anaemia	3	376.7	125.55	17.52	2.27e-10 ***
Residuals	254	1819.8	7.16		

Hypothesis:

Null Hypothesis (H0): There is no significant difference in the mean 'Age at the marriage' across different levels of Anaemia status.

Alternative Hypothesis (H1): There is a significant difference in the mean 'Age at the marriage' across different levels of Anaemia status.

Interpretation:

The low p-value ($2.27e-10 < 0.05$) suggests strong evidence to reject the null hypothesis. Therefore, we conclude that there is a statistically significant difference in the mean 'Age at the marriage' across the various levels of Anaemia status.

There is need of further deep examination so, Tuckey HSD test was used. Tukey's Honestly Significant Difference (HSD) test is a post-hoc test commonly used after an ANOVA to identify which specific groups differ from each other in terms of their means. It helps determine pairwise differences between multiple group means while controlling for the family-wise error rate. The output is as follows:

> # Perform Tukey's HSD post-hoc test

> posthoc_tukey <- TukeyHSD(model);posthoc_tukey

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = `Age at the marriage` ~ Anaemia, data = data_pregnant)

\$Anaemia

	diff	lwr	upr	p adj
1-0	-0.3017045	-1.389975	0.7865663	0.8902996
2-0	-0.7454545	-1.856422	0.3655125	0.3075875
3-0	-4.3121212	-5.871671	-2.7525719	0.0000000
2-1	-0.4437500	-1.687674	0.8001742	0.7928107
3-1	-4.0104167	-5.667309	-2.3535239	0.0000000

3-2 -3.5666667 -5.238554 -1.8947790 0.0000005

Interpretation: In summary, for Anaemia Level 3 (i.e. severe anaemia) compared to Levels 0, 1, and 2, (i.e. no, mild, moderate respectively) there are statistically significant differences in the mean ‘Age at the marriage’. For the other comparisons, no significant differences were observed. These results imply that Anaemia Level 3 has a distinctive impact on the ‘Age at the marriage’ compared to other Anaemia levels within the dataset. Therefore, severe anaemia with age at marriage was examined.

Table 7.6 Average of Age at the marriage with anaemia in pregnant WRA.

	No anaemia	Mild	moderate	severe
Average of Age at the marriage	20.14545455	19.8438	19.4	15.833333

The descending trend in the average ‘Age at the marriage’ across Anaemia severity levels (‘No anaemia’ > ‘Mild’ > ‘Moderate’ > ‘Severe’) suggests a potential association between the severity of Anaemia and the timing of marriage. Specifically, it indicates that individuals with more severe Anaemia tend to marry at a comparatively younger age on average, while those without Anaemia marry at a relatively later age. However, this observation is based on the averages and may not capture individual variations within each group.

Relationship between weight of WRA and the Anaemia status:

Table 7.7 Average of weight(kg) of pregnant women.

	No anaemia	Mild	moderate	severe
Average of weight(kg)	58.29090909	53.4063	55.766667	45.833333

The descending trend in the average weight across Anaemia severity levels (‘No anaemia’ > ‘Moderate’ > ‘Mild’ > ‘Severe’) suggests a potential association between the severity of Anaemia and average body weight. Specifically, individuals with more severe Anaemia tend to exhibit a notably lower average weight, while those without Anaemia demonstrate a higher average weight within the studied population. This observation implies a potential relationship between Anaemia severity and body weight, suggesting that individuals with more severe Anaemia levels may have a tendency to have lower body weights on average, while those without Anaemia might tend to have higher body weights on average.

BMI is continuous factors as it was defined by ration of weight in kilograms and height square in meter. Table 7.8 displays the average BMI among all categories of Anaemia in pregnant women in this research.

Table 7. 8 Table of Average BMI of pregnant WRA among anaemia categories.

	No anaemia	Mild	moderate	severe
Average of BMI	24.87	24.37	24.45	19.83

The observed descending trend in average BMI across Anaemia severity levels ('Mild' > 'No anaemia' > 'Moderate' > 'Severe') suggests a potential association between Anaemia severity and BMI. The trend suggests that individuals with 'Mild' Anaemia tend to have the highest average BMI, possibly indicating a higher average body fat percentage within this group. Conversely, individuals with 'Severe' Anaemia tend to have the lowest average BMI, which might suggest lower body fat percentages or reduced weight relative to height compared to other Anaemia levels.

Relationship between Anaemia status and eating habits:

The following table seems to display the distribution of taste preferences among individuals categorized by the severity of anaemia they have.

Table 7.9 Distribution table of taste preferences.

	No Anaemia	Mild Anaemia	Moderate Anaemia	Severe Anaemia	Total
spicy	64	24	36	0	124
Sweet	24	18	12	0	54
sour	2	0	0	0	2
Salty, spicy	0	0	2	0	2
Spicy, sweet	14	18	6	12	50
Sweet, salty	0	2	0	4	6
Spicy, sour	0	2	0	0	2
Sweet, sour	2	0	0	0	2
Salty, sour	0	0	0	4	4
Spicy, Sweet, Salty	0	0	0	4	4
Spicy, Sweet, Sour	4	0	4	0	8
Total	110	64	60	24	

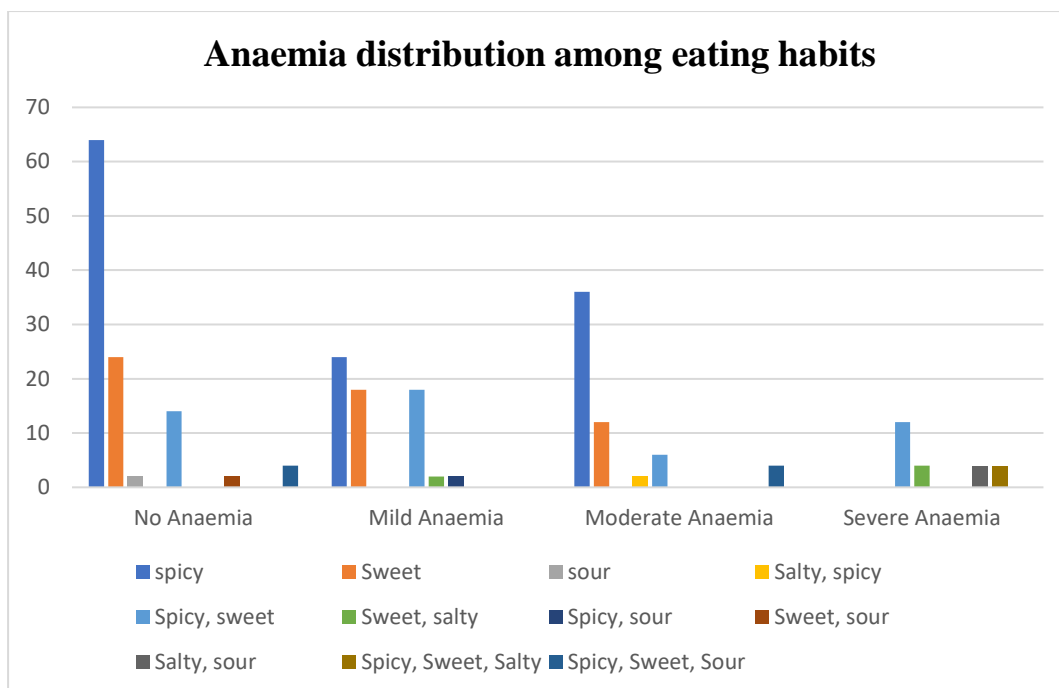


Fig. 7.5 Anaemia distribution among eating habits

Interpretation:

Among those who prefer spicy taste, the majority (64) do not have anaemia, followed by 24 with mild anaemia and 36 with moderate anaemia. None have severe anaemia. Individuals who prefer sweet taste show a similar trend, with 24 having no anaemia, 18 with mild anaemia, and 12 with moderate anaemia. There’s a limited representation of individuals preferring sour taste, with only 2 individuals in the dataset, both without anaemia. Combining taste preferences (e.g. spicy and sweet), different distributions are observed across anaemia levels. Notably, some combinations (like sweet-salty or spicy-sour) have very few individuals, and in some instances, severe anaemia is only present in specific combined taste preference groups. From the table we can’t give any statement about association between eating habits and status of anaemia.

Relationship between alcohol consumption and Anaemia status:

Table 7.10 Alcohol consumption in Pregnant women.

	No anaemia	Mild	Moderate	Severe	Grand Total
No Alcohol	100	56	54	12	222
Alcohol consumption	10	8	6	12	36

The data suggests a potential association between alcohol consumption and anaemia. Individuals who do not consume alcohol generally tend to have higher counts of no anaemia and lower counts of severe anaemia compared to those who consume alcohol. There appears to be a higher prevalence of anaemia across all severities among

individuals who consume alcohol compared to those who do not. However, further statistical analysis such as a chi-squared test could confirm the association between alcohol consumption and the severity of anaemia in this dataset.

To check association between alcohol consumption and Anaemia status chi-square test was used. Following output shows the results of chi-squared test:

```
> t=table(data_pregnant$Anaemia, data_pregnant$`Alcohol Consumption`
+ `)
> t
  0 1
0 100 10
1 56 8
2 54 6
3 12 12
> chisq.test(t)
```

Pearson's Chi-squared test

data: t

X-squared = 29.033, df = 3, p-value = 2.204e-06

In statistical terms, the chi-squared test is used to determine the association between categorical variables. The low p-value (2.204e-06) suggests strong evidence against the null hypothesis. In this case, it indicates that there is a significant association between alcohol consumption and the status of anaemia among the pregnant WRA.

Relationship between age and status of anaemia in pregnant WRA:

Table 7.11 Distribution of average age of pregnant WRA.

	No anaemia	Mild	Moderate	Severe
Average of Age	23.12727273	24.438	25	29.666667

The above table shows that there is a variation in the average age among pregnant WRA across different levels of anaemia severity. Pregnant WRA with no anaemia have an average age of approximately 23.13 years. The average age of Pregnant WRA with mild anaemia is around 24.44 years. Those with moderate anaemia have an average age of 25 years. Pregnant WRA with severe anaemia tend to have a higher average age of about 29.67 years. The data indicates a potential trend where the average age tends to increase as the severity of anaemia becomes more pronounced. This observation could imply that older individuals might be more susceptible to severe forms of anaemia compared to younger Pregnant WRA.

Relationship between Anaemia status and age at menstrual cycle begins:

Table 7.12 Anaemia status and age at menstrual cycle begins with anaemia.

	No anaemia	Mild	Moderate	Severe
Average of Age at menstrual cycle begins	13.45454545	14.1563	14.1	14.666667

The above table indicates that the average age at which pregnant WRA begin their menstrual cycles across different levels of anaemia severity. At a glance, it appears that those with no anaemia have the earliest average age of menstrual cycle onset at approximately 13.45 years. As the severity of anaemia increases from mild to severe, there seems to be a slight trend of a higher average age at the start of menstruation, with the severe category showing the highest average age of approximately 14.67 years. These figures suggest a potential association between anaemia severity and the onset of age at menstrual cycles begins. It was discovered that the delay in menstruation may possibly have severe anaemia.

Relationship between Anaemia status and Number of days of blood flow during menstrual.

Table 7.13 Table of Anaemia status and average number of days of blood flow during menstrual.

	No anaemia	Mild	Moderate	Severe
Average of days of blood flow	4.7	4.65	4.7	2.8

The above table illustrates the average blood flow length throughout menstrual cycles for varying degrees of anaemia severity. It's interesting to note that people classified as having severe anaemia have blood flow on average 2.8 days less frequently than people with no anaemia, mild cases, or moderate cases, whose averages fall between 4.65 and 4.7 days. This discrepancy points to a possible link between a shorter menstrual bleeding length and severe anaemia.

Relationship between Anaemia status and Husband's age:

Table 7.14 Average husband's age of pregnant WRA.

	No anaemia	Mild	Moderate	Severe
Average of husband's age	28.12727273	28.8438	29.8	31.666667

The average age of husbands in a population with several degrees of anaemia severity is shown in the above table. The averages indicate that when anaemia increases from no anaemia to severe anaemia, husband's ages tend to somewhat rise. People who suffer from severe anaemia tend to have husbands who are the oldest on average

roughly 31.67 years old—while people who do not have anaemia tend to have husbands who are the youngest—roughly 28.13 years old.

Relationship between Anaemia status and Number of family members:

Table 7.15 Average family size Vs Anaemia.

	No anaemia	Mild	Moderate	Severe
Average of Number of family members	6.145454545	5.125	5.7	4.833333333

The average number of family members at varying degrees of anaemia severity is displayed in the data table above. There seems to be a correlation between the average family size and the severity of anaemia. Individuals with no anaemia tend to have the largest average number of family members, approximately 6.15, followed by those with moderate anaemia (average of about 5.7 family members), mild anaemia (average of 5.125), and individuals with severe anaemia (average of approximately 4.83 family members). It can be stated that the WRA from small family (nuclear family) have high chance of severe anaemic than that of big family (joint family).

Relationship between Anaemia status and husband’s age at marriage:

Table 7.16 Table of Average of husband’s age at marriage.

	No anaemia	Mild	Moderate	Severe
Average of husband’s age at marriage	25.04545455	25.38	25.833333	25.166667

The information in the table shows the average age of husband at marriage for different degrees of anaemia severity. The averages, interestingly, exhibit a somewhat constant pattern, indicating that the husband’s age upon marriage is generally constant across various anaemia severity groups. People without anaemia had an average husband’s age at marriage of around 25.05 years, and people with severe anaemia have an average husband’s age of about 25.17 years. On the other hand, the average age of those with mild and severe anaemia is marginally higher, at 25.38 and 25.83 years, respectively. These little differences may not suggest a strong relationship between the age at which husbands marry and the severity of their anaemia.

Relationship between Anaemia status and the number of years lives in residential area.

Table 7.17 Average of number of years lives in residential area versus anaemia category in pregnant WRA.

	No anaemia	Mild	Moderate	Severe
Average of Number of years lives in residential area.	5.127272727	5.10938	6.4666667	5.3333333

Based on above table examination, there appears to be very low variation in the number of years spent in the residential area between the various anaemia severity classifications. The longest average stay in residential areas is 6.47 years for WRA with moderate anaemia, followed closely by 5.33 years for those with severe anaemia. In contrast, the average age of WRA without anaemia and moderate anaemia is about 5.13 and 5.11 years, respectively. From the above table it was not get clear interpretation about relationship between Anaemia status and number of years lives in residential area.

Relationship between Anaemia status and Income of the family.

Table 7.18 Relationship between Anaemia status and Income of the family.

	No anaemia	Mild	Moderate	Severe
Average of Income of the family (Rs. Annual)	118527.8182	104063	85100	67500

The information presented in the above tables reflects the relationship between average yearly income of families with different degrees of anaemia severity. WRA with highest average family income have no anaemia, approximately Rs. 118,528 annually. As the severity progresses from mild to moderate and severe anaemia, there is a noticeable decline in the average income, with mild anaemia at an average of Rs. 104,063, moderate anaemia at Rs. 85,100, and severe anaemia at Rs. 67,500 annually. On the conclusion, it seems that the average household income decreases as the anaemia severity rises. In the next association of age of last child in accordance with anaemia category were examined.

Table 7.19 Relationship between Anaemia status and Age of last children in pregnant WRA.

	No anaemia	Mild	Moderate	Severe
Average of Age of last children (month)	4.467272727	4.17188	5.7677778	17

The data presented illustrates the average age of the last child in months for varying degrees of anaemia severity. It's interesting to note that the average age of the last child varies significantly depending on the severity of anaemia. The average age of the last child was much greater in those with severe anaemia (17 months), compared to lower averages of roughly 4.47, 4.17, and 5.77 months in those with no anaemia, mild cases, and moderate cases. This information points to a possible link between an Age of last children and severe anaemia. There may be correlation between anaemia status and the height of women of reproductive age (WRA). Chronic or severe anaemia during critical growth periods may impact height potential due to insufficient oxygen supply

affecting bone growth and development. Therefore, in the next part average height was examined since it gives significant importance while predicting anaemia.

Table 7.20 Relationship between Anaemia status and Height of WRA.

	No anaemia	Mild	Moderate	Severe
Average of Height (meter)	1.530498182	1.48003	1.5123333	1.5250333

The above table shows the relationship between Anaemia status and Height of WRA in meter. According to the figures no any trend was found. So, it was clearly stated that there is no association between anaemia status and height of pregnant WRA.

The relationship between anaemia status and the age at first childbirth can be multifaceted. Anaemia may indirectly influence the age at which a woman has her first child due to its impact on overall health and fertility. Here the Average of Age at first birth of child was studied.

Table 7.21 Relationship between Anaemia status and Age at first birth of child.

	No anaemia	Mild	Moderate	Severe
Average of Age at first birth of child	23	22.0313	22.166667	20.5

The data in the above table illustrates the average age at which WRA gave birth to their first child for each of the different degrees of anaemia severity. A clear pattern appears, pointing to a possible connection between the age of first child-birth and the severity of anaemia. Severe anaemia patients have the lowest average first-childbirth age (20.5 years), suggesting that they tend to give birth to their first child earlier in life. Conversely, the average age at first childbirth for WRA without anaemia was the greatest, at almost 23 years. The average age for WRA with mild and moderate anaemia is 22.03 years, and 22.17 years, respectively. According to this data, there appears to be a pattern where WRA with more severe forms of anaemia typically become parents earlier in life.

The relationship between anaemia status and HIV (Human Immunodeficiency Virus) status among women of reproductive age (WRA) is often interlinked. HIV infection can contribute to the development of anaemia through various mechanisms, including the direct impact of the virus on bone marrow function, decreased production of red blood cells, and increased susceptibility to infections that can cause anaemia. Therefore, this relationship has been studied through following table.

Table 7.22 Relationship between Anaemia status and HIV status of WRA.

		No anaemia	Mild	Moderate	Severe
HIV status	0	110	64	54	4
	1	0	0	6	20

The above table describes the Relationship between Anaemia status and HIV status of WRA. There were total 26 WRA were found to be HIV positive. Out of which 6 are moderately anaemic and 20 are severely anaemic. This result shows that there is strong relationship between HIV status and anaemia status.

The relationship between anaemia status and the number of pads used per day can be connected through menstruation patterns. In some cases, anaemia, particularly iron-deficiency anaemia, might lead to heavier menstrual bleeding (menorrhagia) due to disturbances in the coagulation system and changes in the endometrial lining. Women experiencing heavier menstrual bleeding due to anaemia might require more pads per day to manage their menstruation. To analyse the pattern of this factor following table will be helpful.

Table 7.23 Relationship between Anaemia status and No of pads (per day).

	No anaemia	Mild	Moderate	Severe
Average of No of pads (per day)	2.25	2.34375	2.166666667	3.833333333

The association between the average daily pads usage and anaemia status reveals a slight variation between anaemia severities. Those without anaemia appear to use 2.25 pads on average daily, which is marginally more than those with moderate anaemia who use roughly 2.34 pads daily, according to the data. Remarkably, WRA with moderate anaemia appear to use on an average 2.17 pads a day, less than those without the condition. On the other hand, WRA with severe anaemia use a lot more pads an average of 3.83 pads each day This pattern suggests that when monthly bleeding increases WRA gets severe anaemic. There may be a correlation with increased monthly bleeding and anaemia severity. there is need for menstrual care and awareness.

CHAPTER 8

SUMMARY AND CONCLUSION

8.1 Introduction:

Without women, we are unable to picture how successful life would be in general. They have a major share of the blame for the continued success of life on this planet. The success of sustainable development and family life depends on women. The different roles that women play in the family include those of wife, head of the household, administrator, manager of finances, and last but not least, mother. In the past, they were only thought of as wives and mothers who had to cook, clean, and care for the entire family by themselves. But now that their condition has slightly improved, they have begun engaging in activities other than those with their family and children. She has to take care of herself and family members as daughter, granddaughter, sister, daughter-in-law, wife, mother, mother-in-law, grandmother, etc. By following such a big responsibility in the family, they are fully able to come out and do job for bright future of own, family and country. Women should take care of their own health in addition to their responsibilities, but they frequently overlook it. Being a male or a woman has a significant impact on one's health because of differences in biology and gender. The health of women and girls is a special concern because, in many countries, they are subjected to discrimination because of socio-cultural difficulties. While poverty is a major barrier to optimal health outcomes for both men and women, it tends to have a more detrimental effect on the health of women and girls because of things like the use of hazardous cooking fuels and malnutrition in feeding habits (COPD). In comparison to men, women experience certain healthcare challenges. Chronic illnesses and conditions include heart disease, cancer, diabetes, and anaemia are among the leading causes of death for women.

A disorder known as anaemia occurs when the red blood cells have insufficient haemoglobin or too few red blood cells. Your blood's capacity to carry oxygen to the body's tissues will be diminished if you have either normal or inadequate haemoglobin, red blood cells, or both. Among the symptoms of anaemia include weakness, exhaustion, fainting, and shortness of breath. The optimal haemoglobin concentration needed to meet physiological demands is influenced by a number of factors, including age, sex, elevation of residence, smoking status, and pregnancy. The leading causes of anaemia include nutritional deficiencies in folate, vitamins B12 and A,

haemoglobinopathies, viral diseases such malaria, tuberculosis, HIV, and parasite infections, and dietary inadequacies, particularly iron deficiency.

Pregnant women who have anaemia run the risk of having a number of issues that may harm both the mother and the unborn child. A low birth weight that may cause health issues for the newborn and an increased chance of preterm birth when the baby is born before reaching full term are only two of the negative effects of maternal anaemia. Anaemia has a significant influence on women who are of reproductive age, affecting not only their overall health but also their capacity to bear children. The anaemia may make it more difficult for them to become pregnant and may result in heavy monthly flow, irregular menstrual cycles, and increased tiredness. Preterm birth, low birth weight, and maternal problems are among the pregnancy complications that anaemic women are more prone to encounter. In addition to producing fatigue and a delay in healing, anaemia can exacerbate the healing process following childbirth. Treating and controlling anaemia is crucial for women who are of reproductive age in order to support their overall health, reproductive health, and potential pregnancy outcomes.

Analysing anaemia in women is imperative as it provides crucial insights into the public health impact of this condition, particularly concerning maternal and child health, economic repercussions, and gender inequalities. It helps identify high-risk groups, assess the effectiveness of interventions, and guide research and innovative solutions, all of which are vital for improving the overall health and well-being of women and their communities. Previous research on the prevalence of anaemia in women has utilized a range of statistical methods to provide a comprehensive understanding of the condition's impact. Descriptive statistics, such as mean, median, and standard deviation, have been employed to quantify the central tendency and variability of haemoglobin levels in female populations. Prevalence rates, often expressed as percentages, have been calculated to estimate the proportion of women affected by anaemia in various age groups and regions. Regression analysis has been applied to assess the relationship between anaemia and factors like age, socioeconomic status, and dietary patterns. These statistical approaches have not only quantified the extent of the problem but have also allowed for the identification of vulnerable subgroups and the assessment of risk factors, thereby informing targeted interventions and policy decisions to address anaemia in women. In recent researches mostly, various

statistical techniques have been used to analyse the anaemia in WRA. Advanced machine learning methods were used to predict the anaemia in WRA.

Chapter I elaborately discussed research background, significance, objectives, and scope of the ‘Comparative Study of Machine Learning Algorithms for the Prediction of Anaemia among Women at Reproductive Age’. In the second critical examination and synthesis of existing scholarship and research related to the topic under investigation was done. Through a systematic review and synthesis of existing literature, this chapter not only identifies key theories, methodologies, findings, and debates but also critically evaluates their strengths, limitations, and gaps. It provides a comprehensive overview that highlights the evolution of ideas, the progression of research.

The third chapter is methodology chapter, which is typically encompasses various components, such as research design, data collection methods, participant selection criteria (if applicable), instrumentation or tools employed, and data analysis procedures. It delineates the rationale behind the chosen approach, whether qualitative, quantitative, mixed methods, or other specialized methodologies. For this research the two data sets were used. The pilot study was done on the DHS (2015) data. Initially there were 6,99,686 samples of all India. For Maharashtra state 28,648 samples were extracted. Data pre-processing was done after the data pre-processing 179 samples are taken for analysis purpose. In the final data there were total 46 variables. Dataset covers a wide range of information, including health-related factors like anaemia, pregnancy status, and various dietary habits, such as consumption of specific food items. Socio-economic indicators like household wealth, education levels, and employment status are also included. Demographic details like age, marital status, and family size are key components. Moreover, the dataset encompasses lifestyle choices like alcohol and tobacco consumption. Decision tree and Random Forest algorithm were developed for the classification of anaemia. The second data set used was primary data. Which is gathered from the well-designed questionnaire. The dataset contains information about women at reproductive age (WRA) between the ages of 14 to 49. The data was collected from the hostel girls, pregnant women and non-pregnant women. Therefore, the main data itself has three sub datasets. The parameters or questions are slightly different for the Unmarried women, married women and married women with pregnancy. It encompasses a diverse set of 56 variables that capture a wide range of characteristics related to the participants’ health, lifestyle, and socioeconomic status.

Out of these 56 some are not used or applicable for unmarried women like pregnancy related questions, contraceptive related questions, etc. Similarly, some of questions not applicable for nonpregnant WRA such as pregnancy related questions.

A pilot study is a small-scale preliminary inquiry that is conducted in order to assess and refine research approaches, procedures, and data gathering tools prior to the major research. It serves as a crucial pre-research procedure by providing researchers with insights into the feasibility, applicability, and potential issues of their study. In order to look into the elements that influence anaemia, DT and RF algorithms were designed for the pilot study in the fourth chapter. To assess the model, a confusion matrix was used. A questionnaire was created employing those key criteria in order to collect primary data.

Chapter five consists of ‘Comparing the Performance of Machine Learning Algorithms on Unmarried WRA’. The various ML algorithms were developed/built on the unmarried women dataset to predict anaemia. Among all developed algorithms best algorithm was identified and according to that best algorithm the influential factors associated with anaemia was examined with help of central tendencies, percentage and counts.

In the chapter 6 entitled as ‘Comparing the Performance of Machine Learning Algorithms on Married Non-Pregnant WRA’. The same methodology like chapter 5 was applied on this chapter and key factors associated with anaemia status was identified. In this chapter traditional machine learning models fails to achieve optimum accuracy so, the stacking ensemble techniques was used. ‘Comparing the Performance of Machine Learning Algorithms on Married Pregnant WRA’ is the 7th chapter. Here the machine learning algorithms were developed to study the anaemia prevalence in married WRA. Also examined the significant factors associated with the status of anaemia.

8.2 Summary and Conclusion:

The pilot study was done by using DHS data in 4th chapter. The first objective of this research is to determine prevalence of anaemia among WRA in Maharashtra. Regarding the incidence of anaemia in India. 331,619 of the 684,189 patients were classified as having "No Anaemia." The following three most common categories are "Severe," with 6,950 occurrences, "Moderate," with 82,490 cases, and "Mild," with 263,130 cases. If dividing the population into two categories those with and without anaemia here it was found that, out of the WRA, roughly 61% had anaemia, which is

extremely harmful to their health. According to the Maharashtra data, 53% of the individuals in the group or community being studied do not have anaemia. The category comprises a sizeable portion of the population (36%) who suffer from mild anaemia, a lesser percentage who have moderate anaemia, and a very small number who have severe anaemia. Although it is positive that 53% of women do not have anaemia, DHS statistics show that 47% of WRA were found to be anaemic when comparing the anaemic and non-anaemic categories.

This distribution illustrates a noteworthy incidence of anaemia in Maharashtra's population, with varied degrees of severity. Especially, the high frequency of mild cases indicates a significant health issue in this area. It was remarkable that comparatively high portion of women who do not exhibit any symptoms of anaemia, which accounts for a sizeable segment of the population.

Based on state-specific data, the states with the highest percentage of anaemic WRA include West Bengal, Bihar, Zarkhand, and Haryana. Compared to non-pregnant WRAs, a higher proportion of pregnant WRAs did not have anaemia. Although percentage of mild anaemia was found higher in non-pregnant WRA, Percentage of moderate anaemia was found to be higher in pregnant WRA.

Women with moderate anaemia are at risk, and without proper care and intervention, they face an increased likelihood of progressing to severe anaemia, emphasizing the importance of timely and appropriate medical attention to prevent worsening health outcomes. Anaemia may prevent the foetus from growing as optimally as possible throughout pregnancy. As a result, the baby can have stunted growth, impaired cognitive performance, and increased susceptibility to infections. Moderate anaemia during pregnancy may pose serious health risks to both the mother and the growing child. It may cause problems during pregnancy and labour, such as low birth weight, early delivery, or even maternal death. This issue needs to be addressed in order to safeguard the health and welfare of expecting mothers and their unborn children. Interventions such as a nutritious diet, iron supplements, prenatal care, and medical monitoring are essential to treat and prevent anaemia during pregnancy.

While considering demographic area of the women it was found that the proportion of non-anaemic women is lower in urban areas than in rural areas, the proportion of mild, moderate and severe anaemia was found to be high in rural areas. The higher prevalence of anaemia in rural areas compared to urban areas can be attributed to various factors such as limited access to healthcare, dietary deficiencies,

poor sanitation, and socioeconomic disparities. Addressing the higher prevalence of anaemia in rural areas requires a multi-faceted approach involving healthcare, nutrition, sanitation, education, and community engagement to create sustainable and impactful changes.

According to the pilot study, the RF algorithm outperforms the DT algorithm in terms of accuracy. Therefore, the RF method was used to identify significant determinants for the status of anaemia in WRA. The Random Forest model predicts anaemia severity based on key predictors such as currently working status, husband's job, wealth status, residency years, age at first birth, household size, education and age of WRA. Understanding these factors can guide targeted interventions and public health strategies to effectively address anaemia, highlighting the multifaceted nature of the disease and its impact on socioeconomic, demographic, and health-related factors.

It was found that a working woman can be susceptible to anaemia at varying degrees of severity. The chances of anaemia were high in those WRA whose husband's occupation was job or other. Research has indicated a possible trend, women of reproductive age who suffer from severe anaemia typically have greater average wealth levels than those who have no anaemia, mild anaemia, or moderate anaemia. The average length of residency was longest among women who do not have anaemia. Compared to women with mild or moderate anaemia, those with severe anaemia have an average residence time that was comparatively higher. The research reveals that the anaemia severity was increases as the age of women at first birth was less. Early childbirth might contribute to higher susceptibility to anaemia due to the physiological stress of pregnancy at a younger age, potentially impacting nutritional status and iron reserves. Further research exploring this relationship can aid in developing targeted interventions to support the health of young mothers and reduce the prevalence of anaemia in this demographic. It seems to suggest that WRA with severe anaemia usually have a higher average biological age of the household head than WRA in other anaemia groups. The relationship between anaemia severity and educational years was found to be inverse, with a decrease in educational years there was rise in anaemia severity from mild to severe. With an average age of about 34.45 years, severe anaemia has the highest average age of the groups. It suggests that WRA with higher levels of anaemia are frequently older than those with lower or no anaemia. These conclusions were made on the basis of pilot data.

Using the results of pilot study, the questionnaire was designed to collect primary data. As the anaemia cut-off is different for expecting women and non-pregnant women then original data was divided into three parts unmarried WRA, married non-pregnant WRA and pregnant WRA. In the 5th chapter unmarried WRA was examined. Initially there were 182 unmarried WRA. Since the data was small firstly decision tree model was developed on whole data and the model performance shows 78% accuracy of the fitted decision tree algorithm. Since there are some drawbacks of developing machine learning algorithm without train- test splitting criteria, the new decision tree algorithm was developed according to 70-30 pattern. Where 127 trained data were used to develop decision tree algorithm and 55 test data was used to test the fitted decision tree algorithm. The developed decision tree algorithm shows approximately 51% accuracy. When 127 samples, or a smaller amount of data, were used to train the same model, the accuracy fell to 50%. This disparity suggests that the model may not have been optimally trained on the training set of data. Stated differently, it became excessively narrow and customised to the new features of the 127 samples. The 10-fold cross validation technique was employed to address this problem. Ten folds, or subgroups, were created from the data in order to apply the 10-fold cross validation technique. The CART algorithm was trained, tested, and its performance evaluated using these folds. After observing all cp values the $cp = 0.06451613$ was found to be best as it gives accuracy 66.04%. The decision tree model with cp value 0.06451613 was used to develop the new model. The fitted new model shows 65% accuracy however there was some complications while examining confusion matrix. The confusion matrix indicates that no predictions exist for classes 0, 2, and 3, suggesting that the model is only able to predict class 1 for all cases in the dataset. This could be caused by a number of things, including imbalanced data or an issue with the model training process. To evaluate the model performance the assessment measures such as accuracy, recall, precision, and F1 score can be employed. Using the provided confusion matrix, it is impossible to accurately calculate these metrics since there are no predictions for classes other than 1.

There was indication of necessity of additional data. Due to time limit and availability of resources only 28 additional sample able to collect. The new data consists of 216 WRA. There were some missing values so after cleaning there were 203 samples of unmarried WRA. Out of 203 WRA 97% unmarried WRA were found to be anaemic. Only 3% of unmarried women found to be anaemic. Here all unmarried WRA in the

sample are hostel girls. Due to various problems in hostel there may be health issues in girls. Living in a hostel may have an impact on health of WRA. The erratic meal schedules, a dependence on fast food, or restricted availability of wholesome meals results a inverse impact on the status of anaemia.

To predict anaemia and explore the key factors associated with anaemia decision tree model with 10 fold cross validation were developed with cp value 0.1071429 gives better accuracy than other. In the next section decision tree model with this optimum cp value was developed , it gives approximately 61% accuracy. Still it was not considerable therefore looking forward to better results the SVM algorithm with various kernels were developed and performance was observed. The linear support vector machine gives comparatively high 62.29% accuracy and minimum no. of support vectors. Since the linear kernel gives better accuracy than that of other kernel we can say that the data is linearly separable.

The k-nearest neighbour algorithm was developed to achieve better accuracy. At the beginning to select optimum k value 10-fold cross validation were used. The K-NN algorithm with k=7 gives comparatively better accuracy among k=5,7,and 9. The KNN with k=7 achieves only 60% accuracy. Here, individual machine learning models, such as DT, SVM, and KNN, were first tested with in an attempt to increase accuracy and improve predictive power. These models yielded accuracy levels of 65%, 62.29%, and 60%, respectively. However, in order to strategically go forward and acknowledge the need for additional enhancement, ensemble models must be built. Bagged decision tree with different number of trees were developed, it was observed that bagged decision tree with 100 decision tree gives 65% accuracy. Although using a Bagged Decision Tree to get an accuracy of 64.79% is an acceptable start, there's always space for improvement. Investigating other ensemble procedures is a wonderful option if you're going for greater accuracy. The Random Forest algorithm was applied to the identical set of data in the next section.

The random forest model was constructed with 100 decision trees. Since the OOB is 35.21% that means model misclassifies about 35.21% of the data points that were not part of the training process. After the confusion matrix evaluation approximately 69% accuracy was observed for the random forest algorithm. Even if a RF model's 69% accuracy rate is a good result, more improvements must be taken into account for a better prediction performance. Developing an AdaBoost algorithm might be a strategic step to increase accuracy. Therefore, The boosting model, with a total of

100 boosting iterations ($m_{\text{final}} = 100$), has been applied to the 'Anaemia' prediction. The error rate was calculated to be 0.295082. ADA Boost throws 70% accuracy which is high than all previously developed algorithms. At the top of the list, 'Weight' stands out as the most significant predictor variable. This suggests that a woman's weight has a significant influence on her anaemia status. The following are 'Age' and 'BMI' (body mass index). These predictor variables show how important they are in the context of the discussion by having a major impact on the anaemia status. 'Number of Years Living in Residential Area' and 'Age at Menstrual Cycle Begins' also hint to their influence on the anaemia status. In relation to anaemia, 'Height' and 'Family Income' are also thought to be highly significant. As proceed down the list, further information is added by 'HIV Status,' 'Days of Blood Flow,' 'No. of Pads per Day,' and 'Number of Family Members'. The parameters 'Cooking Fuel,' 'Daily Tea Intake,' 'Occupation,' 'Exposure to Domestic Violence,' and 'Eating Habits' are relevant despite being ranked far lower. The variables at the top of the list are the key drivers of the anaemia in unmarried WRA, while those at the bottom, though less influential, still contribute to the model's overall understanding and predictive power.

Monitoring and maybe changing these variables may have a significant effect on the anaemia in single WRA. In order to possibly lower the anaemia in single WRA, it's critical to take into account therapies or tactics pertaining to these aspects. While examining weight and anaemia status the data illustrates a clear association between the severity of anaemia and the average weight of individuals. As the severity of anaemia increases from mild to moderate and severe, the average weight tends to decrease. These findings could have important implications for healthcare and nutrition interventions, as they suggest that weight monitoring and nutritional support may be particularly crucial for Unmarried WRA with more severe forms of anaemia to improve their overall health and well-being.

The data reveals that an intriguing relationship between the severity of anaemia and BMI. Individuals with more severe forms of anaemia tend to have significantly lower average BMI values. This observation suggests that severe anaemia can be associated with a decreased BMI, potentially indicating a connection between the health status of WRA with severe anaemia and their nutritional well-being. Monitoring BMI in individuals with anaemia, especially severe cases, is crucial for healthcare professionals to tailor appropriate interventions and support to improve their overall health.

When age factor take into account it shows that, the highest occurrences are observed at ages 16-20 years, suggesting that mild anaemia is more common in younger women 'Moderate anaemia' found in age 16-19 years. The data suggests that individuals with mild and moderate anaemia may have a reduced height compared to those without anaemia. However, the trend shifts for individuals with severe anaemia, who exhibit a somewhat taller average height. In the 'Severe Anaemia' category, the average family income notably decreases to Rs. 51,250 per year. There appears to be a higher incidence of severe anaemia among those infected with HIV. This may suggest that, in comparison to HIV-negative people, HIV-positive people are more prone to suffer from severe anaemia. This information might be useful for healthcare professionals and policymakers to better understand the health conditions of individuals with HIV and to tailor interventions accordingly. Further statistical analysis, such as chi-square tests, could provide more insights into the strength and significance of the observed associations. There appears to be a trend where the average number of family members decreases as the severity of anaemia increases. WRA with no anaemia have the highest average number of family members, while those with severe anaemia have the lowest average number.

The analysis of non-pregnant married women was done in the chapter 6. The distribution of anaemia prevalence rates within the population is worrying when looking at the specific categories. It was observed that 70% women affected by anaemia in various degree of severity. In the first stage traditional ML algorithms were developed but no one can give better performance as accuracy between 50-55%. In the next stage ensemble learning approach was used by using bagged decision tree, random forest and Ads Boost but accuracy was remained in the rage 50-55%. Ultimately the stacking ensemble approach were implemented. In the stacking ensemble DT by CART, SVM with linear kernel, KNN, RF, Bagged decision tree, and GBM was trained by 10- fold cross validation techniques. These models are tested on test data and predictions are noted to make a new dataset. The new data has 7 variables one is anaemia status which was dependent variable and remaining 6 considered as the independent variables to predict the anaemia status. That new data was again partitioned into train and test data. Now the meta model by random forest was trained on this new trained data and tested on new test data. Meta model gives 70% accuracy which was accepted for further study. From the variable significance Pboost variable shows highest significance score than other 5. So, the significant factors associated with

anaemia were extracted from the gradient boost algorithm. Based on the given dataset and model, these results suggest that certain physiological characteristics (like reported symptoms, BMI) and demographic data (like age-related factors, residential details) may have a considerable impact on the risk of anaemia incidence. To fully comprehend the intricate interactions between these factors and provide an accurate prediction of anaemia, more investigation and research may be required. At the last it was found that if the WRA feeling weak or dizziness then there is high probability of getting anaemic. This pattern points to a significant correlation between the prevalence of weakness or dizziness and the degree of anaemia. Notably, individuals with severe anaemia had the lowest average BMI, averaging about 27.29. This study suggests that the average BMI tends to decrease as anaemia worsens, suggesting a negative link between anaemia severity and BMI. The WRA with severe anaemia have found to be the lowest average number of years lived in the residential area, with an average of 8 years. Study reveals that older age WRA have high chance of sever anaemic. there was inverse relationship found between anaemia severity and the weight of respective WRA. As we can see here as if weight decreases the anaemia severity increases from mild to severe. From this we can suggest that to avoid anaemia in non-pregnant WRA we have to aware WRA about the weight factor. There is a discernible relationship between women's anaemia severity and declining income levels. Inadequate financial means frequently impede the ability to obtain iron-rich foods, healthcare facilities, and a balanced diet, which greatly increases the incidence and severity of anaemia in lower-class communities.

In chapter 7 pregnant WRA taken into account for research point of view. The same methodology like chapter 5 and 6 was used in this chapter. 258 pregnant WRA were included in this research. The results show that 57.36 % pregnant women found to be anaemic. This shows that the various degrees of anaemia's severity demand attention and possible action as a relevant health problem in this population. But as compared to previous two groups of WRA this group has high prevalence of no anaemia which is quite good. First decision tree by cross validation with 10-fold shows 54% accuracy for 0.04954955 complexity parameter. By using this cp value decision tree was build, after testing the model performance it gives 48% accuracy which was not good enough. Move forward to SVM classifier the cross-validation results show accuracy 85% with the cost parameter is held constant at 1. SVM classifier with C=1 for various kernels were developed and by comparing the accuracy it was discovered that the linear and radial kernel shows highest accuracy (82.69%) among other two the

kernels. It shows the pregnant women data was linearly separable. KNN classifier doesn't perform well to this data as it throws only 54% accuracy. After the individual classifier ensemble techniques were employed. The bagged decision tree with 100 trees were build but it didn't showcase good performance since it has 52% accuracy. Random forest was studied in this context. RF classifier with 100 trees gives 96% accuracy which was highest among all previously fitted classifiers. Since the Ada Boost classifier remain in the list it was developed on same scale of 100 trees it has 85% accuracy which was found to be good.

Since the random forest shows outstanding accuracy, Random Forest algorithm was used to identify significant factors associated with anaemia status. Factors such as age at marriage, weight, BMI, and eating habits were identified as crucial for overall health and lifestyle choices. Reproductive health variables like age and menstrual cycle began and days of blood flow were also considered. Socioeconomic dimensions like family income and education were also considered. Factors like HIV status, daily pad usage, and average rest in the day also contributed to anaemia status.

Based on the observation of these variables, it may be concluded that people with more severe anaemia typically marry younger, while people without anaemia marry later. It is possible that WRA with more severe anaemia may typically weigh less overall, whereas WRA without anaemia may weigh more overall. In line with this, WRA with "Severe" anaemia typically have the lowest average BMI, which may indicate lower body fat percentages or a smaller weight in relation to height than in other anaemia categories. Individuals who do not consume alcohol generally tend to have higher percentage of no anaemia and lower percentage of severe anaemia compared to those who consume alcohol. There were 33% alcoholic WRA are severe anaemic. Women's ages show a possible tendency wherein an increasing average age may be associated with a more noticeable degree of anaemia. It was discovered that the delay in menstruation may possibly have severe anaemia. It's interesting to note that people classified as having severe anaemia have blood flow on average 2.8 days less frequently than people with no anaemia, mild cases, or moderate cases, whose averages fall between 4.65 and 4.7 days. Women who suffer from severe anaemia tend to have husbands who are the oldest on average roughly 31.67 years old while people who do not have anaemia tend to have husbands who are the youngest roughly 28.13 years old. It can be stated that the WRA from small family (nuclear family) have high chance of severe anaemic than that of big family (joint family). It seems that the average

household income decreases as the anaemia severity rises. A clear pattern appears, pointing to a possible connection between the age of first child-birth and the severity of anaemia. Severe anaemia patients have the lowest average first-childbirth age (20.5 years), suggesting that they tend to give birth to their first child earlier in life

After studying all three population of women under study this study reveals that age, BMI, Number of years lives in residential area., Number of family members, annual family income, Days of blood flow, and Number of family members found to be commonly significant while predicting status of anaemia in women. The nature of these factors in accordance with anaemia status was already studied in the respective chapter. If only married women take into account it was found that BMI, Number of years lives in residential area, age, age at marriage, age of last children (in month), husband's age, husband's age at marriage, annual family income, number of days of blood flow during menses, and alcohol consumption were commonly significant for the prediction of anaemia in married WRA.

8.3 Recommendations:

Empower women by providing them with knowledge and resources to take charge of their health. Encourage them to seek healthcare when needed and make informed choices regarding nutrition and family planning. Advocate for policies that support the nutritional needs of women, especially in areas with high rates of anaemia. This could involve subsidizing or providing free iron-rich foods or supplements. Addressing high prevalence of anaemia in women requires a multi-faceted approach, considering socio-economic, cultural, and healthcare-related factors. Implementing these recommendations can significantly improve the situation and reduce the burden of anaemia in these specific states.

To mitigate the impact of hostel life on haemoglobin levels, it's crucial for students to prioritize a balanced diet, incorporate iron-rich foods, manage stress through relaxation techniques or counselling, maintain good hygiene practices, seek prompt medical attention for any health concerns, and strive for adequate sleep and rest despite the challenges of hostel life. Regular health check-ups and awareness about maintaining a healthy lifestyle are essential to monitor and sustain optimal haemoglobin levels.

The greater frequency of mild anaemia may point to a generalised, albeit manageable, health issue, whereas the presence of moderate and severe anaemia in a significant section of the population may necessitate immediate attention and treatment. Policymakers and healthcare providers must make sure that.

This finding emphasises the necessity for focused treatments and healthcare programmes designed with urban populations in mind in order to address and lessen the greater prevalence of anaemia among women living in metropolitan regions.

There was correlation with increased monthly bleeding and anaemia severity there is need for menstrual care and awareness. Ensure easy access to healthcare services for individuals experiencing heavy menstrual bleeding. This includes access to gynaecologists, haematologists, or other specialists who can diagnose and manage menstrual disorders and anaemia effectively.

Pregnancy-related factors contribute significantly to anaemia, advocate for regular antenatal care and screenings during pregnancy. Encourage iron supplementation for pregnant women as per healthcare provider recommendations. Promote awareness about the importance of nutrition and prenatal vitamins during pregnancy.

Early Marriage and Pregnancy: In regions where, early marriage and early pregnancies are prevalent: Advocate for policies and programs that discourage early marriage and promote education for girls. Encourage comprehensive sex education to raise awareness about the risks associated with early pregnancy, including anaemia, and promote family planning.

One of the objectives of this research is to find best model to predict anaemia. As we can see for three different data sets we get three best machine learning models for the prediction of anaemia. For the DHS data Random Forest with 100 trees shows 64% accuracy. Since it was pilot study this accuracy was acceptable. On the first data set which is unmarried WRA the Ada Boost algorithm shows 70.49% accuracy. The Non-pregnant data was critical data since various machine learning models single as well as ensemble were failed to give considerable predictive performance. Stacking ensemble technique was gives useful insights in this condition. Stacking ensemble gives near about 70% accuracy. For Pregnant WRA random forest model with 100 tress gives 96.15% accuracy which was impressive.

While comparing all the models the research shows that the hyper-parameter tuning was very necessary to enhance the predictive performance of model. Finding the ideal set of hyperparameters for a particular model in order to maximise accuracy or performance on a validation dataset is the key step towards the fetching the best predictive model. In dealing with real-world data, the utilization of ensemble techniques emerges as a strategic solution. The inherent imbalance within the dataset, often

unavoidable in real-life scenarios, poses a challenge for traditional machine learning models. However, ensemble methods, such as Random Forests, Gradient Boosting Machines (GBM), AdaBoost, or XGBoost, prove invaluable in handling this imbalance without necessitating direct alterations to the dataset. By leveraging the collective wisdom of diverse base models, ensemble methods amalgamate predictions from multiple sources, allowing the model to glean insights from both majority and minority classes. This robust approach not only helps in mitigating the impact of class imbalance but also enhances predictive performance. Through techniques like weighted voting, boosting, or bagging, these ensemble models intelligently balance the influence of each class, effectively addressing the challenge of imbalanced data in a real-world context.

8.4 Future scope:

1. **Technology and Innovation:** Developing innovative diagnostic tools, mobile health applications, or wearable devices for early detection and monitoring of anaemia in resource-limited settings, enabling timely interventions.
2. **Menstrual Health and Bleeding Disorders:** Investigating the prevalence, causes, and impact of menstrual disorders leading to heavy bleeding and subsequent anaemia, as well as evaluating the effectiveness of various treatments and interventions.
3. **Psychosocial and Cultural Determinants:** Exploring the psychosocial impacts of anaemia on women's quality of life, mental health, and productivity, considering cultural beliefs and social determinants influencing treatment-seeking behaviour.
4. **Big Data Analytics:** Utilizing big data analytics to process large datasets from healthcare records, national surveys, or electronic health records to identify trends, patterns, and risk factors correlated/associated with anaemia in women across different demographics and geographical regions.
5. **Meta-Analysis and Systematic Reviews:** Conducting comprehensive meta-analyses and systematic reviews to synthesize existing research findings, evaluate the overall impact of interventions, and identify gaps or inconsistencies in the literature on anaemia in women.
6. **Survival Analysis:** Employing survival analysis techniques to assess the time to recovery from anaemia, recurrence rates, and factors influencing the duration of anaemia episodes among women, considering various covariates.

8.5 Limitations of research:

1. **Data Availability and Quality:** Due to time and availability of resources study was done on small scale.

2. Sampling Bias: The study sample not entirely representative of the entire population of women, leading to sampling bias. The present study was only including participants from Baramati city and nearest villages.

CHAPTER:9

REFERENCES

1. Uddin, N., Khabir Uddin Ahamed, Md., Uddin, M. A., Manwarul Islam, M., Alamin Talukder, Md., & Aryal, S. (2023). An Ensemble Machine Learning Based Bank Loan Approval Predictions System with a Smart Application. *International Journal of Cognitive Computing in Engineering*. <https://doi.org/10.1016/j.ijcce.2023.09.001>
2. Yadav, J., & Nilima, N. (2021). Geographic variation and factors associated with anaemia among under-fives in India: A multilevel approach. *Clinical Epidemiology and Global Health*, 9, 261–268. <https://doi.org/10.1016/j.cegh.2020.09.008>
3. Polamuri, S. R., Veerababu, P., Rao, S., & Prof, A. (n.d.). Study on Machine Learning Techniques to Predicting Heart Disease. <https://www.researchgate.net/publication/356171813>
4. Radha, P. (2021). Disease Classification and Prediction Using Ensemble Machine Learning Classification Algorithm. Article in *International Journal of Recent Technology and Engineering*, 9, 2277–3878. <https://doi.org/10.39621/ijrte.F5507.039621>
5. Nishat, M. M. (2021). Performance Assessment of Different Machine Learning Algorithms in Predicting Diabetes Mellitus. *Bioscience Biotechnology Research Communications*, 14(1), 74–82. <https://doi.org/10.21786/bbrc/14.1/10>
6. Mog, M., & Ghosh, K. (2021). Prevalence of anaemia among women of reproductive age (15–49): A spatial-temporal comprehensive study of Maharashtra districts. *Clinical Epidemiology and Global Health*, 11. <https://doi.org/10.1016/j.cegh.2021.100712>
7. Tirore, L. L., Mulugeta, A., Belachew, A. B., Gebrehaweria, M., Sahilemichael, A., Erkalo, D., & Atsbha, R. (2021). Factors associated with anaemia among women of reproductive age in Ethiopia: Multilevel ordinal logistic regression analysis. *Maternal and Child Nutrition*, 17(1). <https://doi.org/10.1111/mcn.13063>
8. Bari Antor, M., Jamil, A. H. M. S., Mamtaz, M., Monirujjaman Khan, M., Aljahdali, S., Kaur, M., Singh, P., & Masud, M. (2021). A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease. *Journal of Healthcare Engineering*, 2021. <https://doi.org/10.1155/2021/9917919>
9. Sunuwar, D. R., Singh, D. R., Adhikari, B., Shrestha, S., & Pradhan, P. M. S. (2021). Factors affecting anaemia among women of reproductive age in Nepal: A multilevel and spatial analysis. *BMJ Open*, 11(3). <https://doi.org/10.1136/bmjopen-2020-041982>

10. Khan, J. R., Chowdhury, S., Islam, H., & Raheem, E. (2021). Machine Learning Algorithms To Predict The Childhood Anaemia In Bangladesh. *Journal of Data Science*, 17(1), 195–218. [https://doi.org/10.6339/jds.201901_17\(1\).0009](https://doi.org/10.6339/jds.201901_17(1).0009)
11. Yeruva, S., Gowtham, B. P., Chandana, Y. H., Varalakshmi, M. S., & Jain, S. (2021). Prediction of Anaemia Disease Using Classification Methods (pp. 1–11). https://doi.org/10.1007/978-981-33-4046-6_1
12. Karagül Yıldız, T., Yurtay, N., & Öneç, B. (2021). Classifying anaemia types using artificial learning methods. *Engineering Science and Technology, an International Journal*, 24(1), 50–70. <https://doi.org/10.1016/j.jestch.2020.12.003>
13. Shakya, S. (n.d.). Heart Disease Prediction using Ensemble Model. <https://www.researchgate.net/publication/355615496>
14. Alrifai, M. F., Ahmed, Z. H., Hameed, A. S., & Mutar, M. L. (2021). Using Machine Learning Technologies to Classify and Predict Heart Disease. *International Journal of Advanced Computer Science and Applications*, 12(3), 123–127. <https://doi.org/10.14569/IJACSA.2021.0120315>
15. Yefet, E., Yossef, A., & Nachum, Z. (2021). Prediction of anaemia at delivery. *Scientific Reports*, 11(1). <https://doi.org/10.1038/s41598-021-85622-7>
16. Renugadevi, G., Asha Priya, G., Dhivyaa Sankari, B., & Gowthamani, R. (2021). Predicting heart disease using hybrid machine learning model. *Journal of Physics: Conference Series*, 1916(1). <https://doi.org/10.1088/1742-6596/1916/1/012208>
17. Dutta, S., Karkada, I. R., Sengupta, P., & Chinni, S. v. (2021). Anthropometric Markers With Specific Cut-Offs Can Predict Anaemia Occurrence Among Malaysian Young Adults. *Frontiers in Physiology*, 12. <https://doi.org/10.3389/fphys.2021.731416>
18. Rahman, M., Islam, M. J., Haque, S. E., Saw, Y. M., Haque, M. N., Duc, N. H. C., Al-Sobaihi, S., Saw, T. N., Mostofa, M. G., & Islam, M. R. (2017). Association between high-risk fertility behaviours and the likelihood of chronic undernutrition and anaemia among married Bangladeshi women of reproductive age. *Public Health Nutrition*, 20(2), 305–314. <https://doi.org/10.1017/S136898001600224X>
19. Teshale, A. B., Tesema, G. A., Worku, M. G., Yeshaw, Y., & Tessema, Z. T. (2020). Anaemia and its associated factors among women of reproductive age in eastern Africa: A multilevel mixed-effects generalized linear model. *PLoS ONE*, 15(9 September). <https://doi.org/10.1371/journal.pone.0238957>
20. Ferjani, M., & Ferjani, M. F. (n.d.). Disease Prediction Using Machine Learning. <https://doi.org/10.13140/RG.2.2.18279.47521>

21. Zhang, J., & Tang, W. (2020). Building a prediction model for iron deficiency anaemia among infants in Shanghai, China. *Food Science and Nutrition*, 8(1), 265–272. <https://doi.org/10.1002/fsn3.1301>
22. Mohammed, S. J. M., Ahmed, A. A., Ahmad, A. A., & Mohammed, M. S. (2020). Anaemia Prediction Based on Rule Classification. *Proceedings - International Conference on Developments in ESystems Engineering, DeSE, 2020-December*, 427–431. <https://doi.org/10.1109/DeSE51703.2020.9450234>
23. Woldu, B., Enawgaw, B., Asrie, F., Shiferaw, E., Getaneh, Z., & Melku, M. (2020). Prevalence and Associated Factors of Anaemia among Reproductive-Aged Women in Sayint Adjibar Town, Northeast Ethiopia: Community-Based Cross-Sectional Study. *Anaemia*, 2020. <https://doi.org/10.1155/2020/8683946>
24. Jamnok, J., Sanchaisuriya, K., Sanchaisuriya, P., Fucharoen, G., Fucharoen, S., & Ahmed, F. (2020). Factors associated with anaemia and iron deficiency among women of reproductive age in Northeast Thailand: A cross-sectional study. *BMC Public Health*, 20(1). <https://doi.org/10.1186/s12889-020-8248-1>
25. Dutta, S., Bandyopadhyay, S., & Kumar Samir, B. (2020). Diabetes Prediction Using Ensemble Classifier A review on bio medical image processing View project Medical Imaging View project Diabetes Prediction Using Ensemble Classifier. In *International Journal of Medical and Health Sciences Journal Home* (Vol. 9, Issue 2). <https://www.researchgate.net/publication/341775512>
26. Chauhan, R. H., Naik, D. N., Halpati, R. A., Patel, S. J., & Prajapati, M. A. D. (2008). Disease Prediction using Machine Learning. *International Research Journal of Engineering and Technology*. www.irjet.net
27. Kwon, J. myoung, Cho, Y., Jeon, K. H., Cho, S., Kim, K. H., Baik, S. D., Jeung, S., Park, J., & Oh, B. H. (2020). A deep learning algorithm to detect anaemia with ECGs: a retrospective, multicentre study. *The Lancet Digital Health*, 2(7), e358–e367. [https://doi.org/10.1016/S2589-7500\(20\)30108-4](https://doi.org/10.1016/S2589-7500(20)30108-4)
28. Javed, F. M., & Shamrat, M. (n.d.). An Analysis On Breast Disease Prediction Using Machine Learning Approaches Related papers. www.ijstr.org
29. Rajdhan, A., Agarwal, A., & Sai, M. (n.d.). Heart Disease Prediction using Machine Learning. In *IJERT Journal International Journal of Engineering Research & Technology*. www.ijert.org
30. Anand, P., & Sharma, A. (n.d.). Prediction of Anaemia among children using Machine Learning Algorithms. <https://www.researchgate.net/publication/341853966>

31. Gautam, S., Min, H., Kim, H., & Jeong, H. S. (2019). Determining factors for the prevalence of anaemia in women of reproductive age in Nepal: Evidence from recent national survey data. *PLoS ONE*, 14(6). <https://doi.org/10.1371/journal.pone.0218288>
32. Jaiswal, M., Srivastava, A., & Siddiqui, T. J. (2019). Machine learning algorithms for anaemia disease prediction. *Lecture Notes in Electrical Engineering*, 524, 463–469. https://doi.org/10.1007/978-981-13-2685-1_44
33. Noor, N. bin, Anwar, M. S., & Dey, M. (2019, December 1). An Efficient Technique of Haemoglobin Level Screening Using Machine Learning Algorithms. 2019 4th International Conference on Electrical Information and Communication Technology, EICT 2019. <https://doi.org/10.1109/EICT48899.2019.9068812>
34. Rahman, A. K. M. S., Javed, F. M., Shamrat, M., Tasnim, Z., Roy, J., Rahman, S., & Hossain, S. A. (2019). A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, 8(11). www.ijstr.org
35. Agarwal, R., & Sagar, P. (2019). A Comparative Study of Supervised Machine Learning Algorithms for Fruit Prediction. In *Journal of Web Development and Web Designing* (Vol. 4, Issue 1).
36. Dithy, M. D., & Krishnapriya, V. (2019). Anaemia selection in pregnant women by using random prediction (Rp) classification algorithm. *International Journal of Recent Technology and Engineering*, 8(2), 2623–2630. <https://doi.org/10.35940/ijrte.B3016.078219>
37. Sow, B., Mukhtar, H., Ahmad, H. F., & Suguri, H. (2020). Assessing the relative importance of social determinants of health in malaria and anaemia classification based on machine learning techniques. *Informatics for Health and Social Care*, 45(3), 229–241. <https://doi.org/10.1080/17538157.2019.1582056>
38. Young, M. F. (2018). Maternal anaemia and risk of mortality: a call for action. In *The Lancet Global Health* (Vol. 6, Issue 5, pp. e479–e480). Elsevier Ltd. [https://doi.org/10.1016/S2214-109X\(18\)30185-2](https://doi.org/10.1016/S2214-109X(18)30185-2)
39. Nguyen, P. H., Scott, S., Avula, R., Tran, L. M., & Menon, P. (2018). Trends and drivers of change in the prevalence of anaemia among 1 million women and children in India, 2006 to 2016. *BMJ Global Health*, 3(5). <https://doi.org/10.1136/bmjgh-2018-001010>

40. Bansal, D., Chhikara, R., Khanna, K., & Gupta, P. (2018). Comparative Analysis of Various Machine Learning Algorithms for Detecting Dementia. *Procedia Computer Science*, 132, 1497–1502. <https://doi.org/10.1016/j.procs.2018.05.102>
41. Moor, M. A., Fraga, M. A., Garfein, R. S., Rashidi, H. H., Alcaraz, J., Kritz-Silverstein, D., Elder, J. P., & Brodine, S. K. (2017). Individual and community factors contributing to anaemia among women in rural Baja California, Mexico. *PLoS ONE*, 12(11), 1–13. <https://doi.org/10.1371/journal.pone.0188590>
42. Lilare, R. R., & Sahoo, D. P. (2017). Prevalence of anaemia and its epidemiological correlates among women of reproductive age group in an urban slum of Mumbai. *International Journal Of Community Medicine And Public Health*, 4(8), 2841. <https://doi.org/10.18203/2394-6040.ijcmph20173333>
43. Harding, K. L., Aguayo, V. M., Namirembe, G., & Webb, P. (2018). Determinants of anaemia among women and children in Nepal and Pakistan: An analysis of recent national survey data. *Maternal and Child Nutrition*, 14. <https://doi.org/10.1111/mcn.12478>
44. Debelo, O., & Shiferaw, Y. A. (2016). 53) CORRELATES OF ANAEMIA STATUS AMONG WOMEN OF REPRODUCTIVE AGE IN ETHIOPIA (Vol. 7, Issue 2).
45. Le, C. H. H. (2016). The prevalence of anaemia and moderate-severe anaemia in the US population (NHANES 2003-2012). *PLoS ONE*, 11(11). <https://doi.org/10.1371/journal.pone.0166635>
46. Abdar, M., Niakan Kalhori, S. R., Sutikno, T., Much, I., Subroto, I., & Arji, G. (2015). Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases. *International Journal of Electrical and Computer Engineering (IJECE)*, 5(6), 1569–1576.
47. Green, R., & Dwyre, D. M. (2015). Evaluation of Macrocytic Anaemias. In *Seminars in Hematology* (Vol. 52, Issue 4, pp. 279–286). W.B. Saunders. <https://doi.org/10.1053/j.seminhematol.2015.06.001>
48. Patavegar, B. N., Kamble, M. S., & Langare-Patil, S. (2014). Online) An Open Access. In *International Journal of Basic and Applied Medical Sciences* (Vol. 4, Issue 2). <http://www.cibtech.org/jms.htm>
49. Verma, R., Kharb, M., Deswal, S., Arora, V., & Kamboj, R. (2014). Prevalence of anaemia among women of reproductive age group in a rural block of Northern India Corresponding Author Citation. In *Suppl* (Vol. 26).

50. Laha, F. (n.d.). Anaemia “a silent killer” among women in India: Present scenario. www.scholarsresearchlibrary.com
51. Alquaiz, J. M., Abdulghani, H. M., Khawaja, R. A., & Shaffi-Ahamed, S. (2012). Accuracy of Various Iron Parameters in the Prediction of Iron Deficiency Anaemia among Healthy Women of Child Bearing Age, Saudi Arabia. In Iranian Red Crescent Medical Journal Iran Red Crescent Med J (Vol. 14, Issue 7).
52. Mishra, P., & Ahluwalia, S. K. (2012). The Prevalence of Anaemia among Reproductive Age Group (15-45 Yrs) Women in A PHC of Rural Field Practice Area of MM Medical College, Ambala, India. *Journal of Women’s Health Care*, 01(03). <https://doi.org/10.4172/2167-0420.1000113>
53. Rohner, F., Tschannen, A. B., Northrop-Clewes, C., Kouassi-Gohou, V., Bosso, P. E., & Mascie-Taylor, C. G. N. (2012). Comparison of a possession score and a poverty index in predicting anaemia and undernutrition in pre-school children and women of reproductive age in rural and urban Côte d’Ivoire. *Public Health Nutrition*, 15(9), 1620–1629. <https://doi.org/10.1017/S1368980012002819>
54. Sharieff, W., Dofonsou, J., & Zlotkin, S. (2008). Is cooking food in iron pots an appropriate solution for the control of anaemia in developing countries? A randomised clinical trial in Benin. *Public Health Nutrition*, 11(9), 971–977. <https://doi.org/10.1017/S1368980007001139>
55. Pala, K. (2008). Prevalence & risk factors of anaemia among women of reproductive age in Bursa, Turkey. In Article in *The Indian Journal of Medical Research*. <https://www.researchgate.net/publication/23567283>
56. Sanders, J. (2016). Defining terms: Data, information and knowledge. *Proceedings of 2016 SAI Computing Conference, SAI 2016*, 223–228. <https://doi.org/10.1109/SAI.2016.7555986>
57. Girija PL. Anaemia among women and children of India. *Anc Sci Life*. 2008 Jul;28(1):33-6. PMID: 22557295; PMCID: PMC3336343.
58. <https://pib.gov.in/PressReleasePage.aspx?PRID=1795421>
59. Mog, M., & Ghosh, K. (2021). Prevalence of anaemia among women of reproductive age (15–49): A spatial-temporal comprehensive study of Maharashtra districts. *Clinical Epidemiology and Global Health*, 11. <https://doi.org/10.1016/j.cegh.2021.100712>
60. Elizabeth Heger Boyle, Miriam King and Matthew Sobek. IPUMS-Demographic and Health Surveys: Version 9 [dataset]. IPUMS and ICF, 2022. <https://doi.org/10.18128/D080.V9>

61. DHS report Anaemia. (n.d.).
62. Cornell Statistical Consulting Unit Ordinal Logistic Regression models and Statistical Software: What You Need to Know Statnews #91. (n.d.).
63. Mog, M., & Ghosh, K. (2021). Prevalence of anaemia among women of reproductive age (15–49): A spatial-temporal comprehensive study of Maharashtra districts. *Clinical Epidemiology and Global Health*, 11. <https://doi.org/10.1016/j.cegh.2021.100712>
64. Meh C, Sharma A, Ram U, Fadel S, Correa N, Snelgrove JW, Shah P, Begum R, Shah M, Hana T, Fu SH, Raveendran L, Mishra B, Jha P. Trends in maternal mortality in India over two decades in nationally representative surveys. *BJOG*. 2022 Mar;129(4):550-561. doi: 10.1111/1471-0528.16888. Epub 2021 Sep 15. PMID: 34455679; PMCID: PMC9292773.
65. Anand A. Anaemia -- a major cause of maternal death. *Indian Med Trib*. 1995 Jan 15;3(1):5, 8. PMID: 12179189.
66. Totade, M., Gaidhane, A., & Sahu, P. (2023). Interventions in Maternal Anaemia to Reduce Maternal Mortality Rate Across India. *Cureus*. <https://doi.org/10.7759/cureus.46617>.
67. Sarin AR. Indian women's reproductive health -- challenges and remedies. *Indian J Matern Child Health*. 1991;2(1):1-2. PMID: 12288701.