"STATISTICAL STUDY OF EXPLORING ANEMIA PREVALENCE AND RISK FACTORS AMONG 6-59 MONTHS CHILDREN IN INDIA"

Anekant Education Society's

Tuljaram Chaturchand College of Arts, Science, and Commerce, Baramati (Autonomous)

413102

A Project Report On

"Statistical Study of Exploring Anemia Prevalence and Risk Factors Among 6-59 Months Children in India"

Submitted To

Department of Statistics

Tuljaram Chaturchand College of Arts, Science, and Commerce, Baramati (Autonomous)

By

Mr. Dham Tejas Manik. 16678

Mr. Pawar Pranav Vilas. 16701

Mr. Salgude Rohan Changdeo. 16705

Under the Guidance of

Mrs. Pooja S. Gaikwad.

(2023-24)

DEPARTMENT OF STATISTICS

CERTIFICATE

This is to certify that partial fulfillment of curriculum MSc-II Students **Mr. Dham Tejas Manik, Mr. Pawar Pranav Vilas, Mr. Salgude Rohan Changdeo** have successfully completed the project work in the statistics entitled **"Statistical Study of Exploring Anemia Prevalence and Risk Factors Among 6-59 Months Children in India"** prescribed by Tuljaram Chaturchand Collage of Arts, Science and Commerce Baramati during academic year 2023-2024.

This is a record of Bonafede work carried out by them under my supervision and guidance.

.

Mrs. Pooja. S. Gaikwad.
Project Guide

Examiner

Prof. Dr. V.C. Kakade.
Head
Department of Statistics

Place: Baramati

Date: / / 2024

ACKNOWLEDGEMENT

We have satisfaction upon completion of the project entitled **“Statistical Study of Exploring Anemia Prevalence and Risk Factors Among 6-59 Months Children in India”** at the Department of Statistics in Tuljaram Chaturchand College of Arts, Science, and Commerce, Baramati during the academic year of 2023-24.

We take this opportunity to express our sense of gratitude and due respect to our Principal Dr. Jagtap. A.S. Sir for their continuous encouragement, positive support, and inspiration.

Also, thanks to Prof. Dr. Vikas. C. Kakade, Head of Department of Statistics for providing us all necessary facilities, good co-operation and their valuable discussion during completion of our project. The timely and successful completion of this project could hardly be possible without the guidance and assistance of our guide Mrs. Pooja S. Gaikwad.

We sincerely acknowledge the excellent advice, criticisms, encouragement and confidence of our guide. We appreciate very much their precious guidance and support in the preparation of this project. Thank you for inspiring us to learn and explore the statistics application in our project.

And we thankful to our PG Coordinator Mr. Swami. C. P. Sir and all faculties at the department for their Valuable guidance and great Co-operation.

We also thank to classmates and friends at the department for their support and help during the project, Last but Not the Least, we would like to thank non-teaching staff of department of statistics for providing laboratory at any time during the project.

INDEX

ABSTRACT

Children are considered as an important asset for any country and their health as one of the important indicators of a healthy country. Anemia is one of the most prevalent disease. There are several factors affecting children's health and anemia is one of them. In these groups we are study on the anemia in children's.

The National Family Health Survey (NFHS) and the Demographic and Health Surveys (DHS) are comprehensive, large-scale surveys conducted in India and other countries. NFHS provides crucial information on fertility, maternal and child health, family planning, nutrition, and more. It offers ready-to-use data for researchers and policymakers. Similarly, DHS surveys cover a wide range of health and demographic indicators globally. Both programs contribute valuable insights to public health and inform evidence-based policies and interventions.

We access our dataset on government site. The Demographic and Health Surveys (DHS). Data has been taken from the latest fifth round of National Family Health survey (2019–21). A total of 3,11,182 children aged 6 to 59 months were included in the study.

The aim of our project is to examine the effect of anemia on children. We Used the machine learning model for fulfill the motive of our Study. and understood from these the logistic regression model and decision tree model both model gave significant role. and also we use chi-square test to determine the association between anemia and its associated covariates. and understood from these the logistic regression model and decision tree model both model gave significant role. from that model we get Child age in month, wealth quantile, age of mother, mothers education level these five variables are significant form these supervised learning technique.

For the project on anemia prevalence among 6-59 months' children in India, it is suggested to: Enhance nutritional programs focusing on iron-rich diets and supplements. Implement regular anemia screening and treatment, especially in high-risk zones and states. Educate communities on nutrition and healthcare to prevent anemia from early childhood.

Keywords: Anemia, Children, Health Survey, Hypothesis Testing, Logistic Regression, Decision Tree.

MOTIVATION

Anemia is highly prevalent in India. and it is important problem among 6 to 59 months old children in India. India has implemented several programs to tackle anemia. Prevalence of anemia is very important in India. Anemia can affect the children health. Anemia is the most common nutritional deficiency disorder in the India. And in most of the people are not aware about Anemia Prevalence, that's why we are decide to work on this topic. From Our project, find the prevalence of anemia in children in age 6 to 59 months.

We live in India, in our country most of the people living in rural areas, but they are not having much more knowledge about these types of diseases. Identifying risk factors allows for targeted interventions. we can tell them the which factors are important for prevalence of anemia in children.

So our project purpose is that, to give information of how to awere about prevalence of anemia in children.these study suggest the parents education is the most important factor for the better health of children.if the parent is well educated then they have the right knowledge of how to take care of there children and there proper diet.

The main motivation to do project on prevalence of anemia among 6 to 59 month old children in india is after the time of topic serching when we read the letest report of national family health survey.

This study aimed to assess the severity level of anemia and investigate the socioeconomic and demographic elements connected with childhood anemia. This study also focused on revealing a better-fitted model for the data. Moreover, this research can help public health policymakers to determine priorities for invention for reducing the anemia severity level.

INTRODUCTION

Children are considered as an important asset for any country and their health as one of the important indicators of a healthy country. There are several factors affecting children's health and anemia is one of them. Anemia is a critical health concern that affects millions of children worldwide, with India being one of the most impacted countries.

Anemia, an indicator of poor nutrition and health, is a major public health problem in India. The causes of anemia are multifactorial, which includes iron deficiency, nutritional deficiencies, chronic infections, inherited blood disorders, obesity and non-communicable diseases. Anemia has been implicated with growth retardation, impaired motor and cognitive development, and childhood morbidity and mortality. Anemia among children is a public health problem globally.

Among the developing countries, India was the largest contributor to child anemia in the last decades. It has many harmful effects on children's lifestyle as it reduces learning capacity, attentiveness and intelligence. The study attempts to identify the spatial prevalence and detect the clustering of anemia in India based on National Family Heath Survey-V, 2019–21 and also tries to identify the determinants of anemia among children (6–59 months).

WHO estimated that the prevalence of anemia worldwide in 2019 was 39.8% among children aged 6 to 59 months were anemic in 2019. According to WHO's dashboard, In the case of children (6 to 59 months), the prevalence in India was 53.4%. According to the survey report, at least 67 percent children (6-59 months) have anemia as compared to 58.6 per cent in the last survey conducted in 2015-16.

There are many types of anemia. Which are as follows

1. **Iron deficiency anemia:** This is not enough iron in the blood. Iron is needed to form hemoglobin. This is the most common cause of anemia.

2. **Megaloblastic anemia:** This is when red blood cells are too large from a lack of folic acid or vitamin B-12. One type of megaloblastic anemia is pernicious anemia. In this type, there is a problem absorbing vitamin B-12, important to making red blood cells.
3. **Hemolytic anemia:** This is when red blood cells are destroyed. There are many different causes, such as serious infections or certain medicines.
4. **Sickle cell anemia:** This is a type of hemoglobinopathy, an inherited type of anemia with abnormally-shaped red blood cells.
5. **Cooley's anemia (thalassemia):** This is another inherited type of anemia with abnormal red blood cells.
6. **Aplastic anemia:** This is failure of the bone marrow to make blood cells.

The Indian government has implemented several schemes to address anemia. Let's explore some of these initiatives:

1. **Anemia Mukt Bharat (AMB):**
   - Launched as part of the POSHAN Abhiyan (National Nutrition Mission).
   - Aims to reduce anemia by three percentage points per year among adults, children, and adolescents from 2018 to 2022.
   - Focuses on prophylactic iron and folic acid supplementation, deworming, behavior change communication, and more.
2. **Weekly Iron and Folic Acid Supplementation (WIFS):**
   - A program targeting adolescent girls and boys (10–19 years).
   - Provides weekly iron and folic acid supplements to improve iron status.
3. **National Iron Plus Initiative (NIPI):**
   - Part of the Intensified National Iron Plus Initiative Program.
   - Aims to accelerate the annual rate of decline of anemia.
4. **Operationalization of Blood Bank:**
   - Ensures timely availability of blood for transfusion, which is crucial for managing severe anemia.

5. **Pradhan Mantri Surakshit Matritva Abhiyan (PMSMA):**
   - Focuses on pregnant and lactating women, including measures to prevent and manage anemia during pregnancy.

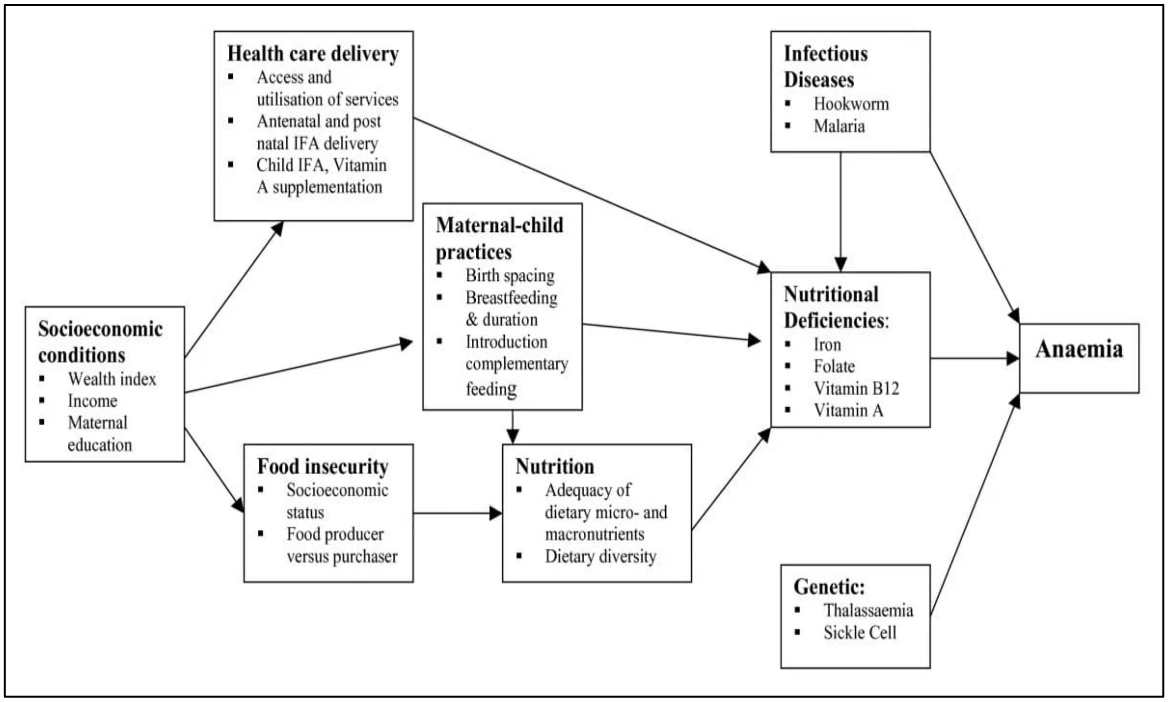
These initiatives aim to combat anemia through a combination of supplementation, awareness campaigns, and improved nutrition. However, continuous monitoring and evaluation are essential to ensure their effectiveness.

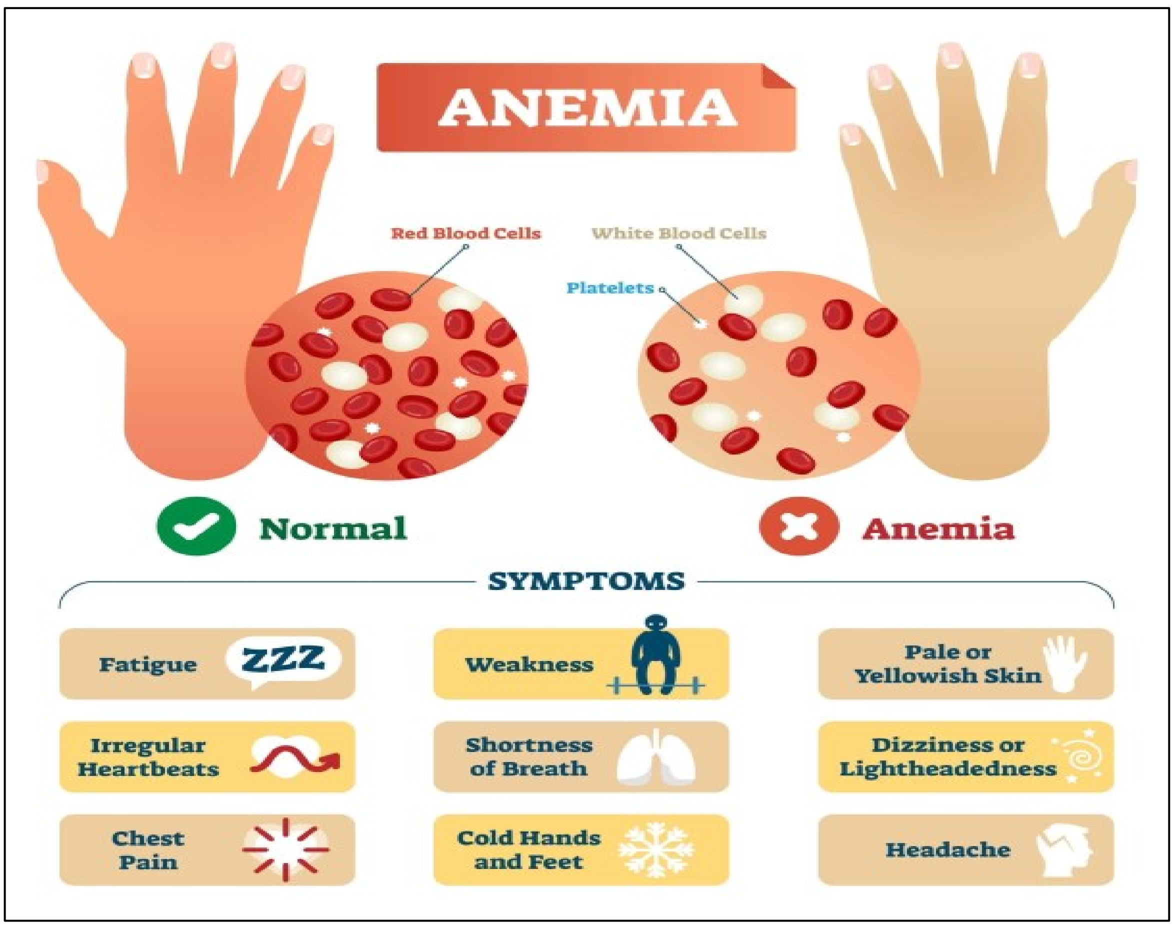
To reduce anemia in India, it is imperative to find out the influencing factors that upsurge the prevalence of anemia. Several studies have shown a comprehensible association between anemia and a list of determinants such as place of residence, states in India, educational status of mother, age of the child, birth order. This statistical study aims to explore the prevalence of anemia and its associated risk factors within this vulnerable age group in India.

This statistical project delves into the prevalence of anemia among children aged 6-59 months, a vulnerable period crucial for growth and development. By analyzing data from national surveys, we aim to uncover patterns and factors contributing to anemia across different regions of India. The insights gained will not only shed light on the current state of child health but also inform strategies to alleviate this condition that hinders the potential of the future generation.

This study will utilize statistical methods to analyze data from various sources, aiming to provide a comprehensive overview of the current state of anemia among children in India. By identifying the key risk factors, the study seeks to inform public health strategies and interventions to reduce the burden of anemia and improve the health outcomes of children in India.

The outcomes of this study will provide valuable insights for developing targeted interventions to reduce anemia and enhance child health.


Health care delivery
Access and utilisation of services
Antenatal and post natal IFA delivery
Child IFA, Vitamin A supplementation
Infectious Diseases
Hookworm
Malaria
Maternal-child practices
Birth spacing
Breastfeeding & duration
Introduction complementary feeding
Socioeconomic conditions
Wealth index
Income
Maternal education
Nutritional Deficiencies:
Iron
Folate
Vitamin B12
Vitamin A
Anaemia
Food insecurity
Socioeconomic status
Food producer versus purchaser
Nutrition
Adequacy of dietary micro- and macronutrients
Dietary diversity
Genetic:
Thalassaemia
Sickle Cell


ANEMIA
Red Blood Cells
White Blood Cells
Platelets
Normal
Anemia
SYMPTOMS
Fatigue
ZZZ
Weakness
Pale or Yellowish Skin
Irregular Heartbeats
Shortness of Breath
Dizziness or Lightheadedness
Chest Pain
Cold Hands and Feet
Headache

OBJECTIVES

1. To find the prevalence of anemia among children in age 6 to 59 months.
2. To develop a machine learning model to predict the child anemia.
3. To determine the factors associated with child anemia.
4. To find rate of anemia according to child age group.
5. To find anemia prevalence according to area.
6. To investigate the Impact of mother's education on children.

METHODOLOGY

Data Collection

Data collection is a systematic process of collecting information about variables. It helps to find answers to questions, hypothesizes to much and evaluate results. Before collection of data decide our research aim. Identify our research objective.

For collection of data we refer National Family Health Survey (NFHS 5) 2019 -21 report in India. were studied thoroughly and get more information about our research topic which is prevalence of anemia in age 6 to 59 months' children in India. These data contain 3,11,182 samples and 55 variables.

Data Cleaning

Data cleaning is the way to improvise the data or remove incorrect, corrupted or irrelevant data. As in our dataset, there are some columns that are not important and irrelevant for the model training. So, we can drop that column before training. There are 2 approaches to dealing with empty/null values

- We can delete the column/row (if the feature or record is not much important).

  Filter the data for delete unwanted variables or observations.

- Filling the empty slots with mean/mode/0/NA/etc. (depending on the dataset requirement).

In our dataset we are used various types of techniques to clean our data for more accurate results. In beginning there are 3,11,182 samples and 55 variables are in the dataset but in that there is much more missing values and unwanted variables are present. we used some statistical techniques and delete that unwanted variables and missing data values. after the data cleaning we can get the 9995 samples and 16 variables for our research study.

Data Description:

The Attributes are defining as follows:

- ANEMIASTATUS - Anemia Status of Children.

  1 - Anemic   0 - Not Anemic

- URBAN - Urban-Rural status

  1 - Urban   2 - Rural

- RESIDENT - Usual resident or Visitor

  1 - Usual Resident   2 - Visitor

- ETHNICITYIA - Ethnicity, India

  0 - Scheduled Caste   1 – Scheduled Tribe   2 – Other Backward Caste

- RELIGION – Religion

  0 – Muslim   1 – Hindu   2 – Other

- MOTHERAGE – Age of Mother When Children Born
- WELTHQ – Household Wealth

  1 – Poorest   2 – Middle   3 – Richest

- KIDSEX – Sex in Children

  1 – Male   2- Female

- KIDAGEMO – Age of Children in Months (6 – 59 months)
- ETHNICITYIA2 – Cast or Tribe, India

  1- Cast   2–Tribe

- MOTHERCURRWORK – Mother is Currently Working or Not

  0- Not Working   1-Working

- HEALTHIDX – Number of Children in health history
- MATERNITYIDX - Number of Children in Maternity history
- EDUCLVL – Highest Educational

  0 - No education   1- Primary   2- Secondary   3 -Higher

- KIDBORD – Child's Birth Order Number

- DIARRECENT – Child had Diarrhea

  1-No 2-Yes

Data Pre- Processing:

"Prevalence of anemia in children" is a dataset having sample size more than 3 lakhs with 15 variables. These variables, which served as a features of the dataset, representing the children is anemic or not. the next step was to investigating missing data. Variables with more than 50 % missing data would be removed from the dataset. Any observation which had missing values were also removed from the dataset.

Data Visualization:

Data Visualization is the pictorial or graphical representation of information. It enables to grasp difficult concepts or identify new patterns. This includes creating and investigating visual representations of information.

Data visualization helps us quickly identify trends and patterns that might not be apparent from raw data tables. It is an effective method from drawing conclusion from data effectively presenting information.

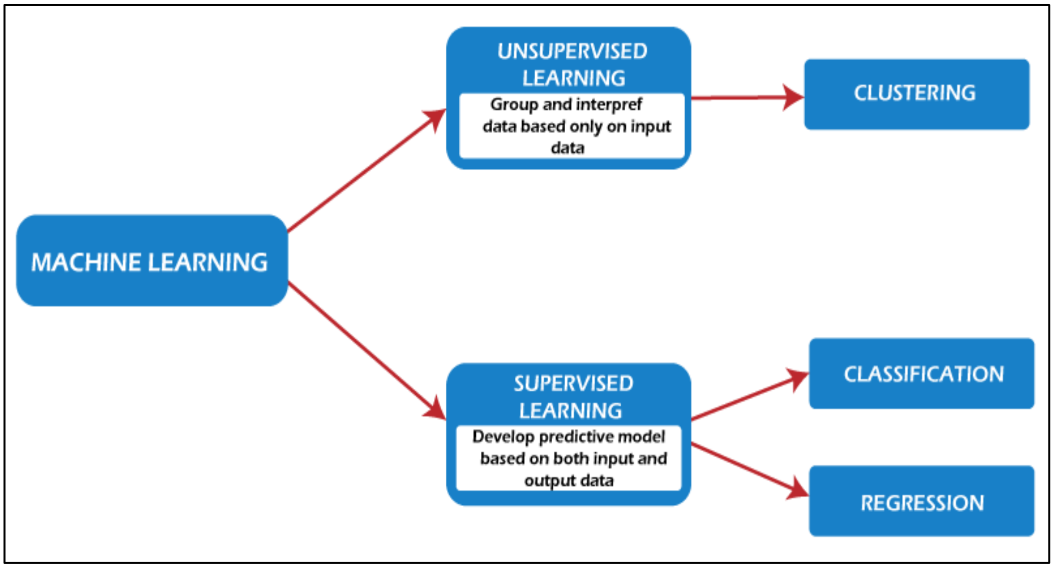
Machine learning techniques

Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to directly "learn" from data without relying on a predetermined equation as a model.

As the number of samples available for learning increases, the algorithm adapts to improve performance. Deep learning is a special form of machine learning.

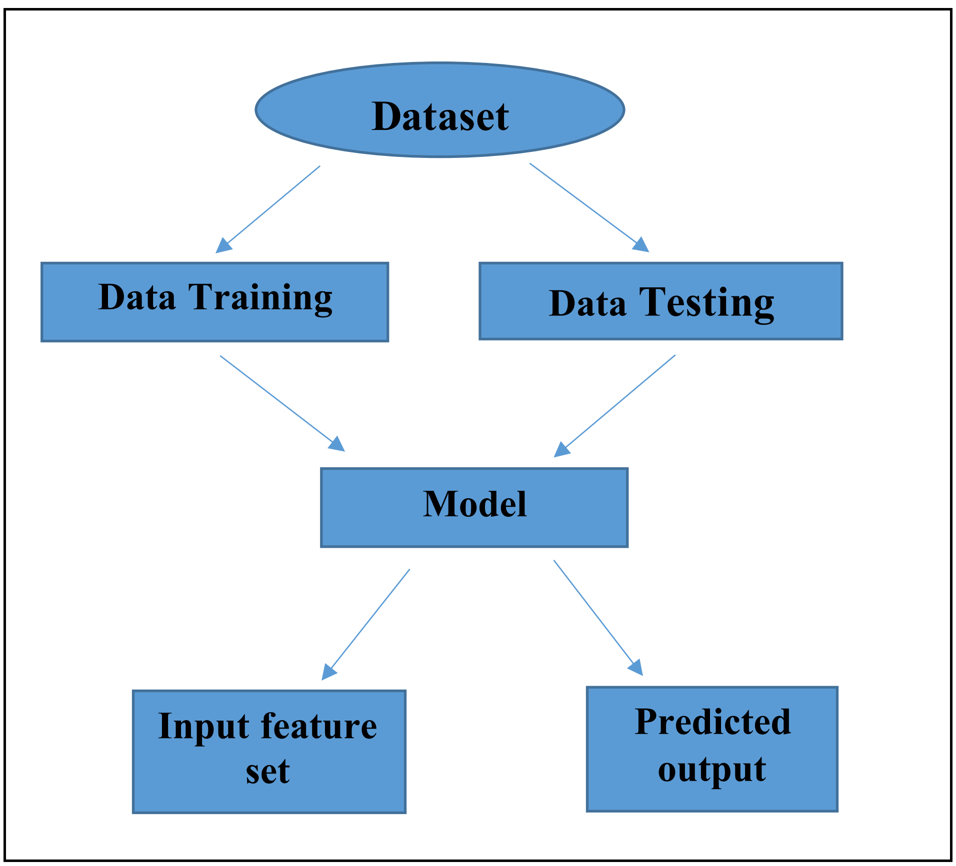
How does machine learning work?

Machine learning uses two techniques: supervised learning, which trains a model on known input and output data to predict future outputs, and unsupervised learning, which uses hidden patterns or internal structures in the input data.



Splitting Dataset into Training and Testing

X and Y splitting (i. e. Y is the anemia status and rest of other columns are X) Here, we use 70-30 rule. Here, 70% of data is training dataset and remaining 30% of the data for testing.



Supervised Learning Algorithm

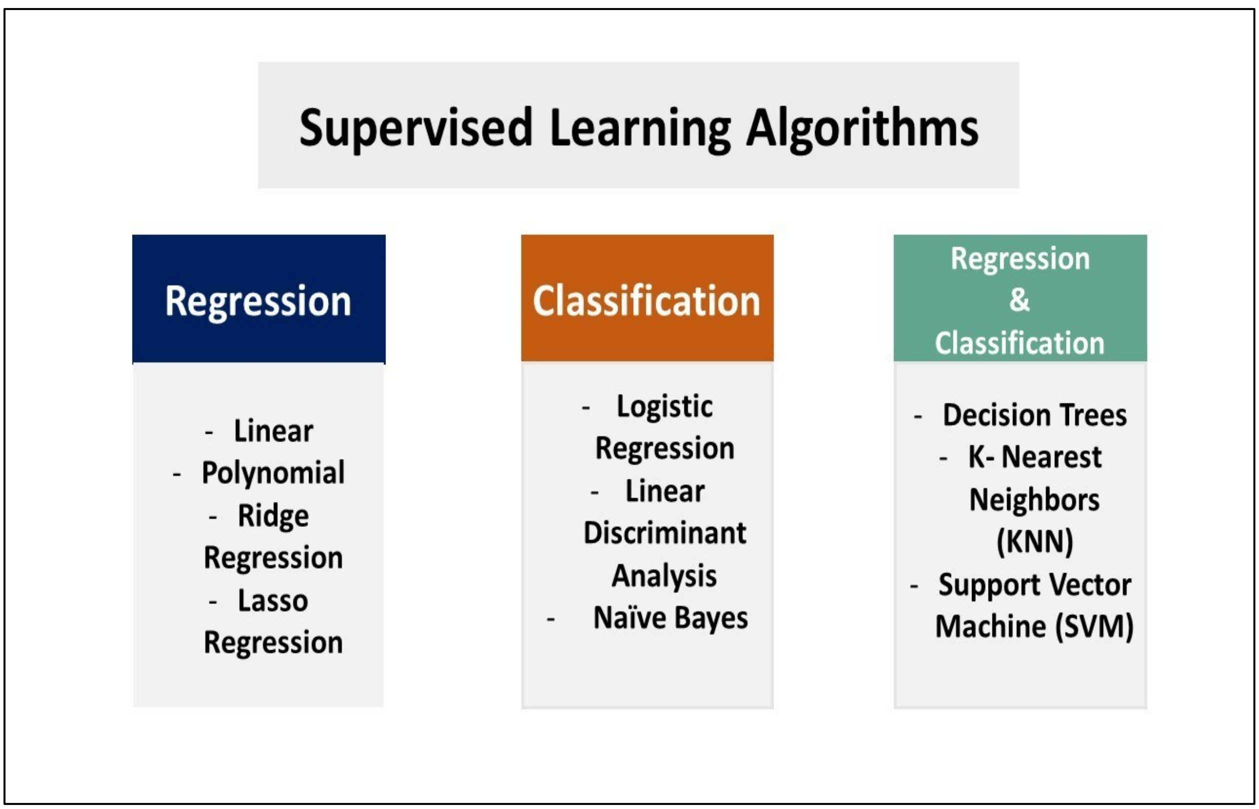
Supervised learning is a type of Machine learning in which the machine needs external supervision to learn. The supervised learning models are trained using the labeled dataset. Once the training and processing are done, the model is tested by providing a sample test data to check whether it predicts the correct output.

The goal of supervised learning is to map input data with the output data. Supervised learning is based on supervision, and it is the same as when a student learns things in the teacher's supervision. The example of supervised learning is spam filtering.

Supervised learning can be divided further into two categories of problem:

- Classification
- Regression

Examples of some popular supervised learning algorithms are Simple Linear regression, Logistic regression, Decision Tree, Logistic Regression, KNN algorithm, etc.



Key Points:

- Supervised learning involves training a machine from labeled data.
- Labeled data consists of examples with the correct answer or classification.
- The machine learns the relationship between inputs (fruit images) and outputs (fruit labels).
- The trained machine can then make predictions on new, unlabeled data.

Here in these study we use logistic regression and decision tree algorithm for the prediction of anemia in child.

Logistic Regression Algorithm:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

Logistic Regression Equation:

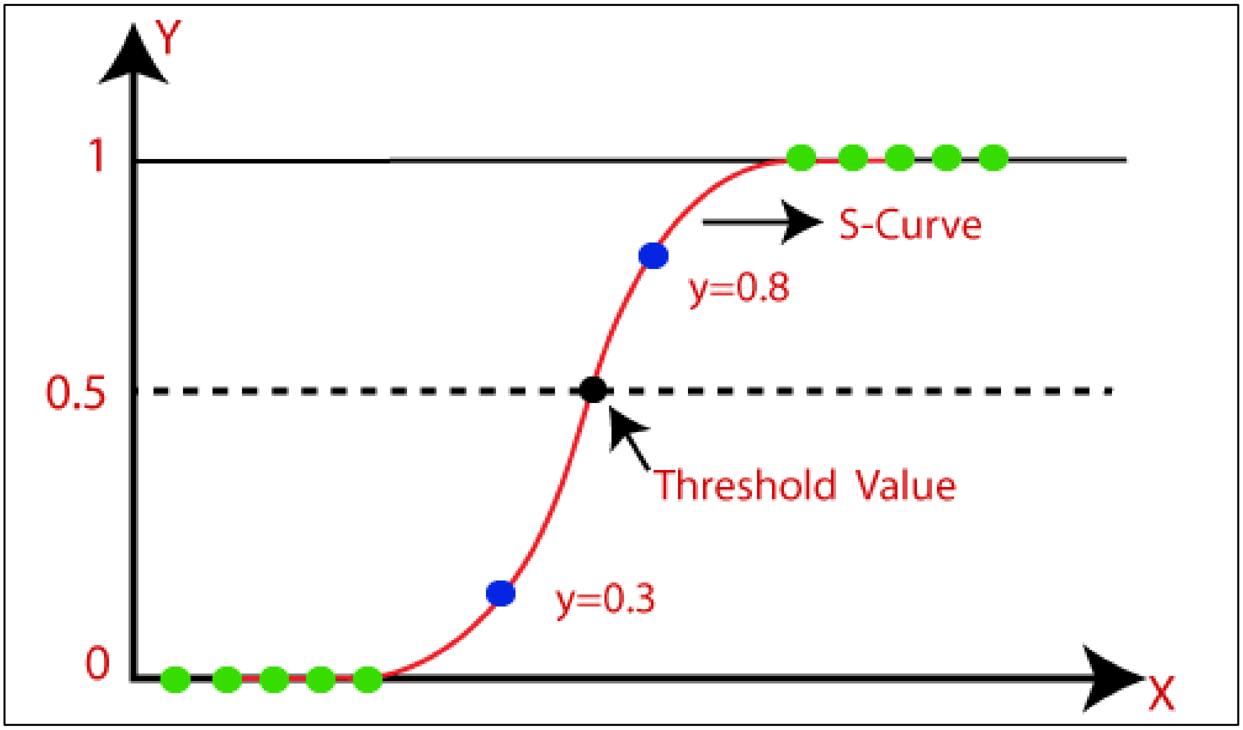
The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

$$y= \frac{\exp\,(\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta iXi)}{1+\exp\,(\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta iXi)}$$

Here,

- X = input value
- Y = predicted output
- $\beta 0$ = bias or intercept term
- βi = coefficients for input (Xi)

This equation is similar to linear regression, where the input values are combined linearly to predict an output values using weights or coefficient values. However, unlike linear regression, the output values modeled here is a binary (0 or 1) rather than a numeric value.



Decision Tree Algorithm

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).

In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Types of Decision Tree:

- **Classification Tree** (Yes/ No type)
- **Regression Tree** (Continuous data types)

Important Terminology related to Decision Tree:

Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.

Splitting: It is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.

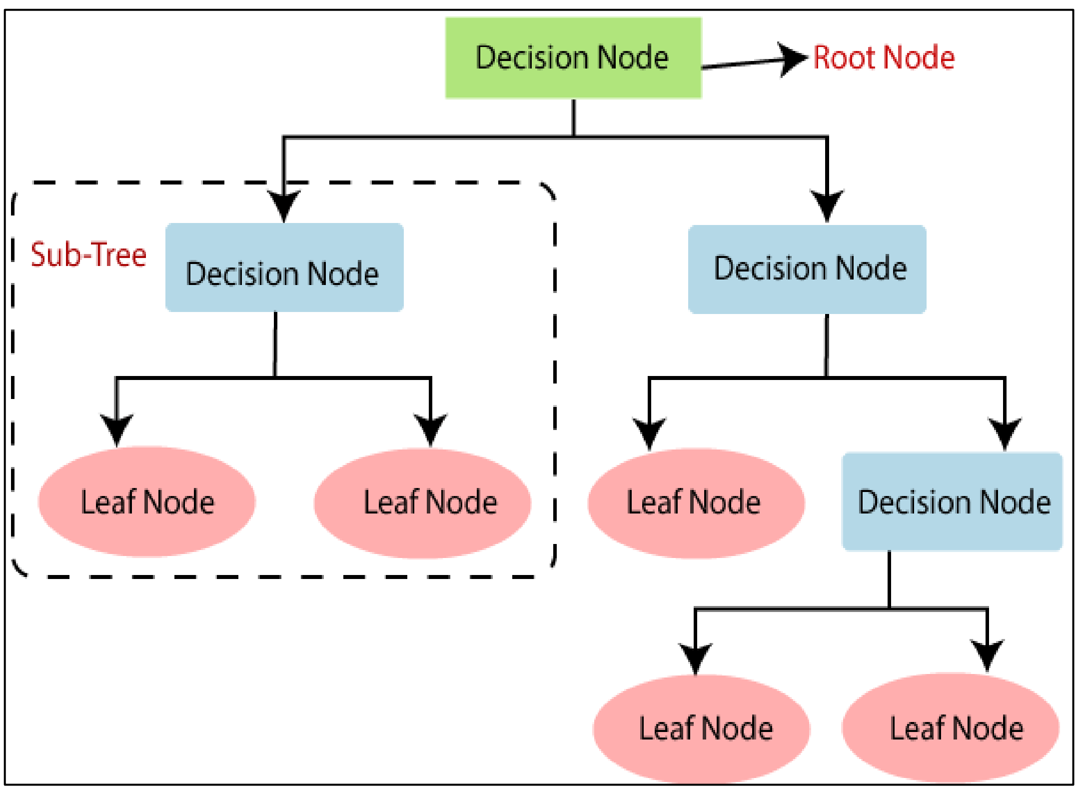
Leaf / Terminal Node: Nodes do not split is called Leaf or Terminal node.

Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.

Parent and Child Node: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

Below diagram explains the general structure of a decision tree:



Model Evaluation

Confusion matrix: A confusion matrix is an N X N matrix, where N is the number of predicted classes. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix. It is a performance measurement for machine learning classification problems where the output can be two or more classes. Confusion matrix is a table with 4 different combinations of predicted and actual values. It is extremely useful for measuring precision-recall, Specificity, Accuracy, and most importantly, AUC-ROC curves.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

- **True Positive:** You predicted positive, and it's true.
- **True Negative:** You predicted negative, and it's true.
- **False Positive:** (Type 1 Error): You predicted positive, and it's false.
- **False Negative:** (Type 2 Error): You predicted negative, and it's false.

Here are a few definitions you need to remember for a confusion matrix:

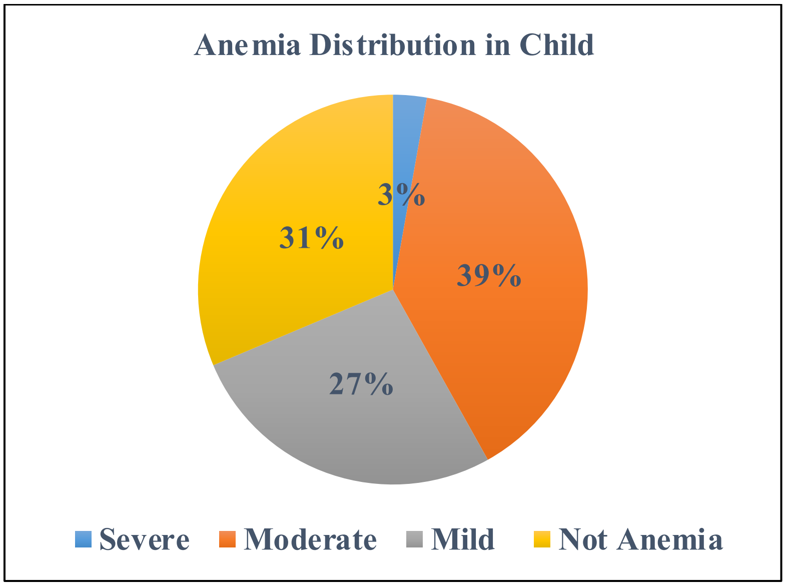
- **Accuracy:** Accuracy is the percentage of predictions that the model makes correctly. It is calculated by dividing the number of correct predictions by the total number of predictions.
- **Precision:** Precision is the percentage of positive predictions that the model makes that are actually correct. It is calculated by dividing the number of true positives by the total number of positive predictions.
- **Sensitivity**: Sensitivity measures the model's ability to correctly identify instances of the positive class out of all actual positive instances.
- **Specificity**: Specificity measures the model's ability to correctly identify instances of the negative class out of all actual negative instances.

- **Recall:** Recall is the percentage of all positive examples that the model correctly identifies. It is calculated by dividing the number of true positives by the total number of positive examples.
- **F1 score:** The F1 score is a weighted average of precision and recall. It is calculated by taking the harmonic mean of precision and recall.
- **ROC Curve:** An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters:
    - True Positive Rate
    - False Positive Rate

Exploratory Data Analysis

Anemia Prevalence in children

Severe	Moderate	Mild	Not Anemia	Total
280	3908	2674	3133	9995

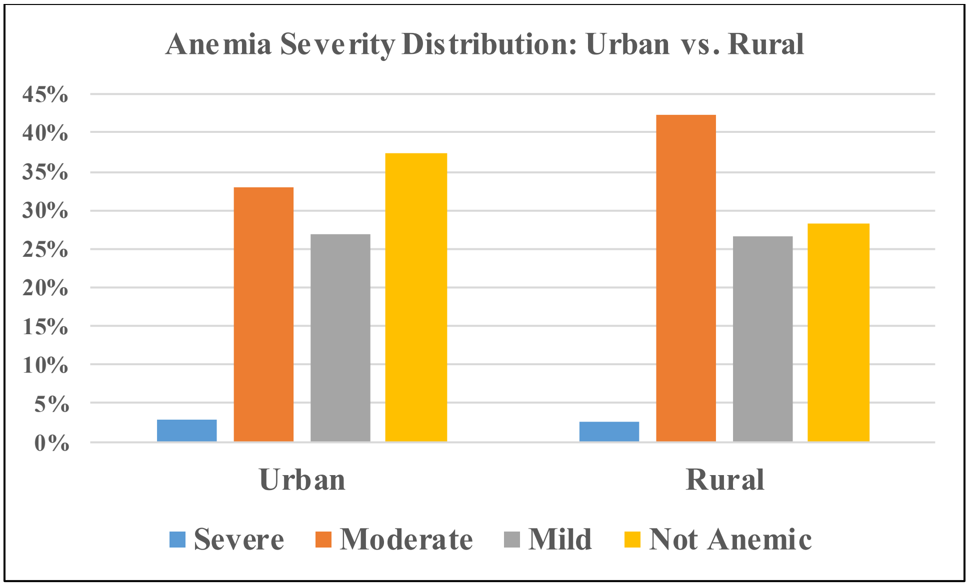


Interpretation:

From the above pie chart, it was discovered that there were only 3% of Children who have Severe Anemia which is good figure but it was sensitive having 39% of Children who have Moderate Anemia, there is 27% of Children who have Mild Anemia, there is 31% of Children who don't have Anemia. But if we consider two categories that is anemia or no anemia we can say that the approximately 69% children found to be anemic and remaining 31% children not suffering from anemia.

Anemia Severity Distribution: Urban vs. Rural

	Severe	Moderate	Mild	Not Anemic
Urban	3%	33%	27%	37%
Rural	3%	42%	27%	28%

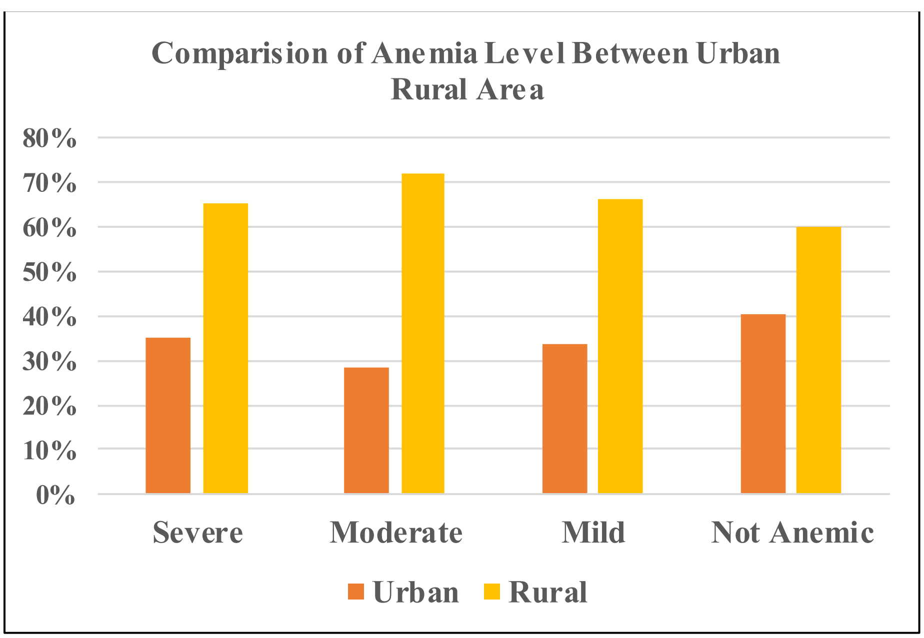


Interpretation:

In urban area there is maximum number of children is not anemic but we see in the rural area there is maximum number of children having moderate anemia. There is a severe condition for rural area because it has maximum number of moderate anemia patients.

Area Wise Comparison Of Anemia Level

	Severe	Moderate	Mild	Not Anemic
Urban	35%	28%	34%	40%
Rural	65%	72%	66%	60%



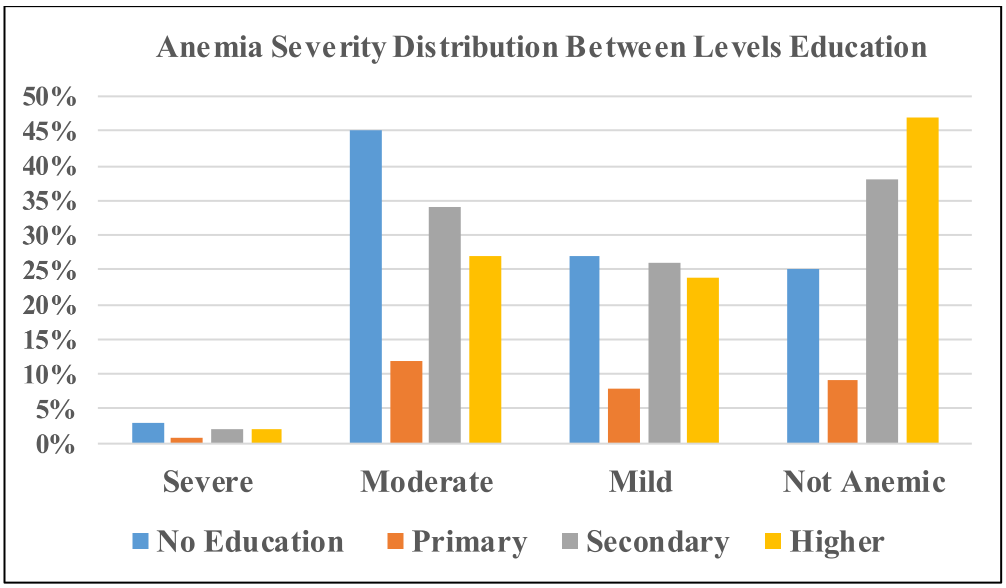
Interpretation:

From these figure here we can have observed that in urban area throughout in all categories Severe, moderate, mild and also in not anemia percentage of children suffering from anemia was higher in rural areas then that of urban area.

In this figure discovers the anemia trend or pattern in the children according to Urban Rural area.

Anemia Severity Distribution Between Levels of Education

	No Education	Primary	Secondary	Higher
Severe	3%	1%	2%	2%
Moderate	45%	12%	34%	27%
Mild	27%	8%	26%	24%
Not Anemic	25%	9%	38%	47%

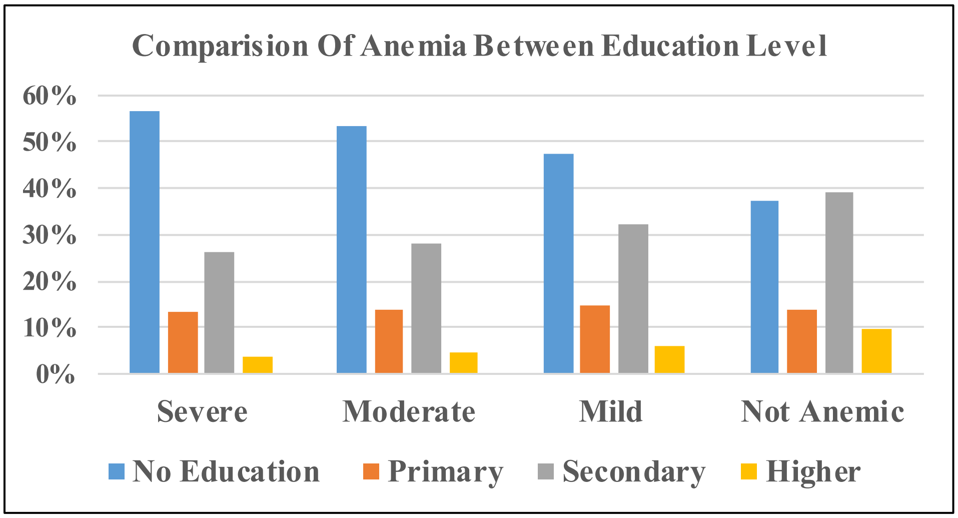


Interpretation:

From the above figure we see that that individuals with higher education appear to have lower rates of severe anemia. Ultimately, we can say that this pattern indicates that education may play a role in health awareness and access to resources for prevention and treatment.

Comparison of Anemia Between Education Level

	No Education	Primary	Secondary	Higher
Severe	57%	13%	26%	4%
Moderate	53%	14%	28%	5%
Mild	47%	15%	32%	6%
Not Anemic	37%	14%	39%	10%

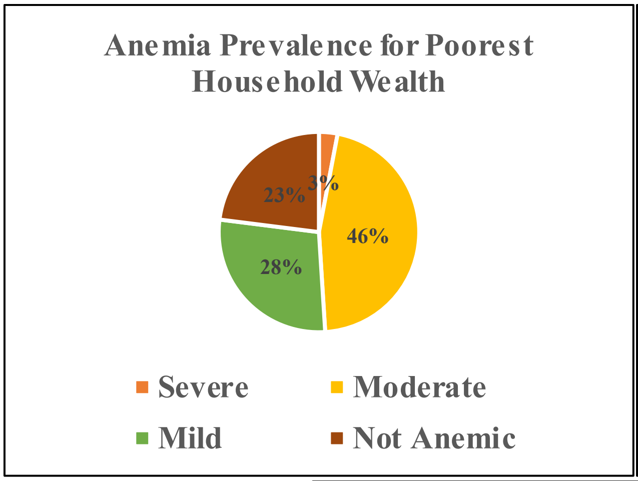


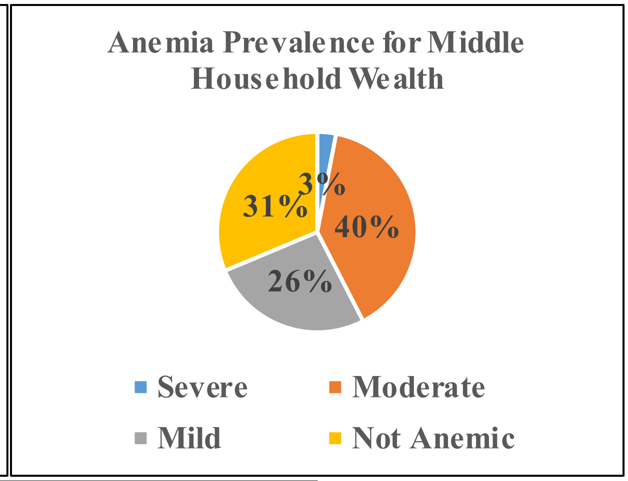
Interpretation:

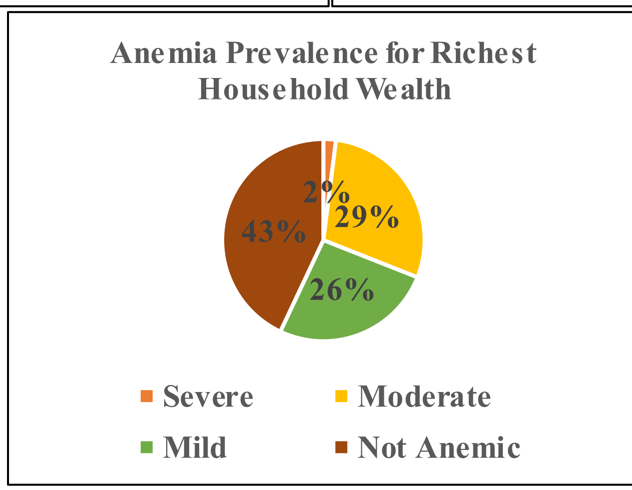
As seen in the above figure, the percentage of Individuals with no education was higher than that of other education level which is primary, secondary and higher education throughout for all anemia levels. Those with higher education are more likely to be not anemic. Although mother is educated or not is important for children anemia status. Mother's education that might impact on a children's health. Therefore, anemia prevalence according to Mother's education status was examine in this plot.

Anemia Severity Distribution: In Wealth Quantiles

	Severe	Moderate	Mild	Not Anemic
Poorest	3%	46%	28%	23%
Middle	3%	40%	26%	31%
Richest	2%	29%	26%	43%







Interpretation:

From figure 1, we can say that there is high portion (46%) of the poorest population suffers from moderate anemia. It highlights that a significant portion (28%) of the poorest population suffers from mild anemia. And very small number of people in these group have severe anemia.

But comparing anemic and not anemic groups 77% population is anemic and remaining 23% not suffer from anemia.

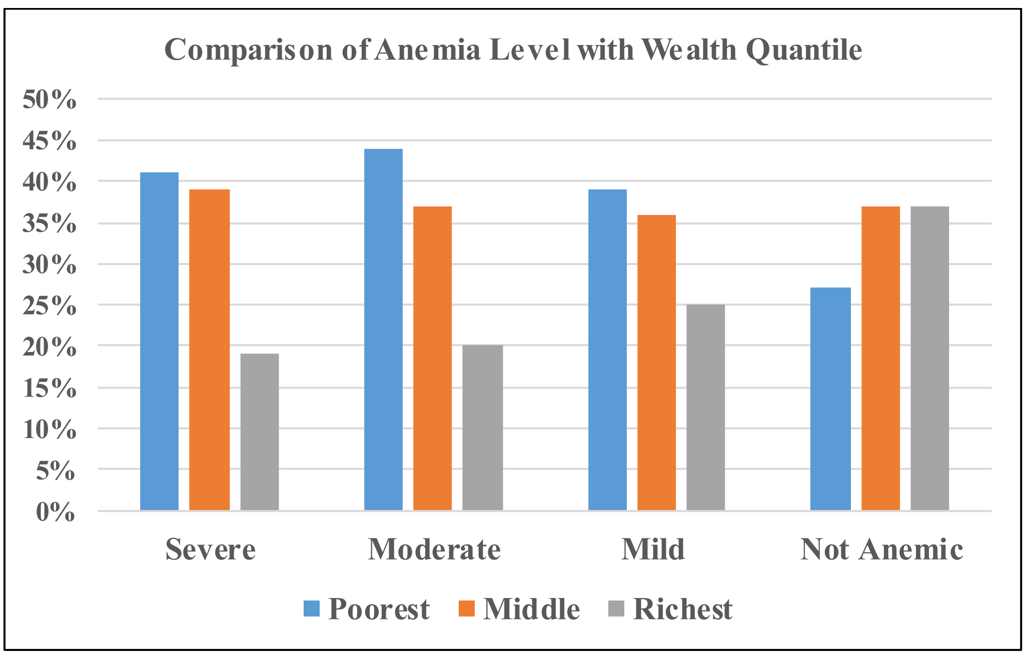
From figure 2, we can say that there is high portion (40%) of the poorest population suffers from moderate anemia. It highlights that a significant portion (26%) of the poorest population suffers from mild anemia. And very small number of people in these group have severe anemia. This visualization underscores the correlation between wealth and health, indicating that economic factors significantly impact anemia prevalence. But comparing anemic and not anemic groups 69% population is anemic and remaining 31% not suffer from anemia.

From figure 3, we can say that there is high portion (43%) of the poorest population are not anemic. It highlights that a significant portion (29%) of the poorest population suffers from moderate anemia. And the portion (26%) is suffered from mild anemia. very small number of people in these group have severe anemia. But comparing anemic and not anemic groups 57% population is anemic and remaining 43% not suffer from anemia.

From overall we can see that the, this visualization underscores the correlation between wealth and health, indicating that economic factors significantly impact anemia prevalence.

Comparison of Anemia Level with Wealth Quantile

	Poorest	Middle	Richest
Severe	41%	39%	19%
Moderate	44%	37%	20%
Mild	39%	36%	25%
Not Anemic	27%	37%	37%

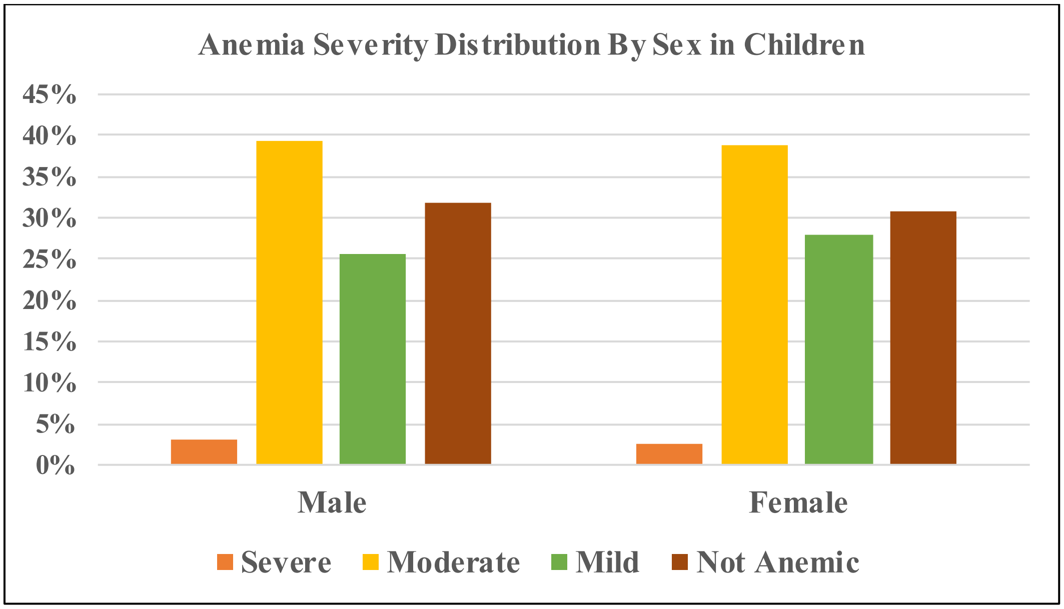


Interpretation:

From the above figure here we can see that the household wealth of a children is impact of children health if the children is poorest or middle household wealth then there is high chance of, they suffer from anemia as compare to Richest household wealth.

Anemia Severity Distribution by Sex in Children

	Male	Female
Severe	3%	3%
Moderate	39%	39%
Mild	26%	28%
Not Anemic	32%	31%

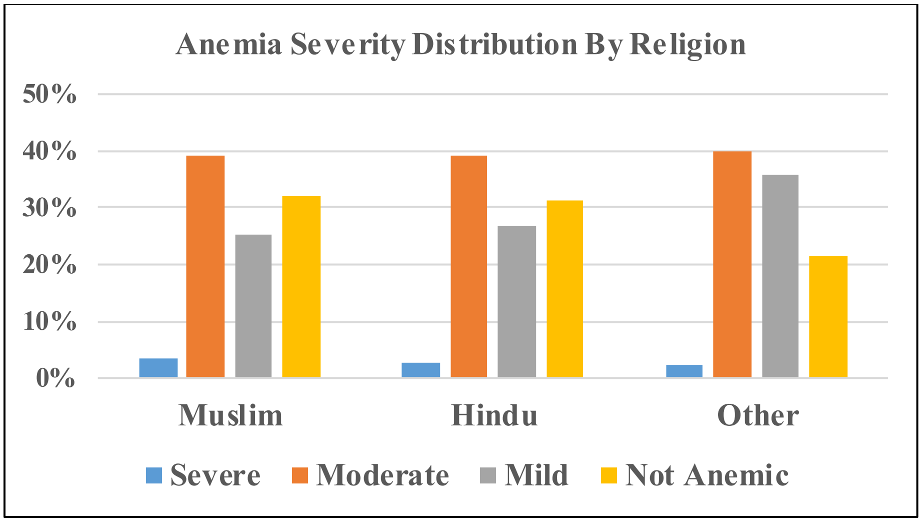


Interpretation:

From the above plot we can say that the anemia severity is approximately equal in male as well as female child. So, there may be no association between anemia severity and the sex in children.

Anemia Severity Distribution by Religion

	Muslim	Hindu	Other
Severe	3%	3%	3%
Moderate	39%	39%	40%
Mild	25%	27%	36%
Not Anemic	32%	31%	22%

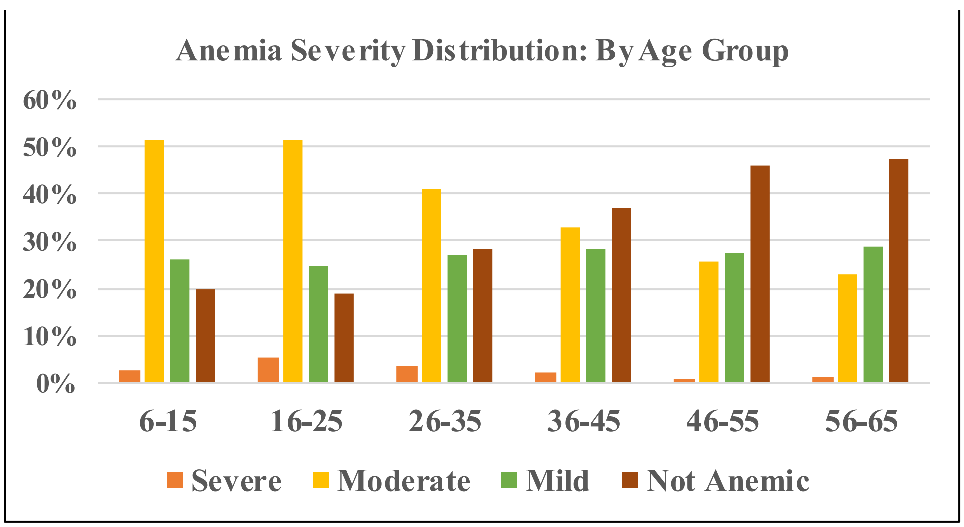


Interpretation:

From the above plot we can say that the anemia severity is approximately equal in all religions. So, there may be no association between anemia severity and the religion.

Anemia Severity Distribution: By Age Group

	6-15	16-25	26-35	36-45	46-55	56-65
Severe	3%	5%	4%	2%	1%	1%
Moderate	52%	52%	41%	33%	26%	23%
Mild	26%	25%	27%	28%	27%	29%
Not Anemic	20%	19%	29%	37%	46%	47%



Interpretation:

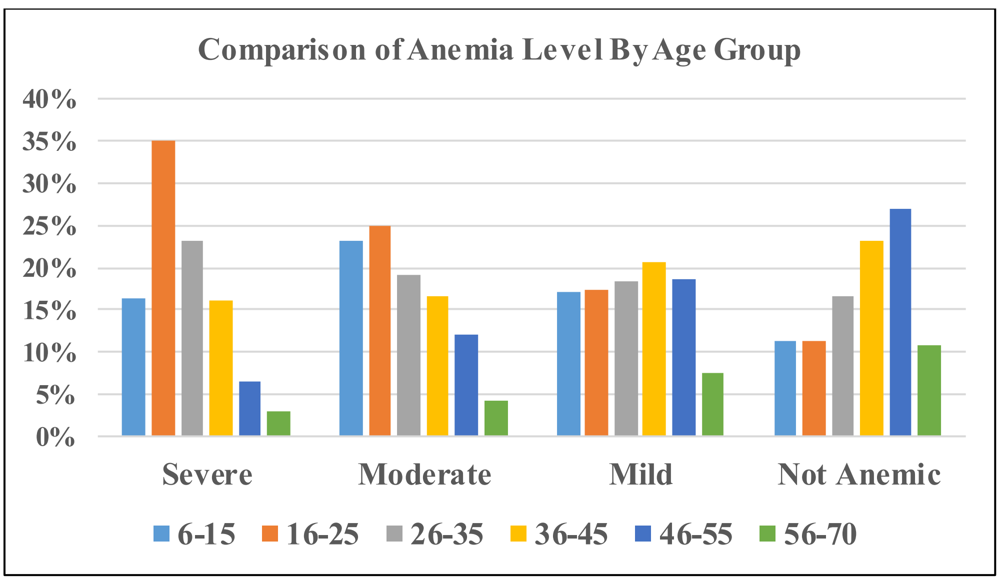
- For age group 6-15, Patients of moderate anemia is highest as compare to other.

- For age group 16-25, Patients of moderate anemia is highest as compare to other.
- For age group 26-35, Patients of moderate anemia is highest as compare to other.
- For age group 36-45, Non anemic patients are more in this group. But cases of moderate anemia are highest as compare to severe and mild.
- For age group 46-55, Non anemic patients are more in this group. But cases of mild anemia are slightly higher then moderate.
- For age group 55-56, Non anemic patients are more in this group. But cases of mild anemia are slightly higher then moderate.

Above plot reveals that there is negative association between anemia severity and the age of the children. As Age increases the anemia severity decreases.

Comparison of Anemia Level by Age Group

	6-15	16-25	26-35	36-45	46-55	56-70
Severe	16%	35%	23%	16%	6%	3%
Moderate	23%	25%	19%	17%	12%	4%
Mild	17%	17%	18%	21%	19%	8%
Not Anemic	11%	11%	17%	23%	27%	11%



Interpretation:

From these plot we can observe that, Patients of severe and moderate anemia is highest in age group 16-25. In age group 36-45 cases of mild anemia is slightly maximum. And in age group 46-55 more number of children has not anemia.

STATISTICAL ANALYSIS

Chi-Squared Test for Independence of Attributes

The main aim of this project is to find the association between status of anemia among all independent variables for this. In this section chi-square test is used.

H_0: The Two Attributes A and B Are Independent

Against

H_1: The Two Attributes A and B Are Dependent.

The R-software Following Command is used for performing the test: -

Chisq. test (y, conf. level=correct=F)

Let's we have to calculate the which variables are associated with anemia status

Variables	**Chi-Square Value**	**P Value**	**Decision**
MOTHERAGE	9990.4	2.20E-16	Reject H_0
KIDAGEMO	607.2	2.20E-16	Reject H_0
RELIGION	5.42	0.06654	Accept H_0
RESIDENT	1.358	0.2439	Accept H_0
ETHENTICITYIA	91.005	2.20E-16	Reject H_0
ETHENTICITYIA2	81.05	2.20E-16	Reject H_0
MOTHERCURRWORK	10.875	0.00097	Reject H_0
WEALTHQ	306.29	2.20E-16	Reject H_0
EDUCLVL	223.27	2.20E-16	Reject H_0
KIDSEX	1.5782	0.209	Accept H_0
MATERNITYIDX	19.537	7.72E-05	Reject H_0
DIARRECENT	19.002	1.31E-05	Reject H_0
KIDBORD	54.225	1.01E-11	Reject H_0
URBAN	87.222	2.20E-16	Reject H_0
HEALTHIDX	19.537	5.72E-05	Reject H_0

Interpretation:

From the Chi Square test table, we can say that the, if the P-Value is less then level of significance (0.05) then we can reject null hypothesis. And from that criteria we can say the variables age of mother, age of children in month, ethnicity, ethnicity2, mothers current work status, household wealth, highest education of mother, number of children in maternal history, Children had diarrhea recently, Child Birth order number, urban rural status, number of children in health history are significantly associated with the status of anemia.

Logistic Regression Algorithm:

> data=read.csv("E:/TD.csv")

> model <- glm (ANEMIASTATUS ~.,data=train_data)

> summary(model)

Call:

glm(formula = ANEMIASTATUS ~ ., data = train_data)

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	1.0728983	0.0566269	18.947	< 2e-16 ***
URBAN	-0.0203175	0.0116378	-1.746	0.080873
MOTHERAGE	-0.0049898	0.0012865	-3.879	0.000106 ***
RESIDENT	0.0074616	0.0178419	0.418	0.675806
RELIGION	-0.0053944	0.0150302	-0.359	0.719674
ETHNICITYIA	-0.0127424	0.0052515	-2.426	0.015266 *
ETHNICITYIA2	0.0581627	0.0138125	4.211	2.57e-05 ***
MOTHERCURRWORK	0.0153304	0.0100548	1.525	0.127368
WEALTHQ	-0.0661085	0.0086309	-7.660	2.04e-14 ***
EDUCLVL	-0.0244475	0.0059296	-4.123	3.77e-05 ***
MATERNITYIDX	0.0116171	0.0073796	1.574	0.115469
KIDSEX	0.0085886	0.0088775	0.967	0.333334
KIDBORD	0.0244260	0.0054816	4.456	8.44e-06 ***
KIDAGEMO	-0.0067884	0.0003083	-22.021	< 2e-16 ***
DIARRECENT	0.0044415	0.0154782	0.287	0.774156

Sign if. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gaussian family taken to be 0.1955034)

Null deviance: 2150.9 on 9994 degrees of freedom

Residual deviance: 1951.1 on 9980 degrees of freedom

AIC: 12068

Number of Fisher Scoring iterations: 2

Logistic regression model

$$\pi(x)=\frac{e^{1.0728983-0.0049898X1-0.0127424X2+0.0581627X3-0.0661085X4-0.0244475\ X5+0.0244260X6-0.0067884X7}}{1+e^{1.0728983-0.0049898X1-0.0127424X2+0.0581627X3-0.0661085X4-0.0244475\ X5+0.0244260X6-0.0067884X7}}$$

Hypothesis Testing:

Hypothesis to be tested,

H_0: Fitted Model is Not Adequate.

v/s

H_1: Fitted Model is Adequate.

From above output,

Null deviance: 2150.9 on 9994 degrees of freedom

Residual deviance: 1951.1 on 9980 degrees of freedom

G= Null Deviance – Residual Deviance

= 2150.9 – 1951.1

G = 199.8

$\chi^2_{7,0.05}$ = 14.067

Here G > $\chi^2_{7,0.05}$

G exceeds the critical value. This indicates that the model significantly improves the fit over the null model, suggesting that at least one of the predictor variables contributes significantly to the prediction of the outcome.

Therefore, we may have reject H_0 at 5% level of significance. Therefore, fitted model is Adequate.

Logistics	Values
Accuracy	0.7012
Sensitivity	0.1878
Specificity	0.9367
f1 Score	0.3130

Accuracy (0.7012): This tells us how often the model is correct overall. In this case, it means the model accurately predicts outcomes about 70.12% of the time.

Sensitivity (0.1878): Also known as the true positive rate, sensitivity measures the proportion of actual positives that are correctly identified by the model. Here, it means the model correctly identifies about 18.78% of the positive outcomes.

Specificity (0.9367): This is the true negative rate, showing how well the model identifies actual negatives. In this study, it means the model correctly identifies about 93.67% of the negative outcomes.

F1 Score (0.3130): This is a combined measure of precision and recall (or sensitivity), providing a balance between them. It's a useful metric for understanding overall model performance, where higher values indicate better performance. Here, the F1 score is about 0.3130.

From this logistic regression model we have seen that the model is good fit for data. Anemia status is mainly depending on Mother Age, Ethnicity, Ethnicity 2, household wealth, Mothers education level, children birth order, children age in months.

Odds Ratio:

Variable	β	ψ
MOTHERAGE	$\hat{\beta}_1$ = -0.0049898	$\widehat{\psi 1} = e^{-0.0049898}$ = 0.9950 < 1
ETHNICITYIA	$\hat{\beta}_2$ = -0.0127424	$\widehat{\psi 2} = e^{-0.0127424}$ = 0.9873 < 1
ETHNICITYIA2	$\hat{\beta}_3$ = 0.0581627	$\widehat{\psi 3} = e^{0.0581627}$ = 1.0598 > 1
WEALTHQ	$\hat{\beta}_4$ = -0.0661085	$\widehat{\psi 4} = e^{-0.0661085}$ = 0.9360 < 1
EDUCLVL	$\hat{\beta}_5$ = -0.0244475	$\widehat{\psi 5} = e^{-0.0244475}$ = 0.9758 < 1
KIDBORD	$\hat{\beta}_6$ = 0.0244260	$\widehat{\psi 6} = e^{0.0244260}$ = 1.0247 > 1
KIDAGEMO	$\hat{\beta}_7$ = -0.0067884	$\widehat{\psi 7} = e^{-0.0067884}$ = 0.9932< 1

Interpretation:

1] One-unit increase in Mothers Age, Anemia Status will decrease by 0.5%

2] One-unit increase in Ethnicity, Anemia Status will decrease by 1.77%

3] One-unit increase in Ethnicity2, Anemia Status will increase by 5.98%

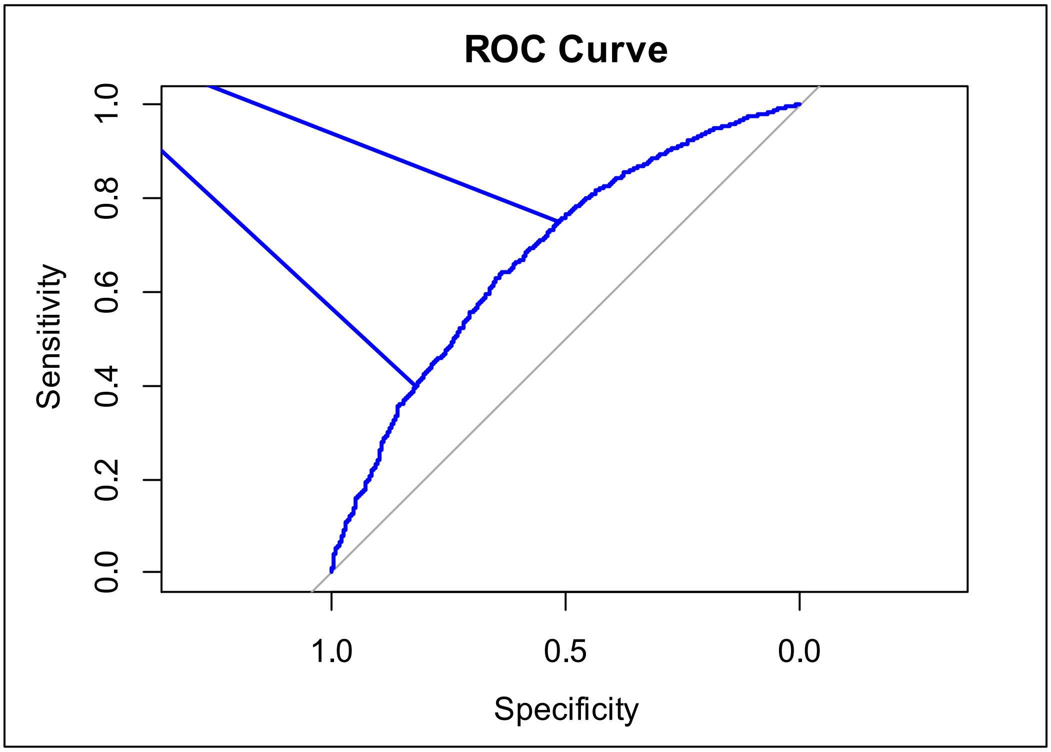
4] One-unit increase in Household Wealth, Anemia Status will decrease by 6.4%

5] One-unit increase in Education level, Anemia Status will decrease by 2.4%

6] One-unit increase in Children birth order, Anemia Status will increase by 2.4%

7] One-unit increase in Age of Children in months, Anemia Status will decrease by 0.6%

ROC Curve:



Interpretation:

An AUC (Area Under the Curve) value of 0.7 indicates that the anemia status of children is moderately effective.

Decision Tree Algorithm

```
> set.seed(2)

> Data=read.csv("E:/TD.csv")

> fit<-rpart(ANEMIASTATUS~., data_train,method='class');fit

> summary(fit)

Call:

rpart (formula = ANEMIASTATUS ~ ., data = data_train, method = "class")

  n= 6872

        CP          nsplit  rel error     xerror       xstd

1 0.02613159      0     1.0000000    1.0000000   0.01791975

2 0.01000000      2     0.9477368    0.9486701   0.01765561
```

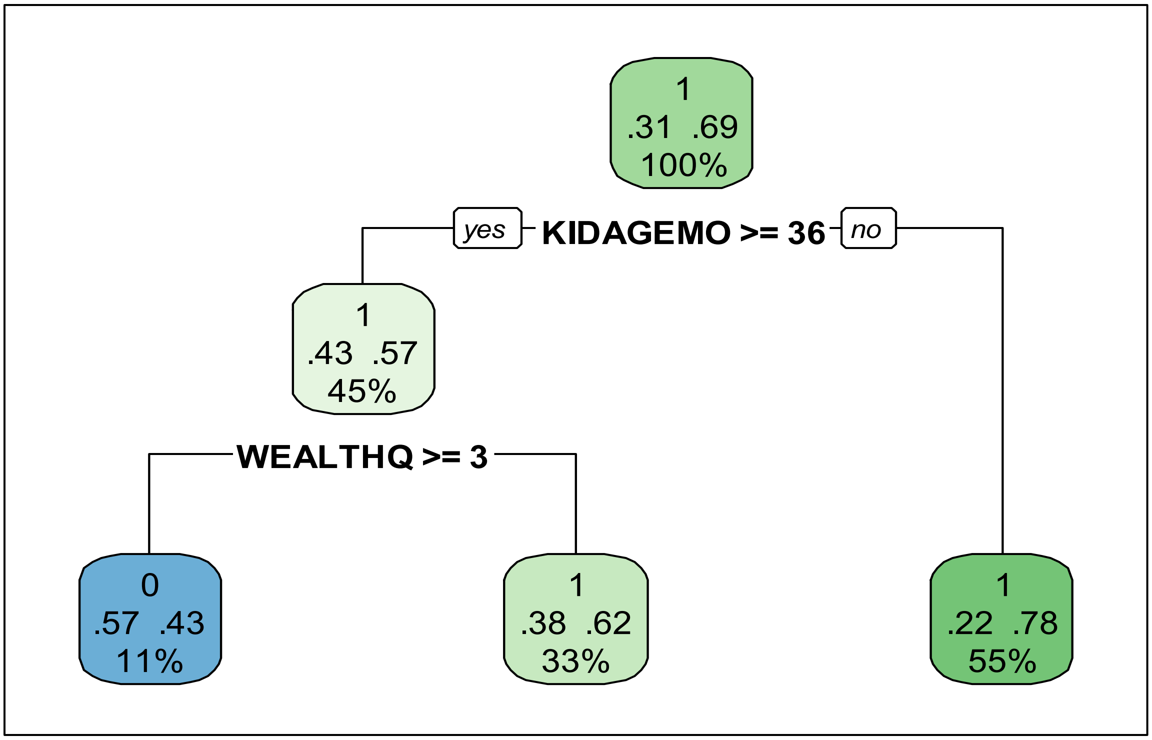
Variable importance

KIDAGEMO	WEALTHQ	MOTHERAGE	URBAN	EDUCLVL
63	**20**	**6**	**5**	**5**

The model predicts anemia status (1 or 0) based on several variables. In this model summary we can get a most important variables according to their score. Variable importance indicates their degree of influence or contribution of each variable towards anemia status.

Here, the age of children in months (KIDAGEMO) are most important because their variable importance score is high as compared to others. After that household wealth (WEALTHQ) is at second place. And the other variables age of mother (MOTHERAGE), urban rural status (URBAN), mother's education level (EDUCLVL) are slightly decreases according their variable importance score.

```
> rpart.plot( fit,extra=104)
```



Decision Tree Plot

Interpretation:

From the above decision tree we can see that the only household wealth and children age in month is associated to anemia. but from variable importance table we can say that the age of children in months, household wealth, age of mother, Urban Rural status, mother's education level is found to be most significant factors associated to the anemia in children according to decision tree. To make more conclusion there was need of model assessment. confusion matrix was drawn to find model accuracy.

> confusionMatrix(table_mat)

Confusion Matrix

	0	1
0	219	771
1	172	1961

> accuracy_test<-sum(diag(table_mat))/sum(table_mat);accuracy_test

[1] 0.6980467

```
> sensitivity=table_mat[2,2]/sum(table_mat[2,]);sensitivity
[1] 0.9193624
> specificity=table_mat[1,1]/sum(table_mat[1,]);specificity
[1] 0.2212121
> precision=table_mat[2,2]/sum(table_mat[,2]);precision
[1] 0.7177892
> f1_score=2*(precision*sensitivity)/(precision+sensitivity);f1_score
[1] 0.8061665
```

Conclusion:

1] From Confusion matrix we can conclude that,

- True positive (TP) = The Children anemia status is correctly classified as children is anemic. Here we observe that 1961 children anemia status is correctly classified as children is anemic by the model.
- True Negative (TN) = The children anemia status is correctly classified as children is not anemic. Here we observe that 219 children anemia status is correctly classified as children is not anemic by the model
- False Positive (FP) = The children anemia status is not anemic but it is classified as children is anemic. Here we observe that 771 children anemia status is misclassified as children is anemic by the model.

- False Negative (FN) = The children anemia status is children is anemic but it is misclassified as children is not anemic. Here we observe that 172 children anemia status is misclassified as children is not anemic by the model.

2] Accuracy (0.6980): This tells us how often the model is correct overall. In this case, it means the model accurately predicts outcomes about 69.80% of the time.

3] Sensitivity (0.9193): Also known as the true positive rate, sensitivity measures the proportion of actual positives that are correctly identified by the model. Here, it means the model correctly identifies about 91.93% of the positive outcomes.

4] Specificity (0.2212): This is the true negative rate, showing how well the model identifies actual negatives. In this study, it means the model correctly identifies about 22.12% of the negative outcomes.

5] F1 Score (0.8061): This is a combined measure of precision and recall (or sensitivity), providing a balance between them. It's a useful metric for understanding overall model performance, where higher values indicate better performance. Here, the F1 score is about 0.8061.
In summary, the decision tree model shows moderate performance in predicting anemia status. It correctly identifies anemia cases (high sensitivity) but has room for improvement in precision and overall accuracy.

Model Comparison

Sr. No	Model	Accuracy
1	Logistic Regression	0.7012337
2	Decision Tree	0.6980467

From the above table we can see that both the models show approximately same accuracy which is 70%. The main goal of this study is to examine the significant factors associated with the anemia in children.

Therefore, in the next section comparison table of variable importance was examined.

Comparison table for variable importance

Chi-Square Test	Logistic Regression Model	Decision Tree Algorithm
1. MOTHERAGE 2. KIDAGEMO 3. ETHENTICITYIA 4. ETHENTICITYIA2 5. MOTHERCURRWORK 6. WEALTHQ 7. EDUCLVL 8. DIARRECENT 9. MATERNITYIDX 10. KIDBORD 11. URBAN 12. HEALTHIDX	1. MOTHERAGE 2. ETHNICITYIA 3. ETHNICITYIA2 4. WEALTHQ 5. EDUCLVL 6. KIDBORD 7. KIDAGEMO	1. KIDAGEMO 2. WEALTHQ 3. MOTHERAGE 4. URBAN 5. EDUCLVL

Interpretation:

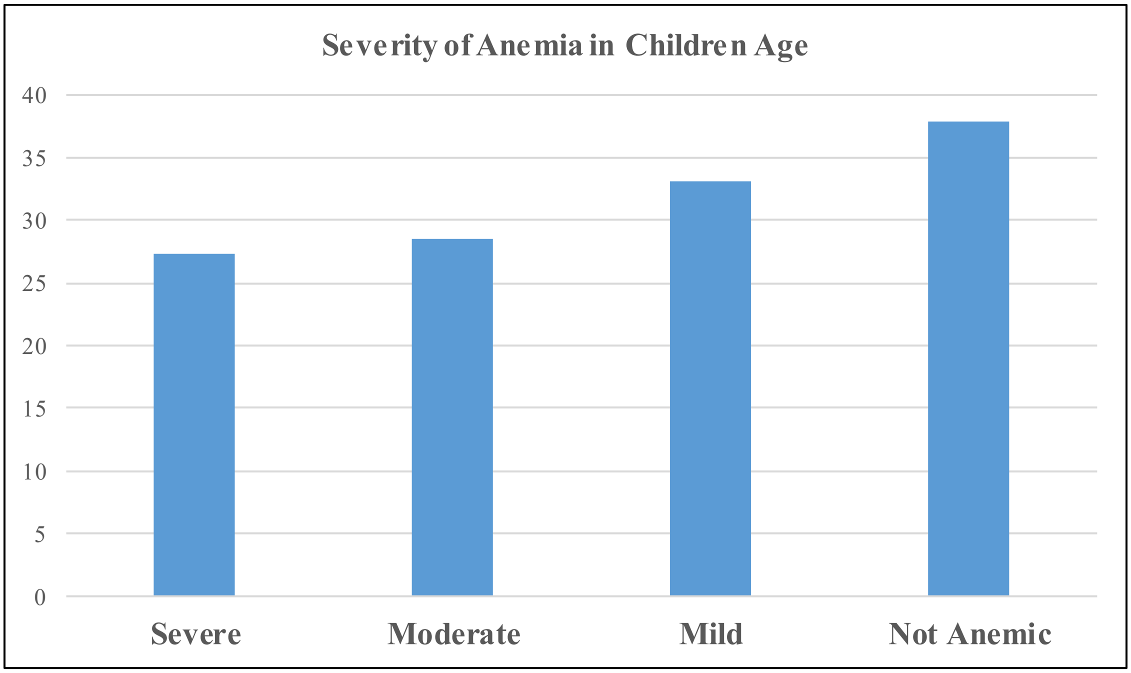
From comparison table of variable importance, we can observe that KIDAGEMO, WEALTHQ, MOTHERAGE, EDUCLVL these four variables are found to be common among all three techniques.

Therefore, in the next section we examine pattern of these variables according to anemia severity.

Examination of pattern between significant factors and anemia severity

Relationship between Age of Children and Anemia Status:

	Severe	Moderate	Mild	Not Anemia
Average of Children Age	27.28928571	28.44984647	33.14061331	37.8078519

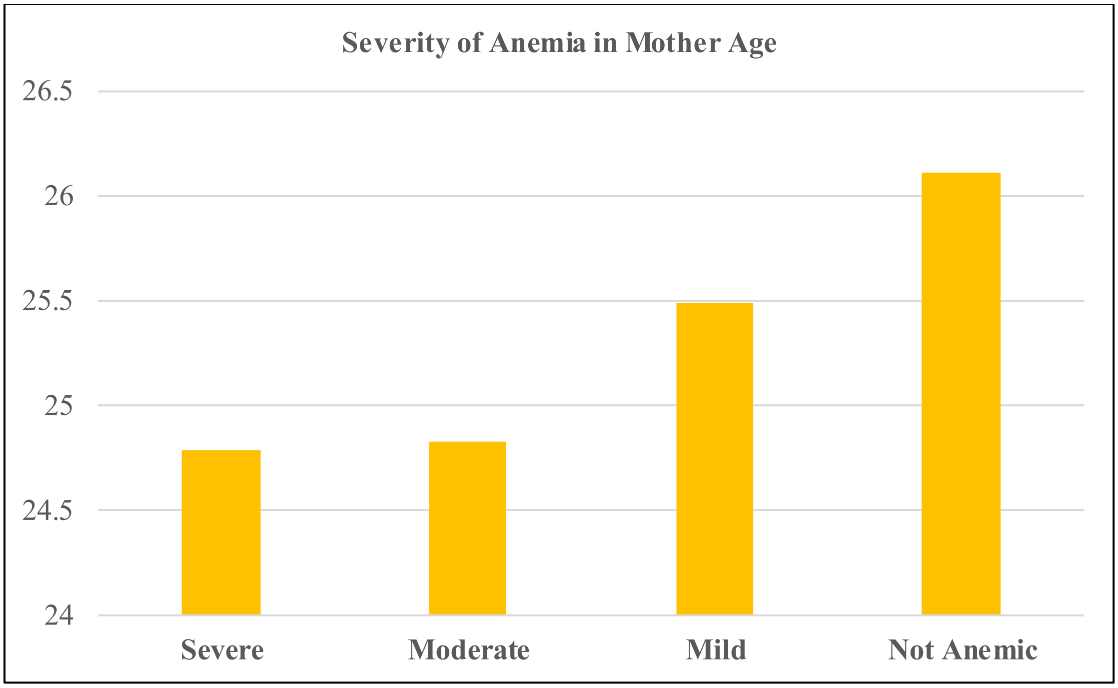


Interpretation:

We observe that as the average age of children increases then the level of anemia in children decreases slightly.

Relationship between Mother Age of Children and Anemia Status:

	Severe	Moderate	Mild	Not Anemia
Average of Mother Age	24.79285714	24.82753327	25.48616305	26.11459



Interpretation:

We observe that as the average age of mother increases then the level of anemia in children decreases slightly.

we can observe that age of children in months (KIDAGEMO), wealth quantile (WEALTHQ), age of mother (MOTHERAGE), education levels of mother (EDUCLVL) these four variables are significant but WEALTHQ, EDUCLVL are already examined in the exploratory data analysis. that's why here we use only two variables for examination of pattern which is age of children and age of mother.

CONCLUSION

- It was discovered that there were only few number of children who have Severe Anemia which is good figure but it was sensitive having more number of children who have Moderate Anemia, as compare to children who have Mild Anemia.

  From this study we have observed that the prevalence of anemia among age group 6 to 59 months' children in India there is approximately 69 % children have anemia.
- There is a severe condition for rural area because it has maximum number of moderate anemia patients.
- Mother's education found to be significantly associated with the child's anemia. As As mother's education level increases anemia severity decreases. Therefore, Mother is educated or not is important for children anemia status. Mother's education that might impact on a children's health.
- According to household wealth we can observe that, there is maximum number of cases of severe, moderate, mild anemia if the children having poorest economical condition. Children household wealth impact on children's health.
- We have observed that, there is a severe condition for age 6-15,16-25 and 26-35 in months because it has maximum number of moderate anemia patients. So overall we say that, As the age of children increase the effect of anemia on children is slightly minimizes.
- From the model comparison we have seen both the models were good fitted for data as have accuracy approximately 70%.
- According to odds ratio, the anemia status of a child is influenced by Mothers age, ethnicity, wealth quantiles, education levels, number of children born, and age of older children all play a role. Specifically: if the mother age is slightly increases then there is less chance of having an anemic child. Some ethnic groups have slightly probability of childhood anemia. Higher wealth quantiles and education levels correlate with decreased anemia risk. Families with more children tend to have higher odds of childhood anemia. The age of children impacts anemia outcomes.
- From examination of pattern between significant factors and anemia severity we can observed that the average of children age and average of mother age increases then the anemia in children is slightly decreases.

SCOPE & LIMITATION

Scope:

1. Integrating biomedical data with social determinants of health to enhance the prediction of anemia occurrence.
2. Collaborating with nutrition science to identify dietary patterns and micronutrient deficiencies associated with anemia risk.
3. Incorporating epidemiological factors such as geographical location and population demographics into predictive modeling for anemia prevalence estimation.
4. Engaging with healthcare informatics to develop user-friendly predictive tools that assist clinicians in early identification and intervention of anemia cases.

Limitation:

Handling datasets from the Demographic and Health Surveys (DHS) comes with its own set of limitations:

1. DHS surveys collect a wide range of demographic and health-related information, but they may lack certain variables of interest for specific research questions, necessitating supplementation with additional data sources.
2. DHS surveys are conducted periodically, with gaps between survey rounds, which may limit the ability to analyze trends over shorter time periods or capture recent developments.
3. Analyzing DHS datasets often requires substantial computational resources and expertise in statistical software, which may pose challenges for researchers with limited access to such resources.

REFERENCE

- Demographic and Health Survey's (DHS) Program.
  https://dhsprogram.com/Data/terms-of-use.cfm
- Statistical Methods & Use of R – Software By Prof P.G. Dixit, Prof P. S. Kapare
- IIPS (2016). Ministry of Health and Family Welfare National Family Health Survey-5, 2019-20. State Fact Sheet West...
- 1. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). Classification and Regression Trees. (Wadsworth and Brooks/Cole).
- Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques, third edition (3rd ed.). Morgan Kaufmann Publishers
- WHO (2014) WHO Global Nutrition Targets 2025: Anemia Policy Brief. URL: https://www.who.int/nutrition/topics/globaltargets_anaemia_policybrief.pdf
- KK Patel · 2021 · Cited by 14 — Overall, the prevalence of anemia was found to be 58 percent.https://www.sciencedirect.com/science/article/abs/pii/S0190740920323409
- Prevalence and Triggering Factors of Childhood Anemia: An Application of Ordinal Logistic Regression Model
  Md. Akhtarul Islam, Sohani Afroja, Md. Salauddin Khan, Sharlene Alauddin, Mst. Tanmin Nahar, and Ashis Talukder
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9159194/
- Determinants and transition of anemia among under-five children from different social groups in India from 2005–06 to 2015-16 2023, Social Science and Medicine
- Anemia Prevalence in India Over Two Decades: Evidence from National Family Health Survey (NFHS)
  December 2021International Journal of Science and Healthcare Research 6(4):335-340
  https://www.researchgate.net/publication/357116282_Anemia_Prevalence_in_India_Over_Two_Decades_Evidence_from_National_Family_Health_Survey_NFHS
- Prevalence of anaemia among 6- to 59-month-old children in India: the latest picture through the NFHS-4
  Published online by Cambridge University Press: 20 May 2019
  Susmita Bharati,Manoranjan Pal and Premananda Bharati

APPENDIX

Code:

*Chi-Squared Test for Independence of Attributes

```
A = scan("clipboard")
B = scan("clipboard")
Mx =c (A, B)
C= chisq. Test (Mx)
```

*Logistic Regression Model

```
# Load necessary libraries
library(datasets)
# Load the dataset
Data=read.csv("E:/TD.csv")
# Split the data into training and testing sets
set. seed (123)
train index <- sample(1:nrow(Data), 0.7*nrow(data))
train_data <- data[train_index, ]
test_data <- data[-train_index, ]
# Fit logistic regression model
model <- glm(ANEMIASTATUS~.,data=train_data)
# Summary of the model
summary(model)
# Make predictions on the test set
predictions <- predict(model, newdata = test_data, type = "response")
# Convert probabilities to class predictions
predicted_classes <- ifelse(predictions > 0.5, 1, 0)
# Evaluate the model
accuracy <- mean (predicted_classes == test_data$ANEMIASTATUS)
cat("Accuracy:",accuracy
```

*Decision Tree Algorithm

```
set. seed (2)
Data=read.csv("E:/TD.csv")
head(Data)
tail(Data)
colnames(Data)
na.omit(Data)
dim(Data)
library(rpart)
library(rpart.plot)
#create train/test dset
library(caTools)
split=(sample.split(Data,SplitRatio = 0.7))
data_train=subset(Data,split==T)
data_test=subset(Data,split==F)
dim(data_train)
dim((data_test))
prop.table(table(data_train$ANEMIASTATUS))
prop.table(table(data_test$ANEMIASTATUS))
fit<-rpart(ANEMIASTATUS~.,data_train,method='class');fit
summary(fit)
rpart.plot(fit,extra=104)
library(e1071)
library(caret)
predict_unseen<-predict(fit,data_test,type="class")
table_mat<-table(data_test$ANEMIASTATUS,predict_unseen);table_mat
accuracy_test<-sum(diag(table_mat))/sum(table_mat);accuracy_test
print (paste('Accuracy for Test',accuracy_test))
# calculate F1 score
f1_score <- function(table_mat) {
  precision <- diag(table_mat) / colSums(table_mat)
```

```
  recall <- diag(table_mat) / rowSums(table_mat)
  f1_score <- 2 * precision * recall / (precision + recall)
  mean (f1_score, na.rm = TRUE)
}
f1_score(table_mat)
confusionMatrix(table_mat)
sensitivity=table_mat[2,2]/sum(table_mat[2,]);sensitivity
specificity=table_mat[1,1]/sum(table_mat[1,]);specificity
precision=table_mat[2,2]/sum(table_mat[,2]);precision
f1_score=2*(precision*sensitivity)/(precision+sensitivity);f1_score
```