

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“JNANASANGAMA” BELAGAVI - 590 018

KARNATAKA



REPORT OF INTERNSHIP/PROFESSIONAL PRACTICE

Carried out in

Eunoia Labs, Bangalore



SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF
BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE & ENGINEERING

Submitted by:

SHEETHAL K S
[1CG20CS092]

INTERNAL GUIDE

Mrs. Rashmi C R M.Tech.,
Assistant Professor,
Dept. of CSE,
C.I. T, Gubbi, Tumkur.

EXTERNAL GUIDE

Mrs. Roopa K S
H R Manager
Eunoia Labs
Bangalore

HOD
Dr. Shantala C P
Professor & Head,
Dept. of CSE
CIT, Gubbi

Channabasaveshwara Institute of Technology

(Affiliated to VTU, Belgaum & Approved by AICTE, New Delhi)

(NAAC Accredited & ISO 9001:2015 Certified Institution)

NH 206 (B.H. Road), Gubbi, Tumkur – 572216. Karnataka



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

2023-2024



Channabasaveshwara Institute of Technology

(Affiliated to VTU, Belgaum & Approved by AICTE, New Delhi)
(NAAC Accredited & ISO 9001:2015 Certified Institution)

NH 206 (B.H. Road), Gubbi, Tumkur – 572216. Karnataka.



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

2023-2024

CERTIFICATE

This is to certify that the internship entitled “**ANEMIA PREDICTION USING MACHINE LEARNING**” has been carried out by **SHEETHAL K S - [1CG20CS092]** bonafide student of **CHANNABASAVESHWARA INSTITUTE OF TECHNOLOGY, GUBBI, TUMKUR**, in partial fulfillment of the requirement for the award of the degree **Bachelor of Engineering** in **COMPUTER SCIENCE & ENGINEERING** from the **Visvesvaraya Technological University, Belagavi** during the year **2023-2024**. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report. The Internship report has been approved as it satisfies the academic requirements in respect of Internship/Professional practice prescribed for the said degree.

Signature of Guide

Signature of Internship Coordinator

Mrs. Rashmi C R M.Tech.,
Assistant Professor,
Dept., of CSE
C.I.T, Gubbi.

Mrs. Rashmi C R M.Tech.,
Assistant Professor,
Dept., of CSE
C.I.T, Gubbi.

Signature of HOD

Signature of Principal

Dr. Shantala C P Ph.D.,
Professor & Head,
Dept., of CSE
C.I.T, Gubbi.

Dr. SURESH D S Ph.D.,
Director & Principal
C.I.T, Gubbi.

External Viva

Examiners Name

Signature with Date

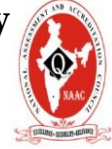
1. _____

2. _____



Channabasaveshwara Institute of Technology

(Affiliated to VTU, Belgaum & Approved by AICTE, New Delhi)
(NAAC Accredited & ISO 9001:2015 Certified Institution)
NH 206 (B.H. Road), Gubbi, Tumkur – 572216. Karnataka.



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

2023-2024

UNDERTAKING

I, **SHEETHAL K S** bearing **1CG20CS092**, student of **VIII Semester B.E.** in **COMPUTER SCIENCE & ENGINEERING, C.I.T, GUBBI, TUMKUR** hereby declare that the Internship carried out in **Eunoia Labs, Bangalore** and submitted in partial fulfillment of the requirements for the award of the degree **Bachelor of Engineering** in **COMPUTER SCIENCE & ENGINEERING** of the **Visvesvaraya Technological University, Belagavi** during the academic year 2023- 2024.

Place: GUBBI

Date:

SHEETHAL K S

1CG20CS092



Channabasaveshwara Institute of Technology

(Affiliated to VTU, Belgaum & Approved by AICTE, New Delhi)
(NAAC Accredited & ISO 9001:2015 Certified Institution)

NH 206 (B.H. Road), Gubbi, Tumkur – 572216. Karnataka.



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

2023-2024

BONAFIDE CERTIFICATE

This is to certify that the Internship carried out in **Eunoia Labs, Bangalore** is a bonafide work of SHEETHAL K S– [1CG20CS092], student of **VIII** semester **B.E.- COMPUTER SCIENCE & ENGINEERING** from **Channabasaveshwara Institute of Technology, Gubbi, Tumkur**, in partial fulfillment of the requirements for the award of degree **B.E., in COMPUTER SCIENCE & ENGINEERING** of **Visvesvaraya Technological University, Belgaum** during the academic year 2023-2024. It is certified that the Internship work carried out was under my supervision and guidance.

Guide

Mrs. Rashmi C R M.Tech.,
Assistant Professor
Dept., of CSE
C.I.T, Gubbi.

Date: 11-09-2023

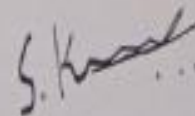
CERTIFICATE OF INTERNSHIP

TO WHOM IT MAY CONCERN

This is to certify that Ms. Sheethal K S (ICG20CS092) studying B.E (Computer Science and Engineering) in Channabasaveshwara Institute of Technology, Gubbi, Tumakuru District has undergone Internship on Machine Learning from 14-August-2023 to 09-September-2023 as a part of BE Programme.

The performance, conduct and character of Ms. Sheethal K S have been very good to the best of our knowledge.

For Eunoia Labs



Tel: +91-9606017353 Email: info@eunoialabs.in

www.eunoialabs.in

Shan, 1st Block, 3rd Main,
Kuvempunagar, Tumkur 572102

#280, 59th Cross, 17th Main, 3rd Block,
Rajajinagar, Bangalore-560010

ACKNOWLEDGEMENT

Several special people have contributed significantly to this effort. First of all, I am grateful to my institution, **Channabasaveshwara Institute of Technology, Gubbi**, which provides me an opportunity in fulfilling my most cherished desire of reaching my goal.

I, acknowledge and express my sincere thanks to our beloved Director & Principal, **Dr. Suresh D S**, for his many valuable suggestion and continued encouragement by supporting me in mt academic endeavors.

I, express my sincere gratitude to **Dr. Shantala C P, Professor and Head, Department of CSE**, for providing her constructive criticisms and suggestions.

I, extend my gratitude to my Internship guide **Mrs. Rashmi C R, Assistant Professor**, Department of CSE, for her guidance, support and suggestions throughout the period of this Internship.

I express my deep sense of gratitude to **Eunoia Labs, Bangalore** for giving such an opportunity to carry out the internship in their esteemed industry/organization.

I sincerely thank **Mrs. Roopa K S, H R Manager, Eunoia Labs, Bangalore** for exemplary guidance and supervision.

Finally, I would like to thank all the individuals who supported me directly and indirectly for the successful completion of this internship work.

Sheethal K S [1CG20CS092]

ABSTRACT

Anemia is a state of poor health where there is presence of low amount of red blood cell in blood stream. This project aims to design a model for prediction of Anemia using machine learning algorithms. In our dataset there are 8,544 rows and 6 columns like Gender, Hemoglobin, MCH, MCHC, MCV and Result. In this project, we investigate the use of two popular machine learning algorithm, Random Forest and Decision Tree for detecting anemia. We begin by preprocessing dataset of anemia, including handling missing values and splitting the dataset. Then I train Random Forest and Decision Tree models on the preprocessed dataset and evaluate their performance using metrics such as accuracy, precision, recall, F1-score. After testing two different algorithms, the accuracies are compared.

CONTENTS

ACKNOWLEDGMENT	i
ABSTRACT	ii
CONTENTS	Page No.
1. COMPANY PROFILE AND TRAINING	
1.1. Company profile	1
1.2. Training	2-3
2. INTRODUCTION	4
2.1. Objectives	5
2.2. Problem Statement	5
3. LITERATURE SURVEY	6-7
4. SYSTEM DESIGN	
4.1. Architecture	8
4.2. Dataflow Diagram	9
4.3. Use Case Diagram	10
4.4. Sequence Diagram	11
5. METHODOLOGY	12-17
6. RESULTS	18-22
7. CONCLUSION	23
REFERENCES	24

CHAPTER 1

COMPANY PROFILE AND TRAINING

1.1 COMPANY PROFILE

Eunoia labs is currently in ‘stealth-mode’ developing products and services in the domain of ‘Education’. It is committed to Individualized and Contextual learning. Eunoia Labs has an experienced pool of experts and engineers with decades of domain experience and expertise in the fields of cloud Services, Open-Source tools, ERP for enhanced teaching-learning process, Foss Consultancy and training.

Eunoia Labs - Training division supports working professionals, students and faculty members to adopt advanced software and technology knowledge. Eunoia Labs has the unique courses designed for different levels of engineers and covers theory + practical + projects and internship program to enhance the skills and experience.

Services

- FOSS Consultancy

Facilitate Adoption of Open-source Software by Educational Institutions for both Academia and Administration and Management. We provide Deployment, Training and Customization of various opensource such as OpenStack, Moodle, Odoo, OSTicket, Simulators and emulators, etc.

- Trainings

Customized training on State-Of-Art technologies such as Containers, GoLang, SDN-NFV, etc.

Products

- Sententia: Sententia is a web-based formal writing assistant. Unlike other tools, Sententia does not focus on just correcting the grammar.
- OSTICKET: Online Ticketing System
- Mapaka: Adaptive Assessment Platform. [Under Development]
- Aurora: Individualized Learning platform. [Under Development]
- Duende: Q&A Based Teaching-Learning Platform. [Under Development]

1.2 TRAINING

Week 1:

In the 1st week of internship, we were assigned with a project based on “Machine Learning” and introduced to a few basics of Python, NumPy, Pandas and other libraries for developing the project. We were addressed by the respective guides and explained some concepts of Python. We were introduced to the concepts of NumPy, Pandas in detail.

- **NumPy:** NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematics, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.
- **Pandas:** Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labelled” data easy and intuitive. It is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring and manipulating data. Pandas allows us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant.

Week 2:

In the second week, we were introduced to the concepts of Matplotlib, Seaborn and Scikit-learn in detail.

- **Matplotlib:** Matplotlib is one of the most popular Python packages used for data visualization. It is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, GTK etc.
- **Seaborn:** Seaborn provides a high-level interface for drawing attractive and informative statistical graphics. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.

- **Scikit-learn:** Scikit-learn (formerly known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. On the second week, Respective guides assigned us a project on “Machine Learning” and explained us regarding the outlook and working of the project.

Week 3:

We were assigned with the task—To work on the algorithm in the machine learning such as Logistic Regression, Support Vector Machines, Decision tree, Random Forest, K-nearest neighbor etc. We were introduced to actual software’s required for the development of the project. We were trained about basics of Anaconda and Jupiter notebook required for the development of the actual project. Jupiter Notebook is a web-based interactive computational environment for creating notebook document.

I started collecting the datasets required for the project, then I apply data pre-processing method to check whether the dataset contains null values or not. Visualization methods are applied using histograms, bar plots to represent different aspects of the data. In the next step the data splitting takes place like they are divided into train and test in the ratio 80:20.

Week 4:

Here I choose the machine learning algorithms like Random Forest and Decision Tree to train the model. After the training of the model the evaluation give the accuracy of the model. At end I compared the accuracies of two models to know the which algorithm is best.

CHAPTER 2

INTRODUCTION

Anemia is one of the most common blood disorders worldwide and can affect all types of people. It is important to keep in mind that the majority of blood disorders are caused by abnormalities in particular genes and can be passed down through families. Blood disorders can also be caused by medical conditions, including drug use and lifestyle choices. According to reports, the most popular blood condition affecting humans is anemia. Anemia is a condition when there are either too few red blood cells or too little hemoglobin in them. A person's blood's ability to transport oxygen to the body's tissues will be reduced if they have insufficient or abnormally few red blood cells or if they have insufficient amounts of hemoglobin, which is required to deliver oxygen. This causes symptoms like weakness, exhaustion, feeling dizzy, and shortness of breath. Age, sex, smoking habits, and the status of pregnancy all affect the ideal hemoglobin concentration needed to meet physiologic needs.

According to the World Health Organization (WHO), one-quarter of the world's population suffers from anemia, which has a particularly negative effect on children between the ages of 6 and 59 months and pregnant women. The number of people who suffer from anemia grew globally by 0.3% in 2019 to 29.90%, it affects pregnant women and children at rates of 47.4% and 41.8%, respectively. As a result, societies where anemia is common suffer enormous economic losses. Although the disease's numerous symptoms make it difficult for people to diagnose the disorder due to its hidden nature, it is a significant and serious problem regardless. To decrease anemia's prevalence, it is essential to spread the necessary knowledge of its causes and symptoms to try to treat this disease as much as possible.

Anemia has many causes; Iron deficiency is thought to be the main cause of anemia worldwide, but other nutritional deficiencies (such as folate, vitamin B12, and vitamin A deficiency) can be the main causes. Additionally, anemia may result from acute or chronic inflammation, parasite infections, inherited or acquired disorders that disrupt the formation of red blood cells, the survival of those cells, or hemoglobin manufacturing for example, the defects in hemoglobin or the synthesis of abnormal hemoglobin cause different and more dangerous types of anemia, such as Sickle Cell Anemia and Thalassemia.

2.1 Objectives

- To develop a Anemia Prediction model using machine learning.
- Comparison of different algorithms to find which algorithm is best for the dataset.

2.2 Problem statement

Traditionally, diagnosing anemia has relied on subjective assessment methods, necessitating the development of machine learning models to provide objective and accurate predictions based on diverse patient data, thereby enhancing early detection and intervention strategies for improved healthcare outcomes.

CHAPTER 3

LITERATURE SURVEY

Machine Learning (ML) techniques have been widely used to detect various diseases over the past ten years. This makes an early diagnosis easier and raises the likelihood of survival.

Machine Learning has been an emerging tool for Prediction of Diseases. The work has figured out that each algorithm has its own strength as well as weakness and its own area of implementation. The authors identify those studies that applied more than one supervised machine learning algorithm on one disease prediction. Algorithms include Random Forest, Decision Tree, ANN, SVM, Logistic Regression, Naïve Bayes and K-nearest Neighbor. It shows that Support Vector Machine (SVM) algorithm is applied most frequently and Random Forest (RF) algorithm showed superior accuracy. The research illustrates that many machine learning algorithms have shown good results. It is so as they identify the related attributes accurately [1] .

The authors investigated about supervised machine learning algorithms Naive Bayes, Random Forest and Decision Tree algorithm for prediction of anemia using Complete Blood Count (CBC) where Naive-Bayes technique performed well in terms of accuracy as compared to Decision Tree and Random Forest. The work determined which individual classifier or subset of classifier in combination with each other achieves maximum accuracy in red blood cell classification for anemia detection showing unique idea of use of subset of classifier and use of ensemble learning techniques. specified anemia type for the anemic patients with dataset from the Complete Blood Count (CBC) which showed J48 Decision Tree as best performer [2].

The research predicted the anemia status of children under five years taking common risk factors as features. The research concluded that ML methods in addition to the classical regression techniques can be considered to predict anemia. The authors constructed some predictive models by using the identified risk factors through machine learning approach predict the anemia status of children under 36 months [3].

In, WEKA is employed to create an appropriate classifier for the creation of a mobile application that can anticipate and diagnose remarks made in hematological data. The J48 and Naive Bayes classifiers were put to the test against neural network classification techniques by the authors. The findings reveal that the 348 classifier has the highest level of accuracy [4].

A decision support system was developed by authors in to diagnose iron-deficient anemia using a decision tree algorithm. This system takes into account ferritin, serum iron-list capacity, and the three hematological parameters The evaluation was grounded in data from 96 cases, and the issues were favorably compared to the doctor's decision [5].

Estimating hemoglobin levels is a crucial stage in any blood analysis work, and it also establishes whether a person is anemic. In research, hemoglobin levels were determined, and anemia was identified using blood test features and a machine learning model. 9004 records make up the dataset, of which 6753 were utilized for training and 2251 for testing Three different machine learning algorithms-DT, NB, and NN-as well as a hybrid classifier, which combines all three methods, were used. Additionally, the performance of the methodology was evaluated using the MAE and RMSE approaches. According to the MAE findings, the hybrid classifier had an accuracy of 0.996% and the best RMSE value of 0.015[6].

In authors used machine learning algorithms such as Random Forest, SVM and others to predict whether the patient is anemic or not. They developed a classification- based ML model and used it to forecast a patient's anemia using crucial CBC test data [7].

As can be seen until now, the predictive power of machine learning has increased incredibly, so many of the studies have used several ML. methods such as Support Vector Machine, Random Forest, Naïve Bayes, Decision trees and Multi-Layer Perceptron to build a model to predict a person is anemic or not, So, in this paper, we are going to use some of these algorithms to make a predictive model and choose the best algorithms according to their accuracy in order to use them or combine it to predict another type of anemia or other diseases soon[8].

CHAPTER 4

SYSTEM DESIGN

4.1 ARCHITECTURE

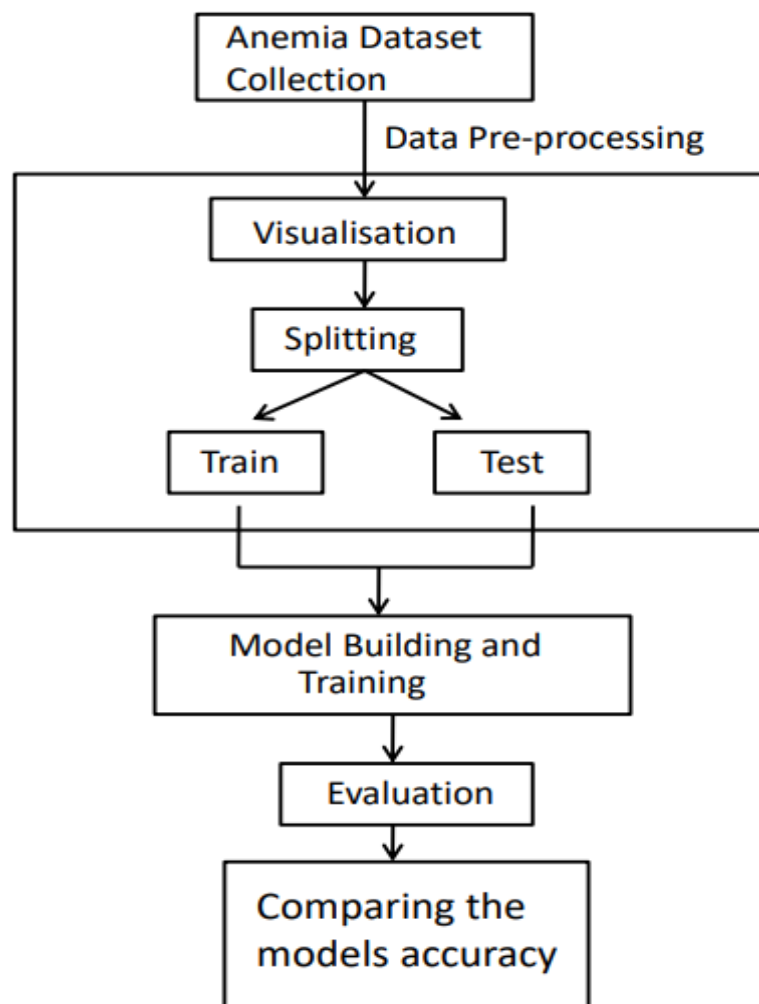


Fig 4.1. architecture of the proposed model

4.2 DATAFLOW DIAGRAM

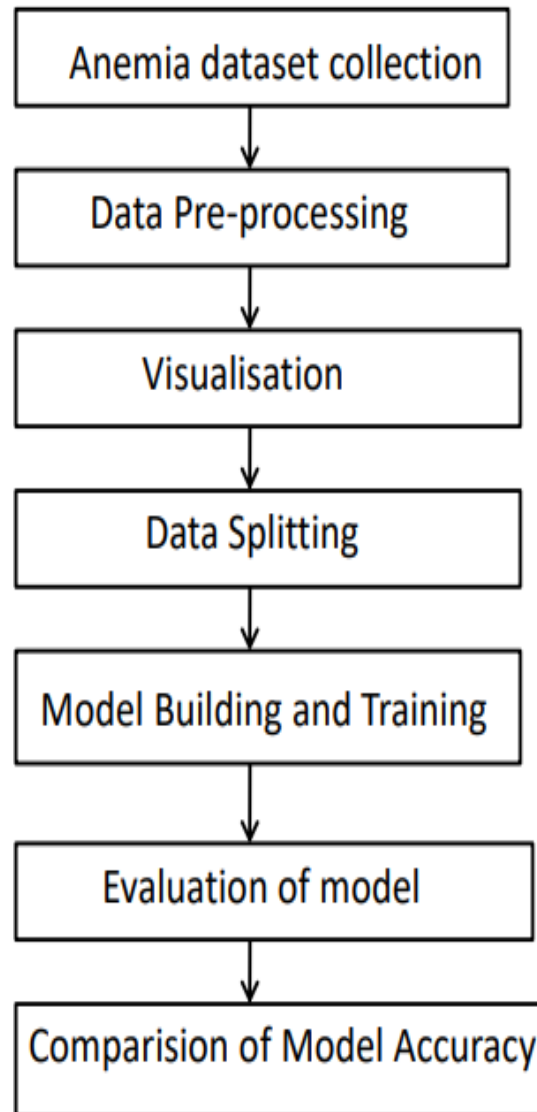


Fig 4.2. Dataflow diagram of proposed system

4.3 USE CASE DIAGRAM

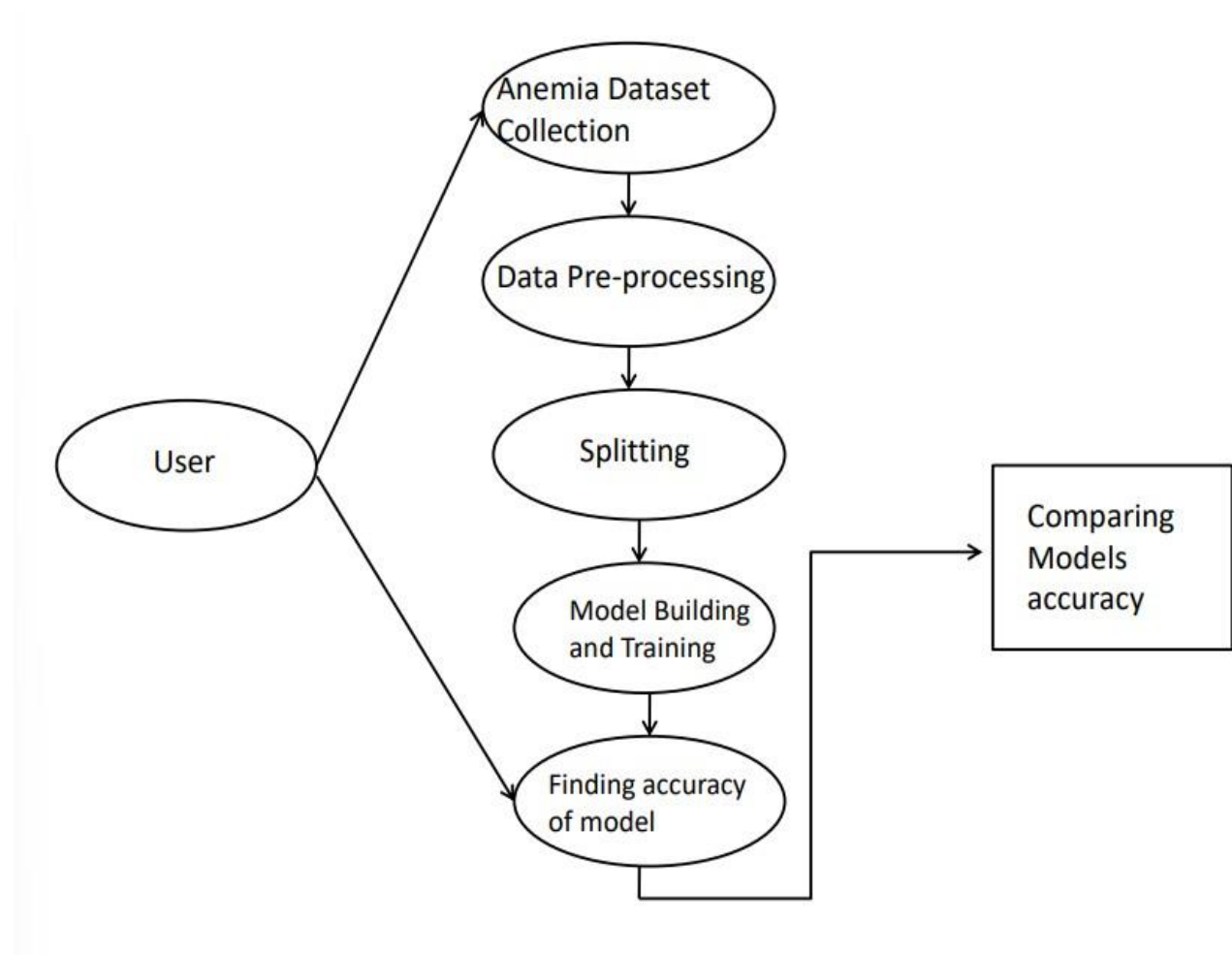


Fig 4.3 use case diagram of proposed system

4.4 SEQUENCE DIAGRAM

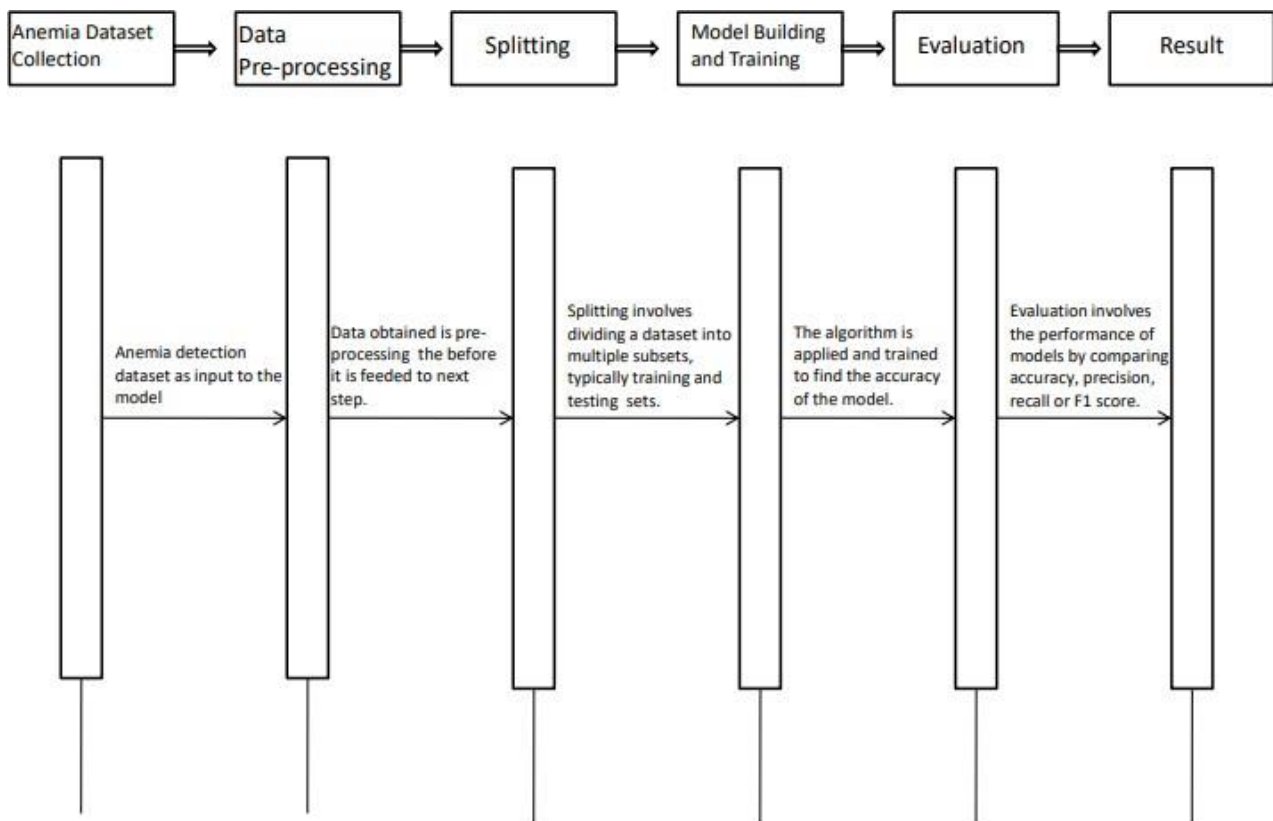


Fig 4.4 Sequence diagram of proposed system

CHAPTER 5

METHODOLOGY

The proposed methodology comprises of following phases:

5.1 Data Collection

The dataset is downloaded from Kaggle. The “Anemia Prediction.csv file” contains anemia quality metrics. The dataset consists of 8545 rows and 6 columns. The 6 columns are divided into Gender, Hemoglobin, MCH, MCHC, MCV and Result.

- MCH- Mean Corpuscular Hemoglobin
- MCHC-Mean Corpuscular Hemoglobin Concentration
- MCV-Mean Corpuscular Volume
- Gender: 0-male, 1-female
- Results:0-not anemic, 1-anemic

5.2 Data Preprocessing:

The data in its normal condition is an unbalanced dataset as the majority of most medical datasets, and as a result, it will affect the prediction models will be used as the models that will be biased towards the majority class only, which here is not anemic (0), and as a result, the predictive Model built won't give a good accuracy in the prediction of the minority class, so with the use of random under sampling we have made a balanced dataset. In addition, the missing values in our dataset. We can use methods like 'isnull ()' or 'isna()' in pandas to identify null values in each column of our DataFrame. This will return a DataFrame of Boolean values indicating the presence of null values.

5.3 Visualization

Visualization is a powerful technique used to represent data visually, allowing patterns, trends, and relationships to be easily understood and interpreted.

There are various types of visualizations, each suited to different types of data and analytical tasks. Common types of visualizations include:

- Bar charts and histograms for displaying frequency distributions or comparing categories. Like for bar chart Hemoglobin vs Result and for histogram for gender or hemoglobin etc.
- Line charts for showing trends or changes over time.
- Scatter plots for visualizing relationships between two continuous variables.

5.4 Data Splitting:

When a machine learning model fits its training data too well and cannot reliably fit fresh data, it is said to be overfit. In order to avoid overfitting, data splitting is widely utilized in machine learning. A machine learning model often divides the initial data into two three. The three sets that are usually utilized are the training set, the validation set, and the testing set. The piece of data used to train the model is known as the training set. In order to improve any of its parameters, the model must observe and learn from the training set. The validation set is a data collection of examples used to alter the settings for the learning process. The objective of this data set is to rate the model's accuracy, which can aid in model selection. The data set that is tested in the final model and contrasted with the earlier data sets is known as the testing set. The testing set serves as an assessment of the chosen algorithm and mode.

Firstly, we must divide the dataset into two parts to make it easier to deal with the data. One part is called X which consists of the 5 parameters of the CBC data (Gender, Hemoglobin, MCH, MCHC, MCV) and the other part is called y and it's our label (Result). Then we can split our data. In this study, the data were split into two parts: train and test with a ratio 80:20

5.5 Feature Selection:

Finding the optimum collection of features that can be utilized to build practical models is the goal of employing feature selection strategies in machine learning. It includes determining each input variable's link to the target variable using a set of evaluation criteria before choosing the input variables with the strongest relationships. Feature selection helps to increase decision accuracy, and shorten the time needed to complete the ML training process.

5.6 Model used:

The classification models used in this project for anemia prediction are:

- **Random Forest (RF)**

Random forest (RF) is a classification method based on "growing" a group of classifiers with a tree-structure. Each classification tree in the forest is used to categorize a new individual using the characteristics of that person. Each growing tree provides a categorization (or "voting") for a class label, and the trees are constructed at random. The choice is based on the votes cast by the majority of the forest's trees. This ensemble approach helps to reduce overfitting and improve generalization performance compared to a single decision tree.

At the core of Random Forest are decision trees, which are hierarchical tree-like structures composed of nodes (representing features), edges (representing decision rules), and leaves (representing class labels or regression values). Decision trees recursively split the data into subsets based on feature values to make predictions.

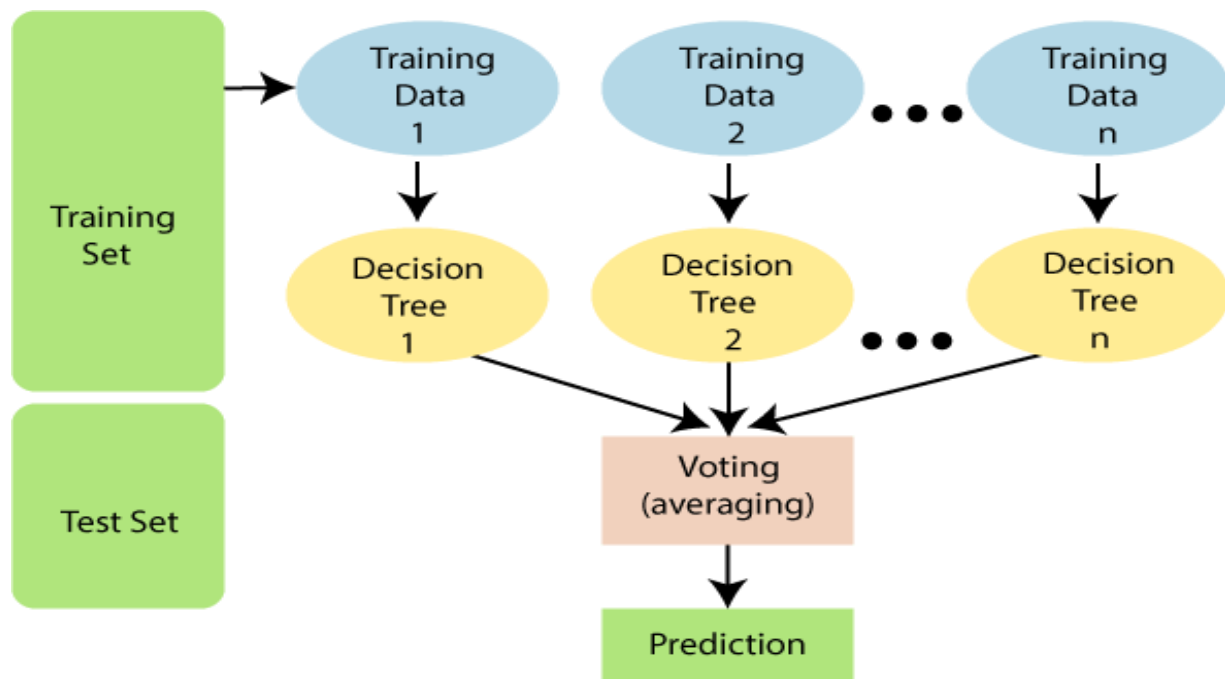


Fig 5.6.2 Random Forest Classifier

ALGORITHM

Step-1 : Select random K data points from the training set.

Step-2 : Build the decision tree associated with the selected data points (Subsets).

$$\hat{y}(x) = \operatorname{argmax}_y \sum_{i=1}^B I(c(x, T_i) = y)$$

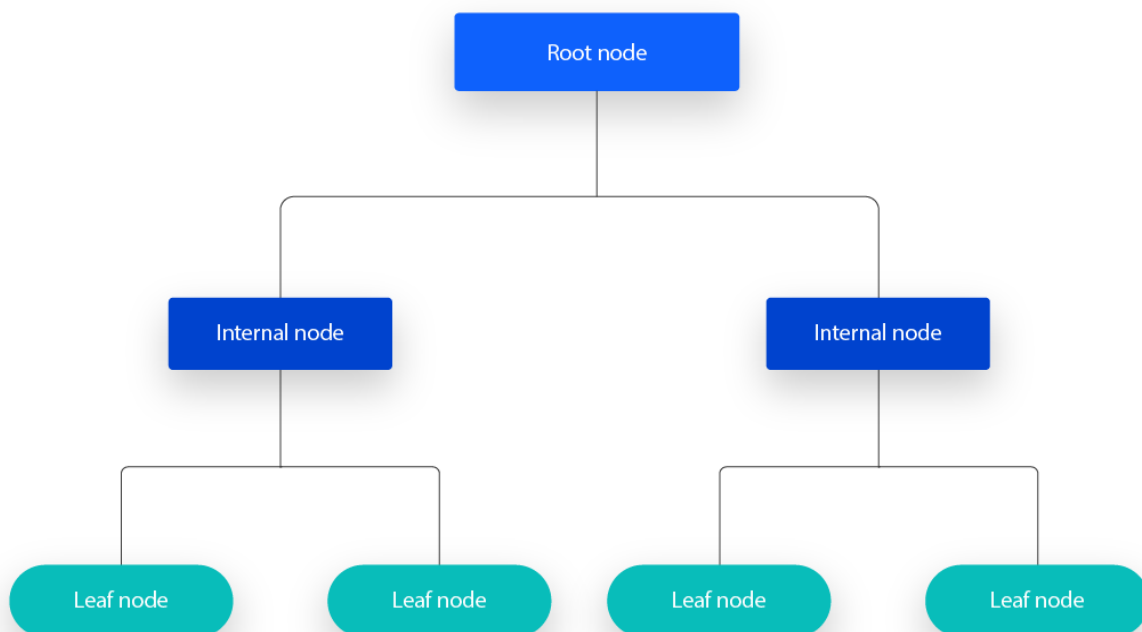
Step-3 : Choose the number N for decision trees that you want to build.

Step-4 : Repeat Step 1 & 2.

Step-5 : For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Decision Tree

Decision tree is one of the supervised learning approaches that is used for classification and regression. It is organized hierarchically, with a root node, internal nodes, and leaf nodes. Each branch node denotes a decision, while each leaf node denotes a choice among several options. The goal of data mining is to uncover as much hidden information as possible; hence, this approach is extensively researched. The subject of medicine is one that still has a lot of unexplored territory.



ALGORITHM

Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

- Calculate the entropy of the target variable.

$$H(Y) = - \sum_{i=1}^n p_i \log_2(p_i)$$

Where P_i is the proportion of data belonging to class i .

- Calculate Information Gain of the target variable.

$$IG(X) = H(Y) - \sum_{i=1}^m \frac{N_i}{N} H(Y|X_i)$$

Where $IG(X)$ is the information gain of feature X ,

N_i is the number of data points in the subset corresponding to the i^{th} value of feature X ,

N is the total number of data points.

Step-3: Select the Best Split with the highest Information Gain. This feature will become the root node of the decision tree.

Step-4: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node. The resulting tree represents a hierarchical structure of decision based on the features of the dataset.

5.7 Model Evaluation

The Predicted values are tested for test data and evaluated.

Accuracy: This is the proportion of correct predictions out of all the predictions made by the model. It is a commonly used metric for binary classification problems like Anemia prediction. However, accuracy can be misleading if the data is imbalanced (i.e., one class is much more prevalent than the other), as a model that always predicts the majority class can have a high accuracy but not be useful.

Mathematically, accuracy is calculated as the ration of the number of correctly predicted instances to the total number of instances:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100 \%$$

Precision:

- Precision, also known as positive predictive value, measures the proportion of true positive predictions (correctly predicted positive instances) out of all instances predicted as positive by the model.
- It focuses on the accuracy of positive predictions and answers the question: "Of all instances predicted as positive, how many are actually positive?"
- Mathematically, precision is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall:

- Recall, also known as sensitivity or true positive rate, measures the proportion of true positive predictions (correctly predicted positive instances) out of all actual positive instances in the dataset.
- It focuses on the model's ability to capture all positive instances and answers the question: "Of all actual positive instances, how many did the model correctly predict as positive?"
- Mathematically, recall is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-score: This is a harmonic mean of precision and recall, and can be useful for imbalanced datasets where both precision and recall need to be considered.

CHAPTER 6

RESULTS

SCREENSHOTS:

First we have to insert the dataset which is present in the local system using pandas library.

```
DataFrame...
      Gender  Hemoglobin   MCH   MCHC   MCV  Result
0         0      15.2    29.9   33.4   89.3       0
1         0      11.9    31.0   32.5   95.4       1
2         0      17.2    31.1   34.3   90.5       0
3         0      12.9    28.0   34.0   82.1       0
4         1      14.5    30.6   33.0   92.8       0
...      ...      ...     ...     ...     ...     ...
8539      0      12.7    27.8   33.3   83.6       0
8540      0      14.9    29.3   32.9   88.9       0
8541      1      12.2    28.9   35.0   82.7       1
8542      1      12.2    29.2   32.8   89.1       1
8543      0      13.7    27.2   33.7   80.7       0

[8544 rows x 6 columns]

Number of rows and column in our Data = (8544, 6)
```

Figure 6.1: Anemia Prediction Dataset

Next about null values of the dataset.

```
Gender      0
Hemoglobin  0
MCH         0
MCHC        0
MCV         0
Result      0
dtype: int64
```

Figure 6.2: Null values

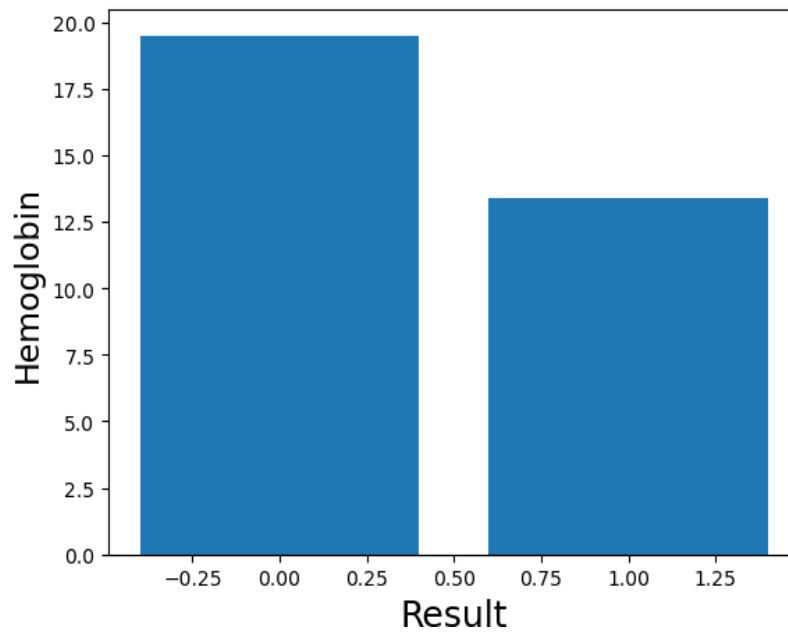


Figure 6.3: plotting Hemoglobin vs Result

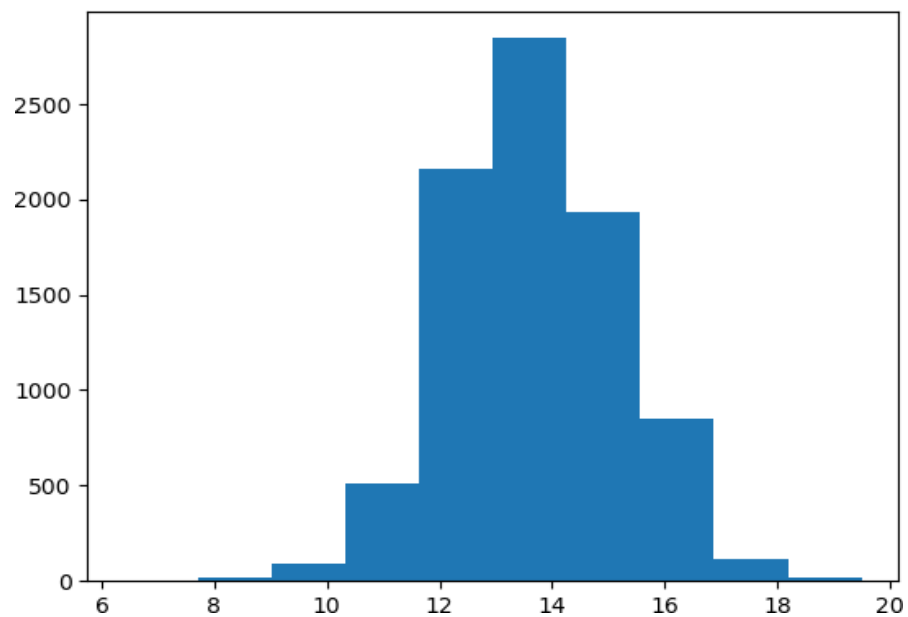


Figure 6.4: plotting Histogram of Hemoglobin

6.5 Random Forest

```
Accuracy: 100.00
Classification Report:
              precision    recall  f1-score   support

     0           1.00        1.00        1.00        1142
     1           1.00        1.00        1.00         567

 accuracy          1.00          1.00          1.00        1709
 macro avg          1.00          1.00          1.00        1709
 weighted avg       1.00          1.00          1.00        1709

[[1142   0]
 [   0  567]]
```

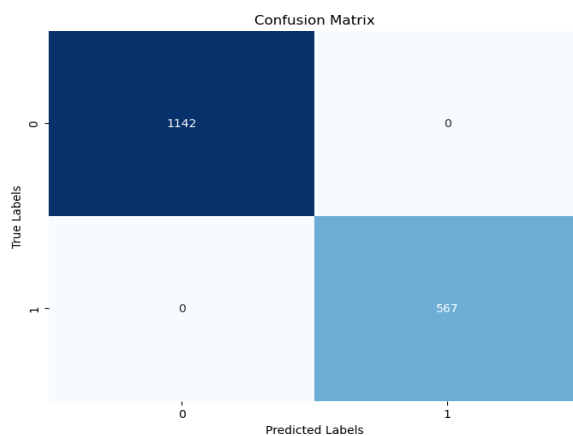


Figure 6.5: Classification Report for Random Forest.

6.6 Decision Tree

```
Classification Report:
              precision    recall  f1-score   support

     0           1.00      1.00      1.00     1142
     1           1.00      1.00      1.00      567

 accuracy          1.00          1.00          1.00     1709
 macro avg          1.00          1.00          1.00     1709
 weighted avg       1.00          1.00          1.00     1709

Accuracy: 100.00
[[1142   0]
 [   0  567]]
```

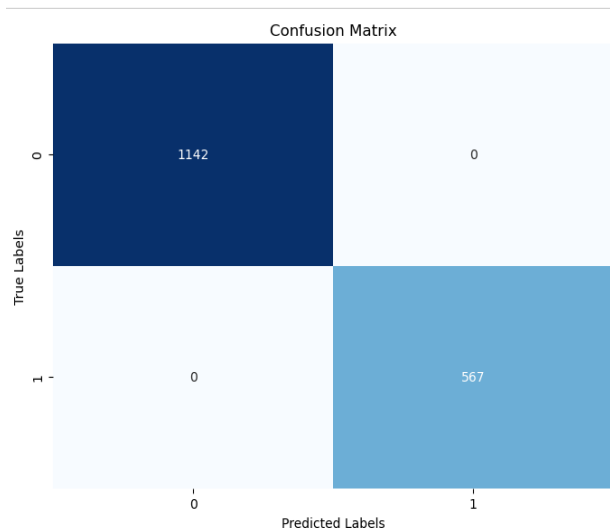


Figure 6.6: Classification Report for Decision Tree

6.7 Accuracy Graph

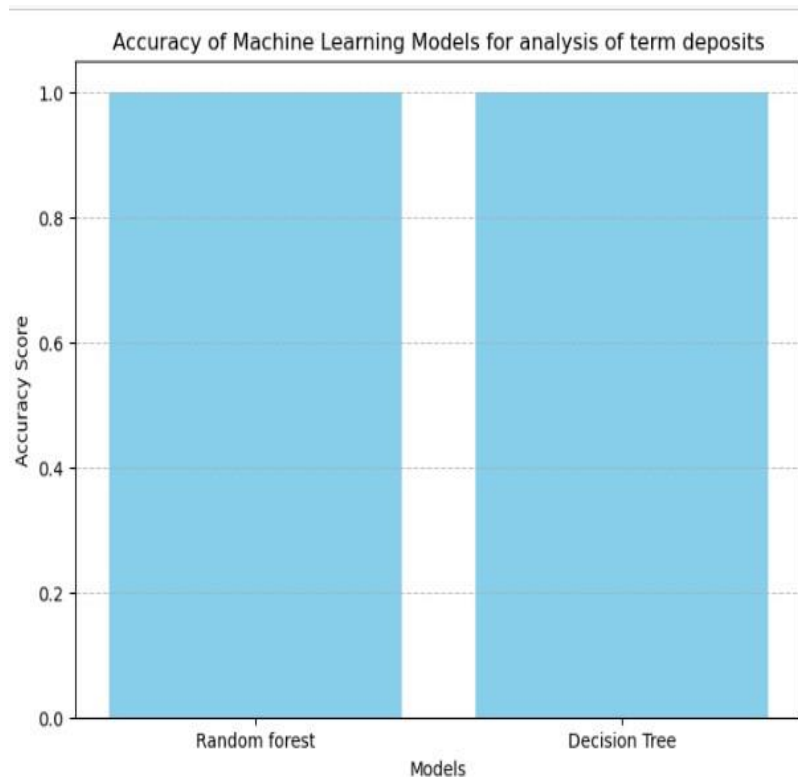


Figure 6.7 Accuracy graph.

CHAPTER 7

CONCLUSION

Anemia prediction can lessen the amount of manual work required for diagnosis. Training and evaluating two classification models, namely Random Forest and Decision Tree classifiers, on the dataset. Both models exhibited high accuracy, precision, recall, and F1-score, indicating their effectiveness in predicting anemia. visualization of model performance through confusion matrices provided insights into the models' ability to correctly classify cases of anemia. After testing two different algorithms, The Random Forest and Decision Tree both the algorithm give the 100% accuracy.

REFERENCES

- [1] Kawo, K.N., Asfaw, Z.G., Yohannes, N.: Multilevel analysis of determinants of anemia prevalence among children aged 6 59 months in Ethiopia: classical and Bayesian approaches. *Anemia* 2018 (2018).
- [2] Feusier, J.E.; Arunachalam, S.; Tashi, T.; Baker, M.J.; Van- Sant-Webb, C.; Ferdig, A.; Welm, B.E.; Rodriguez-Flores, J.L.; Ours, C.; Jorde, L.B.; et al. Large-scale Identification of Clonal Hematopoiesis and Mutations Recurrent in Blood Cancers. *Blood Cancer Discov.* 2021, 2, 226-237. [CrossRef]
- [3] World Health Organization. Hemoglobin concentrations for the diagnosis of anemia and assessment of severity. No. WHO/NMH/NHD/MNM/11.1. World Health Organization, 2011.
- [4] McLean E, Cogswell M, Egli I, Wojdyla D, De Benoist B. Worldwide prevalence of anemia, WHO vitamin and mineral nutrition information system, 1993-2005. *Public health nutrition.* 2009 Apr;12(4):444-54.
- [5] Prevalence of Anemia in Women of Reproductive Age, Our World in Data. Available online: <https://ourworldindata.org/grapher/prevalence-of-anemia-in-women-of-reproductive-age-aged-15-29> (accessed on 28 November 2022)
- [6] Khan, Jahidur Rahman, et al. "Machine learning algorithms to predict the childhood anemia in Bangladesh." *Journal of Data Science* 17.1 (2019): 195-218.
- [7] Obaidy, Midhin AL, et al. "Prevalence and Risk Factors of Anemia among Children Aged 5 months-12 years at Al Anbar Province." *Mosul Journal of Nursing* 9.1 (2021): 131-137.
- [8] Mattiello, Veneranda0, et al. "Diagnosis and management of iron deficiency in children with or without anemia: consensus recommendations of the SPOG Pediatric Hematology Working Group." *European journal of pediatrics* 179 (2020): 527-545.
- [9] Meena, Kanak, et al. "Using classification techniques for statistical analysis of Anemia." *Artificial Intelligence in Medi- cine* 94 (2019): 138-152.
- [10] Mukherjee, K.L., Ghosh, S., 2012. *Medical laboratory Technology. Procedure Manual for Routine Diagnostic Tests. Vol I (Second edition)*, 263-266.
- [11] Lanier, J. Brian, James J. Park, and Robert C. Callahan. "Anemia in older adults." *American family physician* 98.7 (2018): 437-442.
- [12] Jaiswal, Manish, Anima Srivastava, and Tanveer J. Siddiqui. "Machine learning algorithms for anemia disease prediction." *Recent Trends in Communication, Computing, and Electronics: Select Proceedings of IC3E 2018*. Springer Singapore, 2019.
- [13] Verma, Parth, and Vinay Chopra. "A Review on Machine Learning Algorithms for Anemia disease Prediction." (2022).

- [14] Shilpa, S. A., Nagori, M., & Kshirsaga, V. (2011). Classification of anemia using data mining techniques. In *Swarm, evolutionary, and memetic computing* (pp. 113-121). Springer.
- [15] El-kenawy, E.M.T. A Machine Learning Model for Hemoglobin Estimation and Anemia Classification. *Int. J. Compute. Sci. Inf. Secure.* 2019, 17, 100-108.
- [16] Sow, B., et al.: Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques. *Inform. Health Soc. Care* 45(3), 229-241 (2020)
- [17] Amin, N., & Habib, A. (2015). Comparison of different classification techniques using WEKA for hematological data. *American Journal of Engineering Research*, 4(3), 55-61.