EXPLORATION OF MACHINE LEARNING TECHNIQUES IN PREDICTING THE CHILDHOOD ANEMIA

2021
MASTER THESIS
COMPUTER ENGINEERING

QUSAY LUAY SAIHOOD

Thesis Advisor
Assist. Prof. Dr. Emrullah SONUÇ

EXPLORATION OF MACHINE LEARNING TECHNIQUES IN PREDICTING THE CHILDHOOD ANEMIA

Qusay Luay SAIHOOD

T.C.

Karabuk University

Institute of Graduate Programs

Department of Computer Engineering

Prepared as Master Thesis

Thesis Advisor

Assist. Prof. Dr. Emrullah SONUÇ

KARABUK

June 2021

I certify that in my opinion the thesis submitted by Qusay Luay SAIHOOD titled "EXPLORATION OF MACHINE LEARNING TECHNIQUES IN PREDICTING THE CHILDHOOD ANEMIA" is fully adequate in scope and in quality as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Emrullah SONUÇ .....
Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis. June 29,2021

Examining Committee Members (Institutions)	Signature
Chairman : Assist. Prof. Dr. Rafet DURGUT (KBU)	.....
Member : Assist. Prof. Dr. Emrullah SONUÇ (KBU)	.....
Member : Assist. Prof. Dr. Abdullah ELEN (BANU)	.....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Prof. Dr. Hasan SOLMAZ .....
Director of the Institute of Graduate Programs

"I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well."

Qusay Luay SAIHOOD

ABSTRACT

M. Sc. Thesis

EXPLORATION OF MACHINE LEARNING TECHNIQUES IN PREDICTING THE CHILDHOOD ANEMIA

Qusay Luay SAIHOOD

Karabuk University
Institute of Graduate Programs
The Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Emrullah SONUÇ
June 2021, 78 pages

Anemia is the most common disease among children under school age, especially in developing countries, due to a lack of understanding about its causes and preventive measures. In most cases, anemia refers to malnutrition and is closely related to demographic and social factors. Previously, statistical methods were used to predict anemia among children and identify associated factors. It was concluded that this is not a good way. Following the success of machine learning (ML) techniques in exploring knowledge from clinical data in healthcare, it was a good chance to explore the knowledge of social factors associated with childhood anemia. In this study, we compared the performance of eight different ML techniques for predicting anemia in children using social factors to find the most appropriate method. ML techniques achieved promising results in predicting and identifying factors associated with childhood anemia. Multilayer perceptron (MLP) has the best accuracy of 81.67% with all features, while Decision Tree (DT) has the best accuracy of 82.50% when we applied feature selection methods. The explored knowledge of the social factors

associated with anemia can guide nutritional practices and factors essential to child health. Additionally, identified factors can help prevent anemia outbreaks for appropriate intervention by governments and healthcare organizations.

Key Words : Machine learning, classification techniques, anemia in children, iron deficiency, social factors.

Science Code : 92431

ÖZET

Yüksek Lisans Tezi

ÇOCUKLUK ÇAĞI ANEMISINI TAHMIN ETMEDE MAKINE ÖĞRENIMI TEKNIKLERININ ARAŞTIRILMASI

Qusay Luay SAIHOOD

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğrç. Üyesi Emrullah SONUÇ
Haziran 2021, 78 sayfa

Anemi, nedenlerinin ve önleyici tedbirlerin bilinmemesi nedeniyle, özellikle gelişmekte olan ülkelerde okul çağının altındaki çocuklar arasında en sık görülen hastalıklardan birisidir. Anemi birçok vakada yetersiz beslenme nedenli olup demografik ve sosyal faktörlerle de yakından ilişkilidir. Önceleri, çocuklarda anemiyi tahmin etmek ve ilişkili faktörleri belirlemek için istatistiksel yöntemler kullanılmıştır. Fakat bu çalışmalar sonucunda bu yöntemler başarılı olamamıştır. Sağlık hizmetlerinde klinik verilerden elde edilen bilgileri keşfetmede makine öğrenmesi (ML) tekniklerinin başarısını takiben, çocuk anemisi ile ilişkili sosyal faktörlerin bilgisini keşfetmek için bu yöntemlerin faydalı olabileceği yönünde bir fikir oluşmuştur. Bu çalışmada, en uygun yöntemi bulmak için sosyal faktörleri kullanarak çocuklarda anemiyi öngörmek için sekiz farklı ML tekniğinin performansını karşılaştırdık. Bu teknikler, çocukluk çağı anemisi ile ilişkili faktörleri tahmin etme ve tanımlamada umut verici sonuçlar elde etmiştir. Öznitelik seçim yöntemlerini

uyguladığımızda çok katmanlı algılayıcı (MLP) tüm öznitelikler ile %81,67, Karar Ağacı (DT) ise %82,50 doğruluk elde etmiştir. Sonuç olarak anemi ile ilişkili sosyal faktörler, beslenme uygulamalarına ve çocuk sağlığı için gerekli olan faktörlere rehberlik edebilir. Ek olarak, belirlenen faktörler, hükümetler ve sağlık kuruluşları tarafından uygun müdahale için anemi salgınlarının önlenmesine yardımcı olabilir.

Anahtar Kelimeler : Makine öğrenmesi, sınıflandırma teknikleri, çocuklarda anemi, demir eksikliği, sosyal faktörler.

Bilim Kodu : 92431

ACKNOWLEDGMENT

I owe thanks and praise first and foremost to Allah the Almighty for this success and facilitation as I bow to my beloved parents. My dear father, who gave the most precious and valuable things to make me a man of honor. My beloved mother who is good at engineering my heart with her prayers. To my family in which I grow up and its extension that gives me pride and honor. I owe a special thanks to my thesis supervisor, Assist Prof. Dr. Emrullah SONUÇ who spared no effort in providing unlimited advice and guidance until the completion of this thesis to the fullest image.

I also extend my thanks and gratitude to the University of Karabuk, including the wonderful professors, doctors and colleagues who accompanied us throughout our academic journey.

I dedicate this thesis to my beloved country, Iraq. And to the beautiful Turkey, which embraced this scientific experiment and contributed to providing all possibilities to graduate in this distinguished way.

CONTENTS

Page

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

ABBREVIATIONS

AI	:	Artificial Intelligence
ML	:	Machine Learning
DT	:	Decision Tree
SVM	:	Support Vector Machine
RF	:	Random Forest
NB	:	Naïve Bayes
LR	:	Logistic Regression
LDA	:	Linear Discriminant Analysis
KNN	:	K-Nearest Neighbor
MLP	:	Multilayer Perceptron
IDA	:	Iron Deficiency Anemia
RBC	:	Red Blood Cell
WBCs	:	White Blood Cells
CBC	:	Complete Blood Count
RDW	:	Red Blood Cell Distribution Width
2SDs	:	Two Standard Deviations
DHS	:	Demographic and Health Surveys
ALRR	:	ADD-Left Remove-Right
BN	:	Bayesian Network
RP	:	Random Prediction
DNN	:	Deep Neural Network
CKD	:	Chronic Kidney Disease
KNN	:	K-Nearest Neighbor
MLP	:	Multilayer Perceptron
RBF	:	Radial Basis Function
OOB	:	Out-of-bag
ID		Identification

RFE	:	Recursive Feature Elimination
TP	:	True Positives
TN	:	True Negatives
FP	:	False Positives
FN	:	False Negatives
F	:	Frequent

PART 1

INTRODUCTION

1.1. OVERVIEW

Anemia is one of the most common blood diseases worldwide [1]. Anemia is defined as a disordering in which a person's red blood cell count or hemoglobin concentration is lower than the normal range [2]. The fundamental function of hemoglobin is to transport oxygen, and if there is not enough hemoglobin, the blood's ability to transport oxygen to body tissues will be reduced [2]. There are three main mechanisms of anemia: blood loss increased red blood cell breakdown and decreased red blood cell production [3]. Anemia affects people in both developed and developing countries, but it is more widespread in developing countries, where it is estimated that anemia affects 3.5 billion people in these countries [4]. Among the most vulnerable population groups are children under five, adolescents, and pregnant women. Anemia is estimated to affect 42% of children under five and 41.8% of pregnant women worldwide [5,6].

Anemia occurs for many reasons, including nutritional deficiencies such as iron deficiency and lack of vitamins (B12), folic acid, and vitamin (A). In addition, infectious diseases such as human immunodeficiency virus, tuberculosis, and malaria. It can also be caused by hemoglobinopathy and bacterial infections [7].

Iron deficiency anemia is one of the most common types of anemia, especially among children under five years [8]. Several studies indicate the close association between anemia due to iron deficiency and children's social and demographic factors [9]. For example, the study [10] suggests that parents' educational level and socioeconomic status strongly influence the risk of developing anemia in children. Other studies have also confirmed a strong relationship between social and economic factors and the development of anemia in children [11]. Children with iron deficiency anemia may

suffer from many short and long-term health problems, such as low growth index, decreased perception and concentration, mental disorders, and mental retardation [12]. So, early childhood anemia harms academic performance and behavioral development [5]. Also, anemia in children is among the most common causes of high mortality and morbidity rates among children under school age [12]. Moreover, the widespread prevalence of anemia does not only cause health problems but does negatively affect the quality of life. It also leads to significant economic losses for countries that suffer from widespread anemia among children [13].

1.1.1. Anemia in Children and Prediction Techniques

Due to the insidious nature of the disease, people often do not know the condition due to its prevalent symptoms, such as fatigue, headache, paleness of the skin, dizziness, weakness, etc. Still, the disease itself is rampant, and it is a serious and worrying issue. People's ignorance about child nutrition and the importance of social factors in child health are among the main reasons for the high rates of children with anemia at the national and international levels [14]. Therefore, it is crucial to transfer the required knowledge of the causes of anemia to reduce its prevalence [15]. Previously, researchers built many prediction systems by building algorithms based on expert advice. Also, many statistical methods (Chi-square, ANOVA, Logistic regression, and descriptive analysis) have been used to measure the prevalence of anemia [16]. Although such studies may help diagnose the disease, they do not extend to deep data processing, which may detect different relationships and patterns that may be useful in determining the etiology of anemia [15].

Following the success of machine learning algorithms in exploring knowledge from clinical data in healthcare, there was a need to use machine learning techniques to explore social factors associated with childhood anemia [17]. Machine learning is the process of exploring large amounts of data to discover patterns or unknown relationships [18]. Machine learning models can assist in developing models for prediction purposes [19]. These models showed higher performance in solving classification problems compared to the traditional statistical models. Moreover, machine learning is becoming popular in medical and health research, especially

(supervised) classification techniques, among the most critical machine learning techniques [20]. Therefore, there is a need to use machine learning techniques to predict anemia in children and identify the associated factors [15]. This study aims to predict anemia in children and identify the factors using classification techniques; such as Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN) and Multilayer perceptron (MLP). The data used in this study was collected in Iraq for children (600 samples). The data contains correlated factors directly related to the emergence of anemia in children, such as socio-demographic factors, pediatric medical information, child nutrition practice, and mother's nutritional knowledge.

1.1.2. What Is the Definition of Anemia?

Anemia is the state of a deficiency of red blood cells or the hemoglobin content of those cells. So, the blood's capacity to transport oxygen to the body's tissues is reduced if there are few abnormal red blood cells or not enough hemoglobin. Several symptoms appear in a person with anemia, including fatigue, weakness, dizziness, shortness of breath, etc. Several factors affect hemoglobin concentration in the human body, such as age, gender, type of residence, economic level, smoking habits, pregnancy status, etc. [1]. Figure 1.1 shows how the red cells carry oxygen.

Anemia comes in a variety of forms, each with its cause. Anemia can be mild to severe, and it can be temporary or long-term [2].

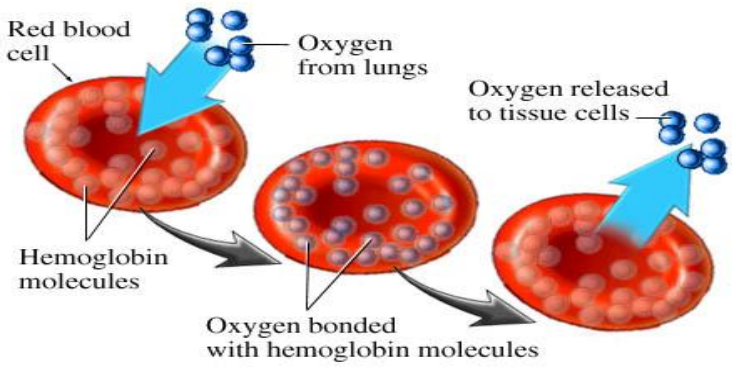

Figure 1.1. Red cells blood carry oxygen [21] .

1.1.3. There Are Several Types of Anemia

- Iron deficiency anemia.
- Thalassemia.
- Vitamin deficiency anemia.
- Aplastic anemia.
- Sickle cell anemia.

This study focuses on the prediction of iron deficiency anemia.

1.1.4. Iron Deficiency Anemia (IDA)

It is one of the most common types of anemia in the world. It is estimated that from 30 to 50% of anemia is caused by iron deficiency, as shown in Figure 1.2. Iron deficiency anemia is widespread, especially in children and pregnant women, leading to 273,000 deaths in the world [13]. Iron deficiency occurs when total or bioavailable iron intake is insufficient to meet iron requirements or compensate for increased losses [22]. When enough iron is in the body, the excess is usually stored as iron storage in the liver or bone marrow for later use by cells [23]. When the required iron overrides the available amount, cells begin to consume the stored iron. So, if buffers are used indefinitely without replenishing the iron supply, the body will eventually become depleted to the point where red blood cell (RBC) formation becomes abnormal. Anemia is the medical term for this condition [24]. Iron is a critical component of hemoglobin, where Fe2 + binds to the protein complex protoporphyrin IX to form hemoglobin. Low blood concentrations and microscopic anemia result from a lack of available iron. Rapid growth, especially during lactation and pregnancy, increases the demand for iron, which explains children's and women's physiological weaknesses [22].

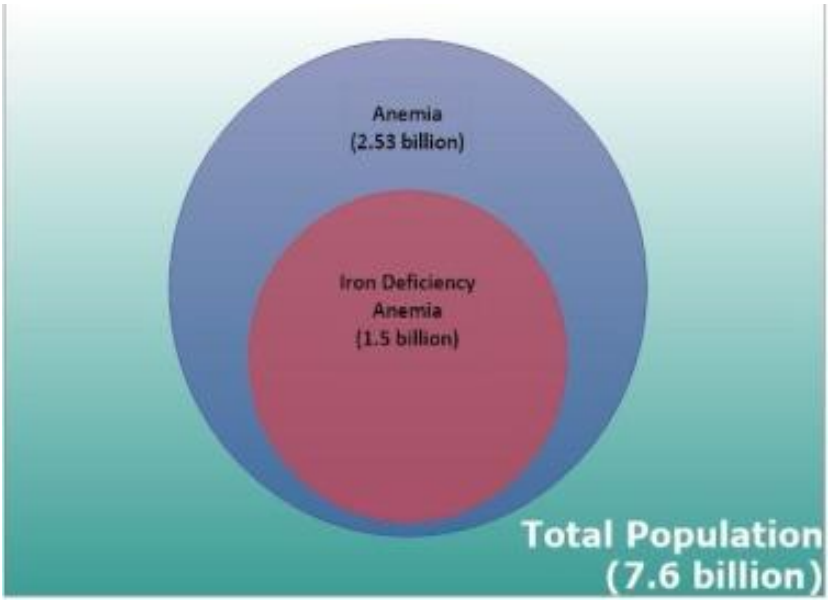

Figure 1.2. Prevalence of anemia and iron deficiency anemia [16].

A complete blood count (CBC) test, including RBCs, white blood cells (WBCs), hematocrit, hemoglobin, and platelets, is usually the first test performed for diagnosis. Lower hemoglobin and hematocrit values are usually enough to diagnose IDA. However, the small size of the erythrocytes can also be considered [16]. Elevated red blood cell distribution width (RDW), which reflects increased variability in red blood cell (RBC) volume, is usually the first indication of IDA with a CBC test [25]. Frequently, the datasets we receive only include hemoglobin level values as a factor, with no other information. However, it is still possible to detect anemia using it as the sole agent, but the error rate is around 5%. Individuals who have two standard deviations (2SDs) under the mean hemoglobin distribution (i.e., healthy range) are at risk for iron deficiency anemia using this method [26]. Table 1.1 shows the mean and two standard deviations below the mean hemoglobin level for infants and children up to six, with the latter being the level at which the child has IDA.

Furthermore, health levels differ significantly in the days following birth and are generally very different for people of all ages. These hemoglobin levels also vary slightly between laboratories that perform these tests. As a result, no generalized value or group can ever be assigned [16].

Table 1.1 Hemoglobin level according to age [16].

Hemoglobin Level	Mean	2 SDS Below Mean
Six Months	(12.6)	(11.1)
Six Months to Two Years	(12.0)	(10.5)
Two Years to Six Years	(12.5)	(11.5)

1.2. PROBLEM STATEMENT

The field of healthcare is significant because it is concerned with human health and life. Hospitals, health facilities, and clinics have enormous data from the clinical records of patients that can be used usefully if appropriate methods are found to explore this data. Relationships and concepts can be extracted from this data. Also, it can be exploited in the early prediction of disease and building decision support systems to help physicians make the right decisions [27].

Also, non-medical external factors such as social factors are decisive factors in many health conditions, especially in children's health. Hence, the need to analyze and understand health and social data to determine the factors causing the spread of disease [15]. For example, anemia is one of the most common diseases among children in developing and developed countries. Childhood anemia is closely related to social factors such as living standards, geographic area, education level, etc. [12].

However, exploring this data requires adequate tools and methods to extract knowledge from this data. In recent years, many statistical methods have been used to measure the prevalence of anemia in children. Still, these methods do not have a solid ability to analyze and predict the disease and find the pathogenesis [16].

Hence, the emergence of the need to use effective machine learning techniques in healthcare to discover hidden information from a patient's history or lifestyle. A deep understanding of the medical and non-medical factors that affect children's health will provide insight into the solution to many children's health problems. Machine learning techniques help predict diseases and determine causal inference from the predictive power of each variable. It includes a method for determining the causal inference

between disease factors known to be variants and disease outcomes [17]. However, this requires a deep knowledge of the disease itself and its associated factors.

1.3. PURPOSE OF THE STUDY

The primary concern of this thesis is the prediction of anemia in children under six years of age from common risk factors. This study analyzes a dataset for a group of children under school age, which contains social, economic, and health factors and nutritional habits associated with the occurrence of this disease. This study improves traditional statistical systems to measure anemia prevalence using artificial intelligence tools such as machine learning techniques and data mining. This study proposes using machine learning algorithms, especially eight classification algorithms, to know the most appropriate method for predicting childhood anemia and identifying the causative factors for the spread of this disease by determining causal inference from the predictive power of each feature. And thus, understanding and identifying the factors affecting the spread of disease to Provide advice to parents, doctors, and health organizations for appropriate intervention to reduce the risk of anemia and know ways to prevent it.

1.4. CONTRIBUTION

This study has eight significant contributions. First, the dataset used in this study was collected by performing a cross-sectional study to obtain the dataset, which consisted of 600 samples and 31 features of risk factors for anemia in children. Second, enter the data obtained into a database. Third, we perform statistical analysis for all the variables in the dataset with the target (anemic, not anemic) to determine the factors that contribute to the spread of the disease and measure the prevalence of anemia for each of the risk factors.

Fourth is implementing three feature selection methods to identify the critical features and know the appropriate feature selection method with machine learning techniques. Fifthly, the main contribution of this study is knowledge discovery in comparing

different machine learning techniques to know the best way to predict anemia in children.

Sixth, predicting anemia by relying on health-related social determinants (variables of socio-demographic, medical information for children, child feeding practice, and nutritional knowledge of the mother). This study assumes that social factors should be considered good predictors for classifying the disease. This is because social determinants of health are critical factors in children's health, leading to endless health problems in developing countries. Seventh, also in this study, anemia is predicted by using only the hemoglobin level because most health professionals and doctors determine anemia in most cases by relying on the hemoglobin level only.

Eighth, this study contributes to identifying the social and health risk factors that cause the spread of anemia among children to provide health instructions for parents to properly care for their children's health. In addition, it contributes to identifying areas that must intervene and improve by local governments and health organizations interested in child health.

1.5. ORGANIZATION OF THESIS

This thesis compares machine learning algorithms' performance to predict anemia in children from risk factors. This thesis consists of six chapters, and the rest of the thesis is organized as follows:

Chapter 2, a literature review deals with predicting and diagnosing anemia and other diseases using AI techniques in this section. In Chapter 3, a brief explanation of Machine Learning, its types, and its techniques. Chapter 4 includes the proposed methodology for the work of this thesis, which is also several parts: data collection, data pre-processing stages, implementation of ML techniques used in this study, and performance measurement metrics. Chapter 5 deals with the results achieved and discuss them. Chapter 6 includes both the conclusion and future work.

PART 2

RELATED WORK

Because of the healthcare sector's high importance, and because it is directly related to human health, there is a constant pursuit of interest in this field and its advancement. Artificial intelligence (AI) techniques, especially machine learning (ML) and data mining, have been the subject of numerous studies in healthcare. Artificial intelligence (AI) techniques have recently achieved great success in exploring knowledge from health data within the healthcare sector. Mainly data mining techniques have been used extensively in diagnosing and predicting diseases.

In this section of this study, a review of the relevant literature was conducted. That includes six investigations about anemia in children, five studies about anemia of all age categories, three studies about heart diseases, one study about chronic kidney disease (CKD), and one study about the nutritional status of children.

Natisha Dukhi et al. [15] proposed using AI methods instead of traditional methods to analyze the prevalence of anemia in children and adolescents in Russia, India, and South Africa. According to this survey, using machine learning is still novel in this domain of medical studies. A machine learning approach to assessing the weight of the cross-linkage between anemia risk indicators will aid policymakers in identifying priority areas for intervention. Traditional statistical methods such as ANOVA, regression, descriptive analysis, and dimensional reduction techniques have been used in most previous studies on the prevalence of anemia in children. But such methods do not include large-scale data processing that could reveal the various relationships and patterns that influence a child's health. As a result, it was proposed to analyze the prevalence of anemia in children using an artificial intelligence approach. So, using AI techniques in this field can attract researchers' and policymakers' attention compared to traditional technologies.

Boubaker Sue et al. [17] focused on exploring social factors that are considered critical in children's health by using machine learning techniques (ANN, SVM, NB, RF, and KNN) to predict anemia and malaria in children from social factors. The dataset used in this study was obtained from Demographic and Health Surveys (DHS) conducted from 2015 to 2016 in Senegal. The machine learning techniques achieved promising results in predicting both anemia and malaria from the same dataset. The ANN technique achieved the best accuracy of 84.17% in predicting anemia and 94.74% in predicting malaria.

Kanak Meena, et al. [16] proposed new data mining methods to replace traditional data analysis methods. They have built a predictive model to predict states of anemia (Not anemia, Mild, Moderate, Severe) in children under the age of five by using data mining techniques (association rule mining and decision tree). The decision tree did not achieve satisfactory results in predicting anemia. Where two decision trees were constructed, the first decision tree, based on the hemoglobin level only, achieved results not realistic with an accuracy of 97%, and the second, based on the relationship between the mother and the child, achieved low results with an accuracy of 44%. But the rules achieved by association rule mining have been considered because each rule has minimum confidence of 0.7, which is desirable.

Priyanka Anand et al. [28] have used five ML techniques (LDA, KNN, RF, DT, and LR) to predict anemia in children under 36 months from the risk factors associated with the occurrence of the disease. The study was conducted on a dataset collected from the outpatient clinics of a hospital in India. RT achieved the highest accuracy of 67.18% among other methods.

Jahidur Rahman Khan et al. [29] also used SVM and the same techniques [28]. ML techniques were tested on the dataset taken by Bangladesh Demographic and Health Survey. RF achieved the highest accuracy of 68.53% among other methods.

Dithy, M. D., and V. KrishnaPriya. [30] proposed using a Gausnominal classification algorithm with a sequential selection process ADD-Left Remove-Right (ALRR) to predict anemia (not anemia, mild, severe, or moderate) in young children and pregnant

women and understand the relationship between iron deficiency and demographic factors. The results showed that the proposed new system performed better with an accuracy of 76.24% than the ANN algorithm, which achieved an accuracy of 65.0%.

Yıldız, et al. [31], using ML techniques (Naïve Bayes (NB), Artificial Neural Networks (ANN), Support Vector Machine (SVM), and Boosted Trees, and Bagged Trees), designed a decision support system to help doctors classify 12 types of anemia. Care was taken to test the system on the same data that the doctor uses, so the system was tested on a data set obtained from a university hospital in Turkey with 25 features. In addition, different feature selection methods were used to create eight different data sets. This system achieved acceptable results for each of the algorithms used in the classification, and the best accuracy of 85.6% was obtained by using Bagged Decision Trees.

Arshed A. AHMAD, et al. [32] Anemia was predicted using four machine learning techniques (Naive Bayes (NB), Bayesian Network (BN), Multilayer Perceptron (MLP), and Logistic Regression (LR)), with and without feature selection techniques. With all the features, the LR algorithm achieved the highest accuracy with 87.3%, followed by MLP, BN, and NB algorithms, achieving 87.1%, 85.1%, and 83.6%, respectively. With the features selected, MLP and LR algorithms achieved the highest accuracy of 86.1%, followed by BN with an accuracy of 85.3% and NB with an accuracy of 84.6%.

Shilpa A. Sanap, et al. [33] predicted anemia using data mining algorithms based on the dataset generated from the CBC test. The C4.5 algorithm obtained the best accuracy of 99.42%, and SVM achieved an accuracy of 88.13%.

Dithy, M. D., and V. KrishnaPriya. [34] used Vect Neighbor classification algorithm with an improved rough set based Fuzzy threshold (RFT) as a feature selection technique to predict anemia in pregnant women (not mild, moderate, and severe anemia). The proposed system achieved better results than the ANN and Gausnominal classifier algorithms in the study [30].

Dithy, M. D., and V. KrishnaPriya. [35] developed a new type of random prediction (RP) with an improvised method for selecting median-based features for predicting anemia in pregnant women. This new system showed better results than ANN algorithms and Gausnominal in study [30] and VectNeighbour in Study [34].

Cardiovascular disease is the leading cause of death worldwide due to the lack of health awareness and the early recognition of the disease. Furthermore, early detection of cardiovascular disease helps in the ability to treat and the speed of recovery from it. Based on these causes, Ashok Kumar Dwivedi, [36] diagnosed heart diseases using six classic techniques (SVM, KNN, ANN, Naïve Bayes, Classification trees, Logistic regression) of machine learning. ML techniques' performance was evaluated on a dataset from the StatLog available in the UCI machine learning laboratory, where the logistic regression technique achieved the highest accuracy of 85%.

D. Panda, et al. [37] compared seven classification techniques (SVM, Gaussian Naive Bayes, DT, RF, KNN, LR, and Ensemble Classification (Extra Trees)) to find the most efficient technique for predicting heart disease. When the classification techniques tested on the data set with five output values (0, 1, 2, 3, and 4), the RF classifier achieved the highest accuracy of 63.33%. In contrast, after reducing the values of the output classes to two values (0.1), the Gaussian Naive Bayes classifier achieved the highest accuracy of 91.66%. This work has observed that classification techniques perform better when the output (target) categories are lower.

Safial Islam Ayon, et al. [27] ML Techniques were used to predict coronary artery heart disease. On two data sets, seven algorithms were compared: SVM, LR, Deep Neural Network (DNN), Nave Bayes (NB), DT, KNN, and RF. The DNN Technique achieved higher accuracy than the other techniques when applied to the Statlog dataset with an accuracy of 98.15%. While being the performance of the SVM Technique was relatively better with the Cleveland dataset with 97.36% accuracy.

Veenita Kunwar, et al. [38] diagnosed CKD from a dataset of 25 features, including age, high blood pressure, diabetes, anemia, etc. Using ANN and Naive Bayes data mining algorithms, the Naive Bayes algorithm performed better at predicting disease

with an accuracy of 100%. In contrast, the ANN algorithm achieved an accuracy of 72.73%.

Zenebe Markos, et al. [39] Using exploration techniques (J48 decision tree, Nave Bayes, and PART classifiers), a method was proposed to predict the nutritional status of children under the age of five and to know the related social and health factors with malnutrition. The 2011 EDHS data set was used. Depending on 17 different features such as the mother's age, region, residence, occupation of mother, education of mother, size of the child at birth and state of anemia of the child, etc. The results were satisfactory as the PART rule induction algorithm achieved the highest accuracy of 92.6%. This study proved data mining techniques could support Predicting the undernutrition status of children. Table 2.1 shows details of all related works that were remembered in this study.

Table 2.1 Related literature review.

Nu.	Methods	Disease /Target	Reference
1	Artificial Intelligence Approach	Anime in children	Natisha Dukhi et al. [15]
2	ANN, SVM, NB, RF and KNN	Anime in children	Boubacar Sow, et al. [17]
3	DT and association rule mining	Anime in children	Kanak Meena, et al. [16]
4	RF, LDA, KNN, DT and LR	Anime in children	Priyanka Anand, et al. [28]
5	RF, LDA, KNN, DT, LR and SVM	Anime in children	Jahidur Rahman Khan, et al. [29]
6	Gausnominal classification.	Anime in children and pregnant women	Dithy, M. D., and V. KrishnaPriya. [30]
7	NB, SVM, ANN and DT	Anemia	Yıldız, et al. [31]
8	NB, BN, MLP and LR.	Anemia	Arshed A. AHMAD, et al. [32]
9	DT and SVM	Anemia	Shilpa A. Sanap, et al. [33]
10	VectNeighbour.	Anime pregnant women	Dithy, M. D., and V. KrishnaPriya. [34]
11	Random Prediction (RP)	Anime pregnant women	Dithy, M. D., and V. KrishnaPriya. [35]
12	SVM, KNN, ANN, NB, DT and LR	Heart diseases	Ashok Kumar Dwivedi, [36]
13	NB, SVM, DT, RF, KNN, LR and Ensemble Classification	Heart diseases	D. Panda, et al. [37]
14	SVM, LR, DNN, NB, DT, K-NN, and RF	Coronary artery heart disease	Safial Islam Ayon, et al. [27]

15	ANN and NB	Chronic kidney disease (CKD)	Veenita Kunwar, et al. [38]
16	J48, NB and PART classifiers	Nutritional status	Zenebe Markos, et al. [39]

PART 3

MACHINE LEARNING TECHNIQUES

ML is one of the essential branches of artificial intelligence concerned with designing algorithms that allow computers to learn independently. ML techniques do not need to program every rule to take a decision or extract a specific pattern. And that, by training it on many data sets that help it understand its concept and its construction. That is, the algorithms are trained on their own [37]. ML is described as the ability of computers to generate accurate predictions based on previous experiences. With the rapid development in computer storage space and processing power, ML has seen significant progress in recent years. ML has numerous advantages, including the ability to determine relationships among large amounts of data. Also, the processing of image-based data with ease, assisting experts with difficult decisions. In addition, it can make for rapid processing of large amounts of data that would be impossible for the human brain to do on short notice [40].ML techniques are widely used in healthcare and a variety of other fields. ML approaches have been developed for the healthcare area due to the complexities and costs of clinical data analysis [41]. When development costs and time are the primary concerns, or when the problem appears too complex to be studied in its entirety, ML can be a viable alternative to traditional methods [42].

There are several types of ML, but the three main types of ML are (supervised learning, unsupervised learning, and reinforcement learning) as shown in Figure 3.1.

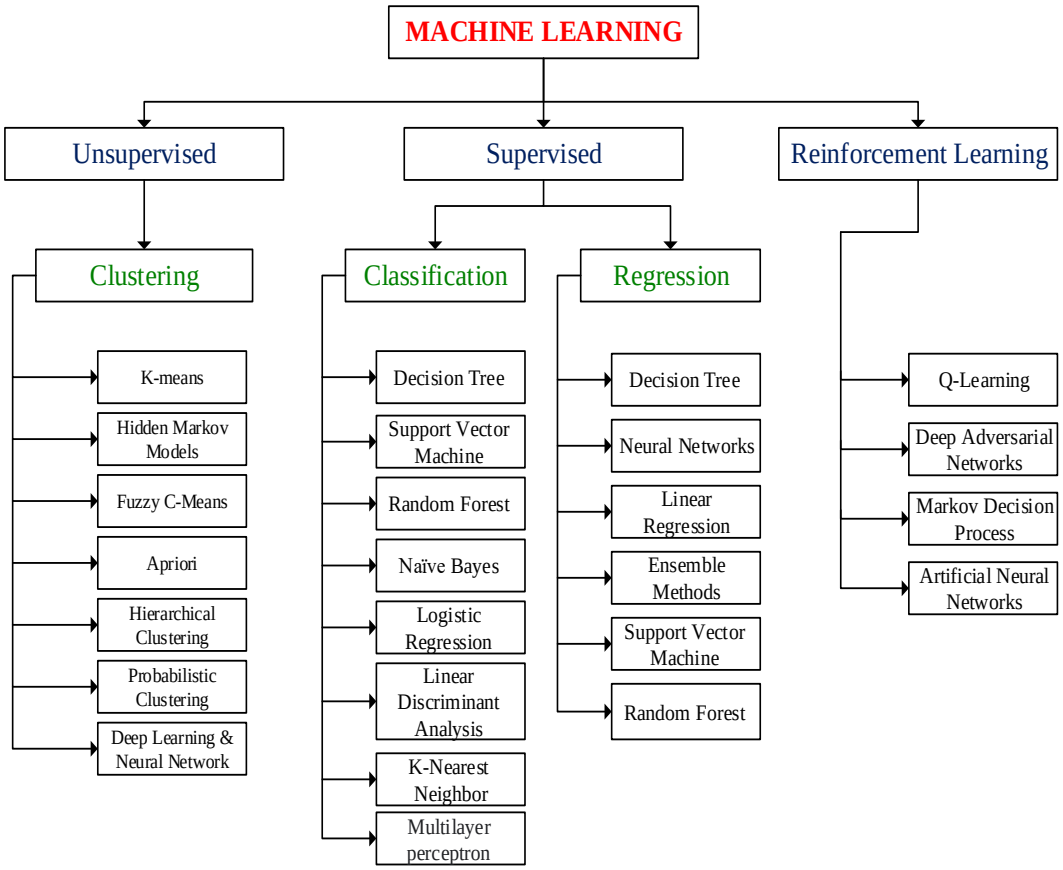

Figure 3.1. Three mine types of ML techniques.

The first type of learning for machine learning algorithms is supervised learning, in which the algorithms are trained on data with predetermined outputs, which means the algorithm is trained on inputs (features) and outputs (target) so that it can later predict new data [37].

As for the second type, unsupervised learning, the algorithms are trained by giving them data without specifying the outputs (target). Through the training process, algorithms build relationships and patterns used to make predictions on new data [37].

The third type, Reinforcement Learning, is distinct from the other two kinds of learning we discussed previously. In this paradigm, an agent explores its environment to achieve a goal. It makes some decisions while exploring its surroundings. Agent receives a positive reward if its decision helps it get closer to his purpose; otherwise, it gets a negative compensation. To put it another way, this can be viewed as a trial-and-error approach [43].

Some studies have also indicated that there are other types of machine learning, such as semi-supervised and reinforcement learning, in addition to supervised and unsupervised learning [43].

In this study, one of the types of supervised learning was focused on classification to predict anemia in children.

3.1. CLASSIFICATION TECHNIQUES

One of the most important and widely used supervised machine learning techniques is classification. There are two types of classification. The first type is used to classify two categories: prediction for, is a child has anemia? So, the prediction is (Anemic or Not Anemic), and this type classified based on two classes is called Binary classification. The second type is the multiple classifications, and in this type, more than two outputs are predicted (Not Anemic, Mild, Moderate, and Severe) [44].

3.1.1. Decision Tree (DT) Technique

It is one of the most critical supervised machine learning algorithms used for both classification and regression problems. The DT looks like a tree-like flowchart where the data is divided continuously depending on a specific parameter, as shown in Figure 3.2. The tree consists of two entities, the decision nodes, to testing a particular feature, and papers that refer to the result of this test, also called the node at the top of the decision tree "root" [45]. Several types of decision tree algorithms differ in the mathematical form of separating training data, as an example (ID3, C4.5, C5, CART (Classification Regression).

Since we are using the scikit-Learn framework, the algorithm used in this study is an improved version of the CART algorithm [46].

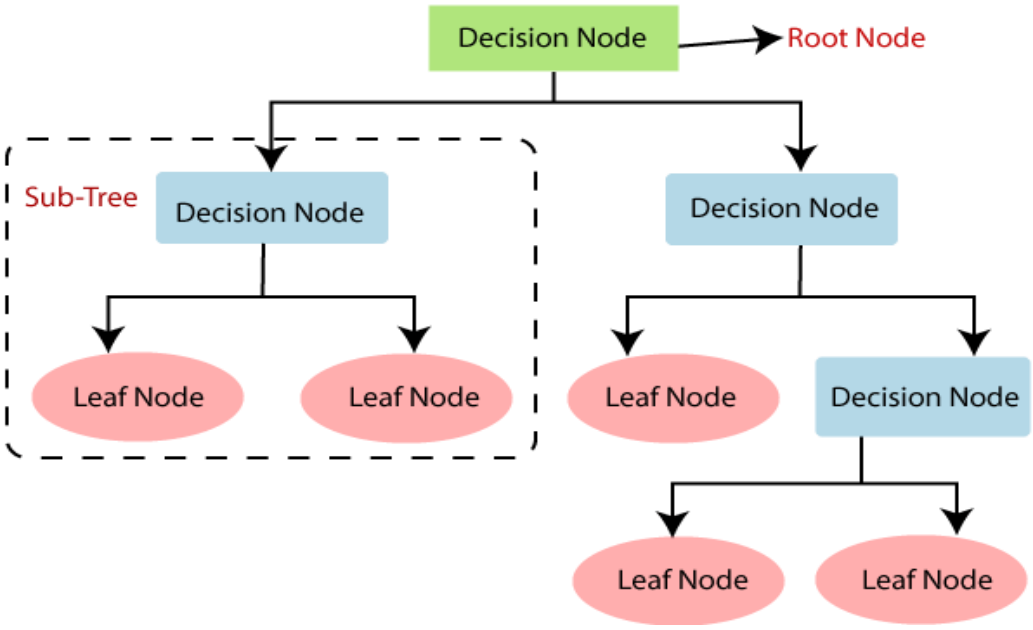

Figure 3.2. The structure of the DT technique [47].

The CART algorithm is used for both classification and regression problems. Depending on the type of dependable variable, its algorithm generates one of these two trees. If the variable is categorical, a classification tree is created; otherwise, a regression tree is made if the variable is numerical. The proposed model's target variable is anemia state (anemic or not anemic), which indicates a categorical variable. Thus, the trees generated in our work are classification trees [16].

The CART algorithm works as follows: The creation of a CART Algorithm is a top-down technique, where the split begins at the top node and progresses through each step, with metrics used to select the optimum partition. For different DT algorithms, several metrics such as Gini impurity, Information gain, Variance reduction, and so on are used. The Gini Impurity, denoted by the letter G, is used by the CART algorithm. It counts the number of times an element in a subset is labeled erroneously (for example, if a child suffering from mild anemia is labeled as a not anemic child). This labeling is a haphazard process that considers the label's dispersion [48]. It can be assessed using the following formula:

$$G = 1 - \sum_{j=1}^{a} (Pi)^2 \qquad (3.1)$$

where G is metrics of Gini impurity, *j* is a number that ranges from one to a *a*, *Pi* is the part of items classified with class *j*.

3.1.1.1. Advantages of DT technique:

- Simple and easy to understand.
- DT algorithm does not necessitate extensive data preparation.
- DT algorithm does not need an oversized cost to build a tree.
- It works with both numerical and categorical data.
- It works well with binary and multiple predictions.
- The decision tree technique is one of the white box techniques, which means that its operation can be easily understood.
- The algorithm's performance can be tested using statistical metrics.

3.1.1.2. Disadvantages of DT technique:

- One of the most common problems of the DT technique is overfitting. Some methods are used to reduce this problem, such as pruning, determining the minimum number of specimens required in a leaf node, and determining the depth of the tree.
- Outliers can create an unstable decision tree. Using decision trees within an ensemble helps solve this problem.
- Decision tree predictions are piecewise constant approximations, not smooth or continuous predictions.
- Some concepts are not easily expressed by DT, such as XOR and equivalence problems.
- If the categories within the data set are not well balanced, it may result in biased trees.

3.1.2. Support Vector Machine (SVM) Technique

It was developed in the mid-1990s by Vapnik et al. [49], based on statistical learning theory, and is one of the most effective supervised machine learning algorithms. The

Support Vector Machine technique can be used for classification and prediction. Still, it is most widely used for classification because it is one of the most efficient machine learning classification techniques [50]. The classification is done by linearly or non-linearly dividing the dataset's input space [31]. It is done by defining the hyperplane in a vector space of N dimensions, distinguishing between two classes of elements, as shown in Figure 3.3. There may be more than one hyperplane separating two classes; in this case, the hyperplane with the largest margin distance is chosen because large margins improve test sample prediction. Support vectors are the points closest to the decision boundary, and these vectors influence the decision boundary's position [49].

This equation determines the hyperplane (H):

$$(H): W * Xi + b = 0 \quad (3.2)$$

where (H) *is* hyperplane, W is a set of weights, b is bias, Xi is input sample features.

The kernel function is one of the most critical parameters for classifier success in the SVM method. There are three main types of Kerner of the SVM algorithm.
Linear kernel: The dot product of two vectors is used to express the linear kernel function, as this equation:

$$K(A, B) = sum(A * B) \quad (3.3)$$

Polynomial kernel: It is defined so that nonlinear and curved input spaces are distinguished. X is the degree of the polynomial. The polynomial kernel is calculated by using this equation:

$$K(A, B) = 1 + sum(A * B) \text{ ^} X \quad (3.4)$$

Radial Basis Function (RBF) Kernel: It is a kernel function commonly used in SVM classifications. RBF can map an input space into infinitely dimensional space. RBF Kernel is calculated by using this equation:

$$K\ (A, B)\ =\ exp\ (-B||\ A - B||\text{^}2) \tag{3.5}$$

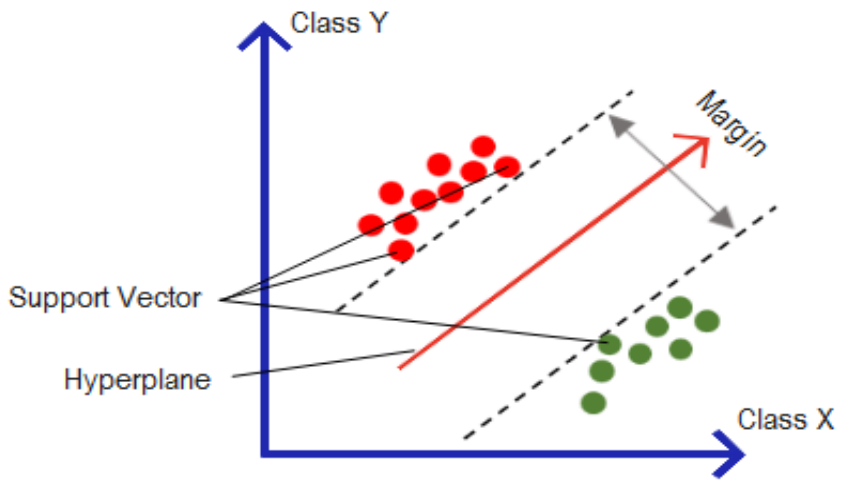

Figure 3.3. Hyperplane, support vector and margin of SVM.

3.1.2.1. Advantages of SVM technique:

- It performs effectively in high-dimensional spaces.
- The approach is still practical when the number of dimensions exceeds the number of samples.
- Because it uses a subset of training points (called support vectors) in the decision function, memory is economical.
- Can be defined various Kernel functions for the decision function, which makes it versatile.

3.1.2.2. Disadvantages of SVM technique:

- If the number of features is substantially more than the number of samples, avoiding over-fitting when choosing Kernel functions and regularization terms is crucial.
- SVM is very time-consuming sometimes because it uses fivefold cross-validation to calculate probability estimates.

3.1.3. Random Forest (RF) Technique

Random Forest was introduced by (Leo, 2001) is a supervised machine learning algorithm that can be used for both classification and regression. Since it grows many decision trees rather than a single decision tree in the model, RF is an ensemble learner. It means more trees which generates a more powerful classifier [51,52]. RF generates many classifiers and aggregates their results to classify a new object, as shown in Figure 3.4. To determine the split, RF will create multiple classification and regression (CART) trees, each trained on a bootstrap sample of the original training data and will search across a randomly selected subset of input variables. CARTs are binary decision trees built by repeatedly splitting data in a node into child nodes, beginning with the root node, which contains the entire learning sample [53]. Each tree in RF will vote for input x, and the majority votes will determine the classifier's output.

The out-of-bag (OOB) error is a crucial feature of RF. For some of the trees, each observation is an out-of-the-box (OOB) observation. As a result, it can be considered an internal validation data set for these trees, as it was not used to build them. The RF's OOB error is simply the average error frequency obtained when the data set's observations are predicted using the OOB trees. The error estimation is less optimistic because of this internal validation, and it is commonly thought to be a good predictor of predicted error for independent data [51].

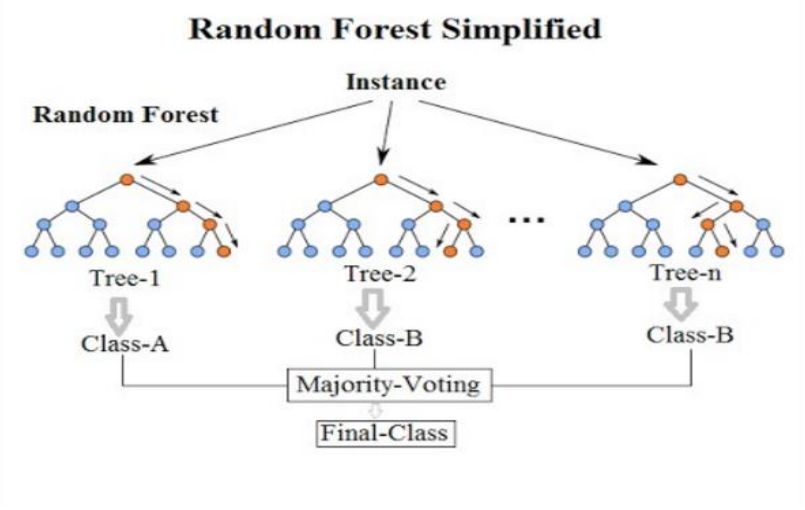

Figure 3.4. The structure of the RF technique [54].

The following steps are used to create the RF:

1. Determined the number of trees in the forest.
2. The number of features subsets to split is determined at each node of the tree.
3. The following standard is used to calculate the number of trees (T):

- A bootstrap with n number of sizes is created, and a random sample Sn is extracted.
- At each node, m attributes are chosen randomly to form a tree, and these randomly selected features are used to find the best split.
- The tree is built to its full potential with no trimming.

4. classifier's output will be determined by the trees' majority votes.

3.1.3.1. Advantages of RF technique:

- Its ability to deal with a large group of data that big dimensions.
- It has a method for estimating missing data that works well and retains precision even though significant data are missing.
- It has an effective method for balancing errors in the dataset where classes are unbalanced.
- It works well with both categorical and numerical features.
- It can estimate the value of all classification features.

3.1.3.2. Disadvantages of RF technique:

- Due to the large number of decision trees combined, RF is more complex.
- It is also more costly when the number of decision trees within the forest is enormous.
- Also, because of its complexity, it takes more time to train the data.

3.1.4. Naïve Bayes (NB) Technique

Naive Bayes methods are a collection of supervised learning algorithms based on Bayes' theorem and the "naive" assumption of conditional independence between every pair of features given the class variable's value [27]. It is instrumental in medical science and text classification problems because a Naive Bayesian model is simple to construct and does not require complicated iterative parameter estimation. Despite its simplicity, the Naive Bayesian classifier frequently performs admirably and is widely used because it often outperforms more complex classification methods.

The Bayes theorem allows you to calculate the posterior probability P (j | i) from P(j), P(i), and P (i | j). According to the Naive Bayes classifier, the effect of the value of a predictor (x) on a given class (c) is independent of the importance of other predictors. Class conditional independence is the term for this assumption [55,56].

$$P(j \mid \mathrm{i}, \ldots, i_n) = \frac{P(j)P(i_1,\ldots,i_n \mid j)}{P(i_1,\ldots,i_n)} \tag{3.6}$$

where $P(j \mid \mathrm{i}, \ldots, i_n)$ is the final probability of conditional probability, $P(j)$ is refer to the prior probability of class, $P(i_1, \ldots, i_n \mid j)$ is the probability of predictor given class, $P(i_1, \ldots, i_n)$ is the prior probability of predictor.

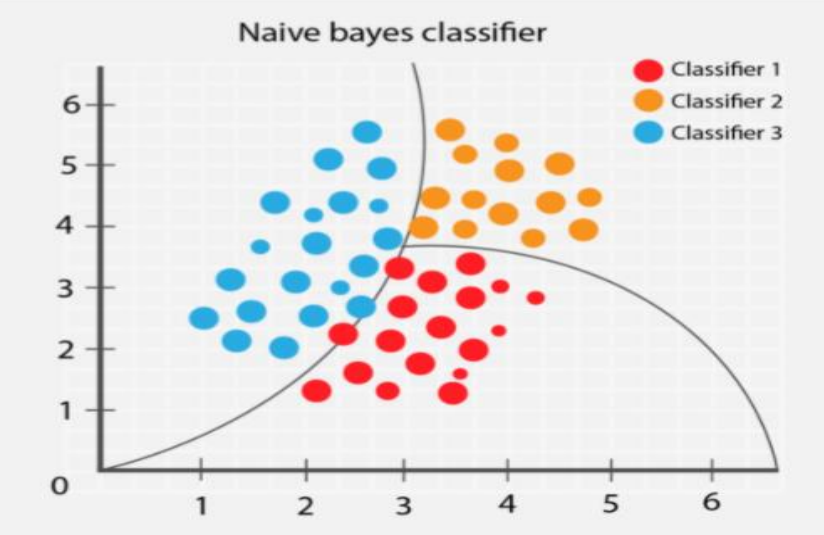

Figure 3.5. Gaussian NB classifier [57].

3.1.4.1. Advantages of NB technique:

- It is relatively easy to understand and build.
- It does not take a long time to implement.
- NB is a simple but fast and accurate method.
- Easy to train even with a small data set.
- It is low cost.
- It can handle a large dataset with ease.

3.1.4.2. Disadvantages of NB technique:

- In practice, obtaining a group of entirely independent predictors is practically impossible.
- The posterior probability is zero when there are no training tuples for a given class. In this instance, the model cannot make any predictions. This issue is referred to as the Zero Probability/Frequency Problem.

3.1.5. Logistic Regression (LR) Technique

LR is a supervised learning classification Technique that uses a statistical model to predict the likelihood of an event by fitting data to a logistic curve, as shown in Figure 3.6. A dichotomous variable is used to assess the outcome. Several expected variables, which can be numerical or categorical, are used in logistic regression. In health care and the social sciences, logistic regression is commonly used. It is also used extensively in marketing to explore consumers' propensity to buy a product or not. The Logistic regression is simply a record of probabilities in favor of a specific event. This function creates an S-shaped curve with probability estimation used to estimate discrete values (0/1 or yes/no) based on a particular set of independent variables [58].

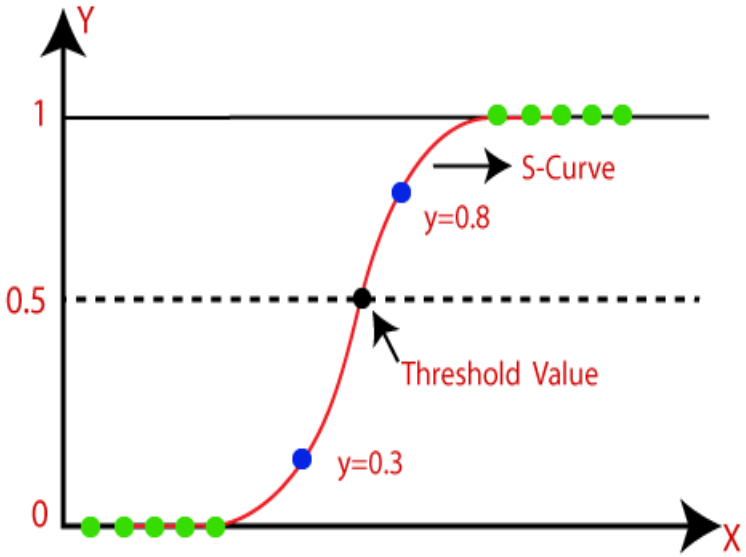

Figure 3.6. Logistic curve of LR technique [59].

Logistic Regression is a multivariable method as well. It tries to establish a functional link between two or more predictor (independent) variables and a single outcome (dependent) variable. Binary LR was used to predict the membership of only two categorical outcomes (Anemic, Not-Anemic) in this study. Although the LR model's primary output is the estimated odds or probability of a binary event, it can also provide additional information that should be used in decision-making. In two-class problems, odds bigger than 50% are classified to a class labeled "1." Other cases are classified with a "0." Although LR is a powerful modeling tool, it assumes that the response variables are linearly related to the predictor variables' coefficients. The coefficients for the independent variables were computed after stepwise incrementing the independent variables [60]. For n independent variables, the general form of the LR functional model is as follows:

$$p(X) = 1/1 + e^{-(B0+B1X1+B2X2+\cdots+BnXn)} = 1/1 + e^{-(b^t * X)} \quad (3.7)$$

Where P(X) is the probability of anemia, *X0, X1, ⋯, Xn* are the predictor variables, *B0, B1, ⋯, Bn* are the regression coefficients, Vector *b*, which for the data set under consideration associates each record (a children) with the probability of anemia, was determined using a binary LR.

3.1.5.1. Advantages of LR technique:

- Because of its simple and efficient nature, it does not require much computing power, is simple to implement, and interpret, and is widely used by data analysts and scientists.
- It also does not necessitate feature scaling. For each observation, logistic regression generates a probability score.

3.1.5.2. Disadvantages of LR technique:

- It is not able to handle many categorical features.
- It is prone to be overfitted.
- Also, LR cannot solve a non-linear problem, so non-linear features must be transformed.
- Independent variables that are not linked to the target variable but are correlated or very similar to each other will not perform well in logistic regression.

3.1.6. Linear Discriminant Analysis (LDA) Technique

LDA is a popular algorithm in supervised and unsupervised machine learning used to identify patterns and solve dimensional problems in data preprocessing steps while building machine learning models [61]. In 1936, R. Fischer suggested linear discriminant analysis (LDA). It involves locating the hyper-projection plane that reduces class variance while increasing the distance between predicted class means, as shown in Figure 3.7. These two objectives can be accomplished by solving the eigenvalue problem with the corresponding eigenvector determining the hyperpigmentation degree, like PCA. This higher level can be used to classify and reduce measurements and clarify the importance of specific features [62,63].
The goal is to locate a linear function.

$$Z = A1\ Xi1 + A2\ Xi2 + \ A3\ Xi3 + \cdots + An\ Xin \tag{3.8}$$

where Z is a discriminant score is a number that is used to predict a case's group membership, *A1, A2, …, An are* coefficients vector that must be determined, *Xi1*, *Xi2, … Xin are* the patients.

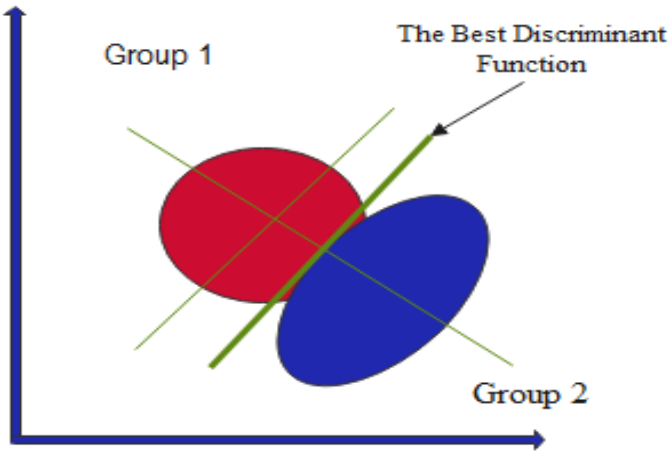

Figure 3.7. The best discriminant function of LDA technique.

3.1.6.1. Advantages of LDA technique:

- Separation of data points linearly.
- Classification of multi-featured data.
- It is discriminating between multiple features of a dataset etc.
- Linear discriminant analysis is a better option than logistic regression whenever multi-category classification is required.
- More stable than logistic regression when the layers are well separated.
- It is also more stable than logistic regression when there are few examples - Logistic regression becomes unstable when there are few cases in which parameters can be calculated. On the other hand, linear discriminant analysis is preferable because it is more stable in such situations.

3.1.6.2. Disadvantages of LDA technique:

- Linear Discriminant Analysis cannot find a new linearly separable axis if the mean values of the distribution are shared between the classes, causing the LDA method to fail, which is one disadvantage of linear discriminant analysis.

- Outliers are extremely sensitive in linear discriminant analysis.

3.1.7. K-Nearest Neighbor (KNN) Technique

The k-nearest Neighbor algorithm (KNN) is a non-parametric classification method first developed by Evelyn Fix and Joseph Hodges in 1951 and later expanded by Thomas Cover [64]. It is one of the simplest algorithms that operate under supervision. It is employed in classification and regression problems. The input in both cases is the k closest training examples in the data set. Whether KNN is used for classification or regression determines the outcome. The output of KNN classification is a class membership. The production of KNN regression is the object's property value [65].

There are two steps to KNN:

- Locate the K training events that are most like the unidentified occurrence.
- Select the most frequently occurring classifications for these K events [27].

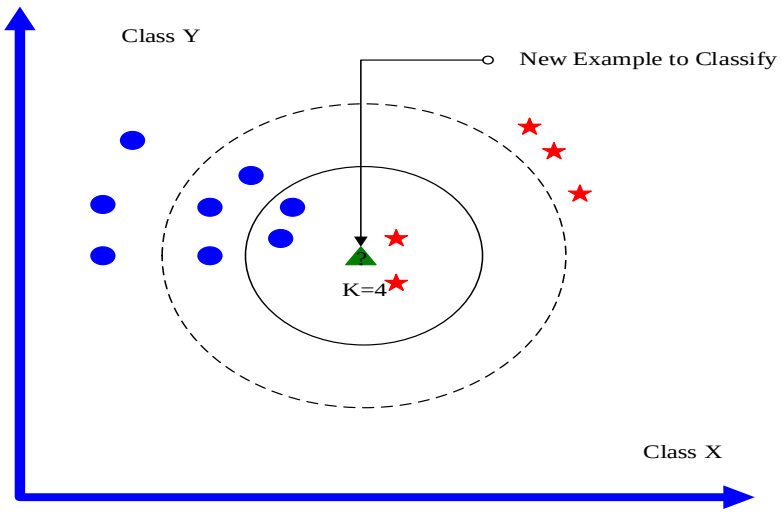

Figure 3.8. The best K of KNN technique.

In this study, the outcomes (target) as categorical, so the KNN algorithm was used as a classification algorithm. Because it makes no assumptions about the possible distributions of the variables used, the KNN classifier is a nonparametric statistic. As

shown in Figure 3.8, where the k symbol indicates the new object's closest neighbors, the new object (target) is classified based on the number of sounds of its neighbors.

For example, if k = 1, then the class of the new object (target) will be categorized by the class of that closest neighbor. There are a lot of ways to choose a suitable k value. However, the simple way is to run the algorithm several times with different values of k, and the best value of k is chosen according to the best result achieved by the KNN algorithm. Because it postpones the generalization decision beyond training examples until a new query is found, KNN is a lazy learning algorithm. It is not necessary to build a model with a training set before using the KNN algorithm. When the input and the k-value are given, KNN uses the training data set to train the inputs and then classifies them. There are two categories, one represented by blue circles and the other by red stars, as shown in Figure 3.8. Each group of these two groups is considered a category (where we can consider that the blue circles represent children who are not anemic, while the red stars represent children with anemia). In the feature space, these classes are represented. The feature can be visualized as a two-dimensional space, such as the data in this study is divided into two categories (anemia and not anemia). They are the x and y coordinates, respectively. If we have three features, however, we will need a 3D space. We also need an N-dimensional space if we have N from space.

If a green triangle represents new data, it must be added to one of the blue circles or red stars set, which is referred to as classification. There is a method for checking the closest neighbors for classification; from the picture, it is a red star; thus, it was added to the red stars category. Because the classification technique is based on the nearest neighbor, this technique is called closest neighbor. However, there is a problem: even though the red star is the closest, what if too many blue circle clusters are nearby? As a result, the blue circles have more power than the red star in this location. As a result, the set of nearest k must be verified. Take, for example, k = 3, which is three samples closer. There is only one blue circle and two red stars, as there is only one blue circle and two red stars. As a result, it should be added to the red stars category once more. If k = 7, however, there are five blue circles and two red stars. It should be added to the blue circle’s category in this case. Accordingly, all changes depend on the value of k.

So, what if k is equal to 4? It indicates that there are two red neighbors and two blue neighbors. Because the classification is based on k-Nearest neighbor, this technique is called k-Nearest neighbor. True, we give importance to k neighbors in the KNN algorithm, but we should give equal weight to all. If you look at the two red stars in the case of k = 4, for example, you will notice that they are closer together than the red circles. Then you should add it to the Red Stars category.

3.1.7.1. Advantages of KNN technique:

- A simple technique to interpret.
- It is a versatile tool that can be used for regression and classification.
- It is high accuracy.
- There is no need to establish more data assumptions, fine-tune multiple parameters, or create a model. It is especially important when dealing with nonlinear data.
- It does not need a long time to implement.

3.1.7.2. Disadvantages of KNN technique:

- It is highly dependent on the quality of the data.
- It may take a long time if the data set is large.
- Sensitive to data volume and irrelevant features.
- All training data must be stored, so it needs a large memory.
- It is computationally expensive because it stores all training instances.

3.1.8. Multilayer Perceptron (MLP) Technique

MLP is a supervised learning technique that connects to a neural network. Unlike other classification techniques like SVM or NB Classifier, MLP Classifier performs the classification task dependent on the underlying neural network [66]. MLP is a type of feedforward artificial neural network (ANN). MLP is ambiguous; it can refer to any feedforward ANN, or it can refer specifically to networks made up of multiple layers of perceptron (with threshold activation). Multilayer perceptron, especially those with

a single hidden layer, is sometimes referred to as "vanilla" neural networks [67]. The multilayer sensory structure of MLP consists of three layers of nodes: the first layer is an input layer, the second layer is a hidden layer, and the third layer is an output layer, as shown in Figure 3.9. For classification using MLP algorithms, the training data is entered into the input layer, and the result is output through the output layer. The number of hidden layers can be increased to improve model performance and obtain higher accuracy [67].

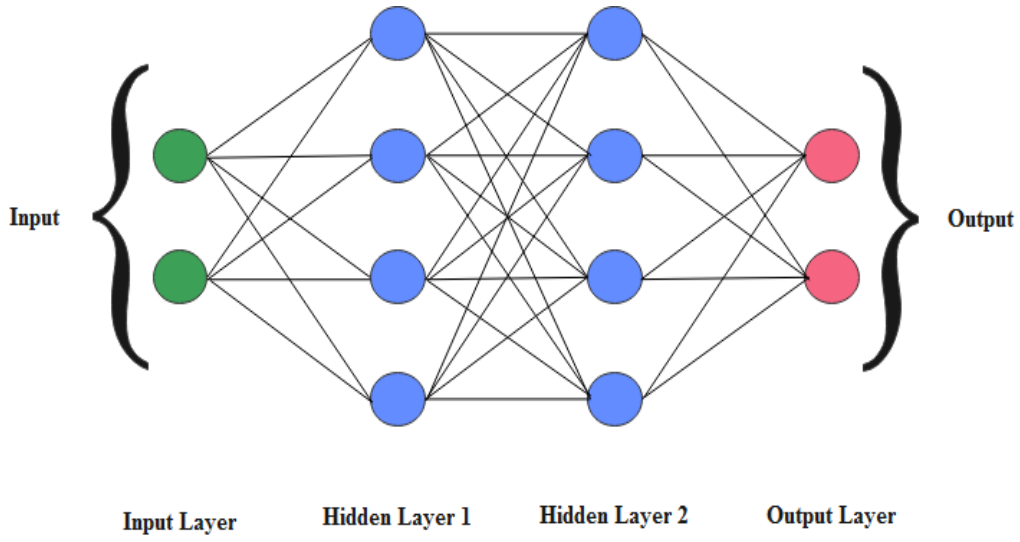

Figure 3.9. Layers of MLP technique.

Multilayer perceptron gets its name from a single classifier called a perceptron (precursor to more enormous neural networks), made up of a single neuron that categorizes the input linearly. As explained by the following equation, the bias is added to the input, a vector multiplied by a certain weight.

$$Y = W * X + B \tag{3.9}$$

where Y is the output, W is the wight, X is the input, X is the bias.

The MLP algorithm addresses the perceptron algorithm's only linear classification limitation and forms more complex functions. The input data is vectorized and fed into the first layer, where it is multiplied by randomly initialized weights, after which some

biases are added. Finally, an activation function is applied to the entire result. The output is passed to the next layer, which repeats the process, with each layer's input data coming from the previous layer except the first. The loss function is calculated after the last layer is reached, as shown in the equation below [68].

$$Loss(\hat{y}, y, W) = -y\ln \hat{y} - (1-y)\ln(1-\hat{y}) + \alpha||W||_2^2 \quad (3.10)$$

where $\alpha||W||_2^2$ is an L2-regularization term, α is a non-negative hyperparameter using to controls the magnitude of the penalty, W is the weights of the input layer and hidden layer, respectively, $\hat{y}$ is actual target for the sample, y is predicted target.

The partial derivative concerning weight in each layer is computed recursively using the calculated error value. The calculated values are updated of the weights, and the process is repeated until the error is as small as possible.

3.1.8.1. Advantages of MLP technique:

- Learning non-linear models is a skill.
- Partial fit can learn models in real-time (online learning).

3.1.8.2. Disadvantages of MLP technique:

- The loss function of an MLP with hidden layers is non-convex, and there are multiple local minimums. As a result, various random weight initializations can result in varying validation accuracy.
- MLP necessitates the adjustment of several hyperparameters, including the number of hidden neurons, layers, and iterations.
- It requires potent data pre-processing, including data scaling and normalization.

PART 4

METHODOLOGY

The proposed system, which is shown in Figure 4.1, consists of three phases.

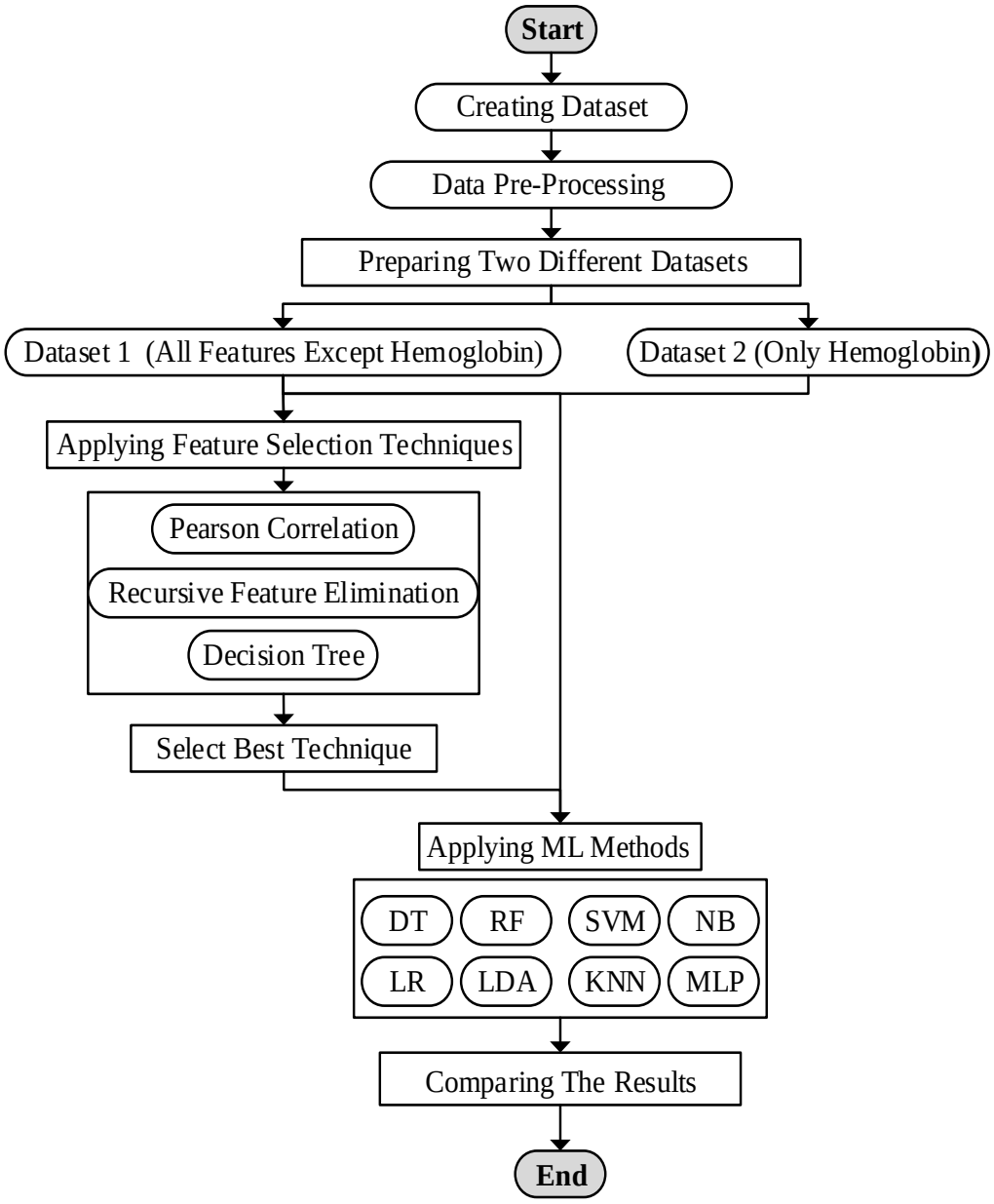

Figure 4.1. Flowchart of the method.

In the first phase, the dataset was used in this study is described. The second phase consists of several steps. First, the data is prepared by pre-processing. Second, the dataset is divided into two subsets. Dataset 1 contains all the features (socio-demographic factors, pediatric medical information, child nutrition practice and mother's nutritional knowledge) without the hemoglobin level feature. Dataset 2 consists of only one feature, which is the hemoglobin level. Third, feature selection is used on Dataset 1. Important features are selected, and features that are not associated with anemia or strongly correlated with another variable are removed. In the third phase, we present classification algorithms to predict anemia and compare the performance of different classifiers by calculating the accuracy of each classifier.

Ethical Authorization: The dataset used in this study was collected by performing a cross-sectional study. It was conducted at Haditha General Hospital and out clinics in Haditha City, Al Anbar Governorate, Iraq, and was overseen by a specialist doctor. This study lasted for three months, from (2020/10/01) to (2021/01/01). The covered population in this study are children from the age of 6 months to under the age of 6 years. The data was collected through a questionnaire containing a set of social factors (socio-demographic factors, pediatric medical information, child nutrition practice, and mother's nutritional knowledge) associated with the occurrence of anemia in children. Written consent was obtained from the child's guardian, the hospital, and the clinic doctors to conduct this study.

Platform Used: The machine learning algorithms used in this study were developed with the Python programming language (version 3.8.3), data pre-processing was done with pandas libraries (version 1.0.5), and machine learning algorithms were developed with scikit libraries (version 0.23.1). The primary platform used for this work is Spyder (Anaconda) on a computer with a 64-bit operating system, x64-based processor, 6.00 GB of RAM, on an Intel (R) Core (TM) i5-3230M CPU @ RAM 2.60GHZ 2.60 GHz.

4.1. DATA COLLECTION

The dataset used in this study has 600 samples collected from Haditha General Hospital and the out clinics in Haditha City in the Iraqi Al Anbar Governorate with a

license from the hospital ethics committee. The features that impact anemia in children were selected according to previous medical studies and the recommendation of an experienced medical professional. All features used in the dataset are explained in Table 4.1.

Table 4.1. Description of dataset features.

Feature Name	Type	Description
1-Id	Numeric	The primary key consists of 600 rows.
2- Child's age	Numeric	From 6 months to under 6 years.
3-Gender	Numeric	0= Female, 1= Male.
4-Mother's age	Numeric	0= Age < 30, 1=Age >= 30.
5-Mother's education level	Numeric	0= Illiteracy,1= Primary, 2= Secondary, 3=University and above.
6-Mother's occupational level	Numeric	0= Un Employee, 1= Employee.
7-Father's education level	Numeric	0= Illiteracy,1= Primary, 2= Secondary, 3=University and above.
8-Father's occupational level	Numeric	0= Un Employee, 1= Employee.
9-Residence	Numeric	0=rural, 1=urban.
10-socio-economic status	Numeric	0= Poor ,1= Moderate,2= Good,
11-Hemoglobin level	Numeric	
12-Short stature for age	Numeric	0= Abnormal ,1=Normal,
13-Fever in the last 15 days	Numeric	0= No ,1= Yes.
14-Previous history of anemia	Numeric	0= No ,1= Yes.
15-Diarrhea in the last 15 days	Numeric	0= No ,1= Yes.
16-Type of breastfeeding	Numeric	0=Abnormal ,1= Normal ,3= Max.
17-Consume milk powder	Numeric	0= No ,1= Yes.
18-Consume sugary drink	Numeric	0= No ,1= Yes.
19-Consume yogurt	Numeric	0= No ,1= Yes.
20-Consume solid/ semisolid food	Numeric	0= No ,1= Yes.
21-Duration of breastfeeding	Numeric	0= Abnormal ,1=Normal,
22-Consumption of meat	Numeric	0= No ,1= Yes.
23- Consumption of dark-green leafy vegetables	Numeric	0= No ,1= Yes.
24-Consumption of foods that are sources of iron	Numeric	0= No ,1= Yes.
25- Consumption of liver	Numeric	0= No ,1= Yes.
26-Know the optimal time of complementary feeding	Numeric	0= Unknow, 1= Know, 2= Not Sure.
27-Know the first complementary food	Numeric	0= Unknow, 1= Know, 2= Not Sure.
28-know the optimum food of supplemental iron	Numeric	0= Unknow, 1= Know, 2= Not Sure.
29-Know nutrients related to anemia	Numeric	0= Unknow, 1= Know, 2= Not Sure.
30-Know the optimal time of breastfeeding	Numeric	0= Unknow, 1= Know, 2= Not Sure.
31-Anemia state	Numeric	0= Not Anemic (171 samples), 1= Anemic (429 samples).

The dataset consists of 31 features, including (the primary key for each child's specific identification). These are: Socio-demographic factors (Child's Age, Gender, Mother's age, Mother's education level, Mother's occupational level, Father's education level, Father's occupational level, Residence, socio-economic status), pediatric medical information (Hemoglobin level, Short stature for age, Fever in the last 15 days, Previous history of anemia, and Diarrhea in the last 15 days), and child nutrition practice (Type of breastfeeding, Consume milk powder, Consume sugary drink, Consume yogurt, Consume solid/ semisolid food, Duration of breastfeeding, Consumption of meat, Consumption of dark-green leafy vegetables, Consumption of foods that are sources of iron and Consumption of liver), And the mother's nutritional knowledge (Know the optimal time of complementary feeding, know the first complementary food, know the optimum food of supplemental iron, know nutrients related to anemia and know the optimal time of breastfeeding). And anemia state (anemic, not anemic) as a target.

4.2. DATA PRE-PROCESSING

Data pre-processing is an essential stage for both data mining and machine learning techniques. Since real-world data tends to be inconsistent, noisy, and may contain missing, redundant, and irrelevant data. It negatively affects the performance of the algorithms and may result in inaccurate knowledge and incorrectly learned. Pre-processing is used to clean data, scaling data, and transform data into a format that matches the algorithms used. In addition, feature selection to select the best features [30]. Figure 4.2 shown Data pre-processing stages using in this study.

4.2.1. Data Cleaning

In this stage of the pre-processing steps, the missing data and the duplicate data are checked. The missing values can be handled in several ways, such as replacing the value with the attribute's median or mean or mode [17]. In this study, the missing value of a feature is replaced with the mean value of that feature column. There were seven residence values, and ten short stature values are missed, which was compensated by

using the mean value of the feature column. In addition, it was ensured that there were no duplicate values.

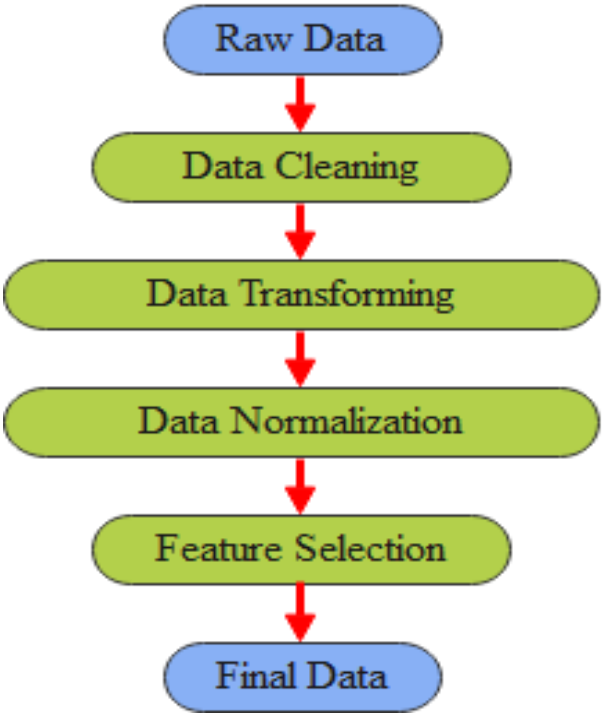

Figure 4.2. Data pre-processing stages.

4.2.2. Data Transforming

In this stage of the pre-processing steps, all categorical features in the dataset are converted to a unified numeric representation. Due to there are categorical and numeric features in the dataset.

4.2.3. Data Normalization.

Normalization converts the values of numeric columns in a dataset to a shared scale without distorting the ranges of values. In this study, data normalization was used with the MLP algorithm only using the Min-Max normalization method [69].

The final dataset after the pre-processing process consists of 600 rows and 31 columns and does not contain any null or duplicate values and is represented in numeric form.

4.2.3.1. Min-Max normalization

Min-max normalization is a linear data transformation that reduces the dataset's range to a single range. Most of the time, attributes are rescaled to fall between -1 and 1 or 0 and 1. In this study, the range of normalized data was between (1,0).
The formula below used to calculate min-max normalization of range between (1,0):

$$Y = a - \min(a) / \max(a) - \min(a) \quad (4.1)$$

Where Y *is* the normalized value, a is an original value.

4.2.4. Feature Selection

Using feature selection techniques in machine learning is to find the best set of features that can be used to construct valuable models. It involves evaluating the relationship between each input variable and the target variable according to specific evaluation criteria and selecting the input variables with the most potent relationship to the target variable. Feature selection is used to improve decision accuracy, minimize the dimensions of the dataset, and minimize the time in implementing the ML training process. There are four main feature selection techniques: filter approach, wrapper approach, embedded approach, and hybrid methods [70].

More than one feature selection technique has been experimented with to choose the best method suiting ML techniques used in this study.

4.2.4.1. Pearson Correlation

Pearson correlation is one of the filter methods to feature selection that uses a statistical method to determine the association between two variables X and Y. It calculates the intensity of the correlation between each function of the input dataset and the category (target). This test's result ranges from 1 to -1, suggesting a weak correlation is indicated by a value near 0. a value indicates a strong positive correlation near 1, and

a value indicates a strong negative correlation near -1 [71]. Below formula used to calculate Pearson correlation:

$$R = \sum(Ai - B)(Ci - D)/\sqrt{\sum(Ai - B)^2 \ \sum(Ci - D)^2} \qquad (4.2)$$

where R is the correlation coefficient, *Ai* is the values of the A-variable in a sample, *B* is the mean of the vB-variable values, Ci is the C-variable values in a sample, *D* is mean of the values of D-variable. Figure 4.3. shown correlation between all features in the dataset.

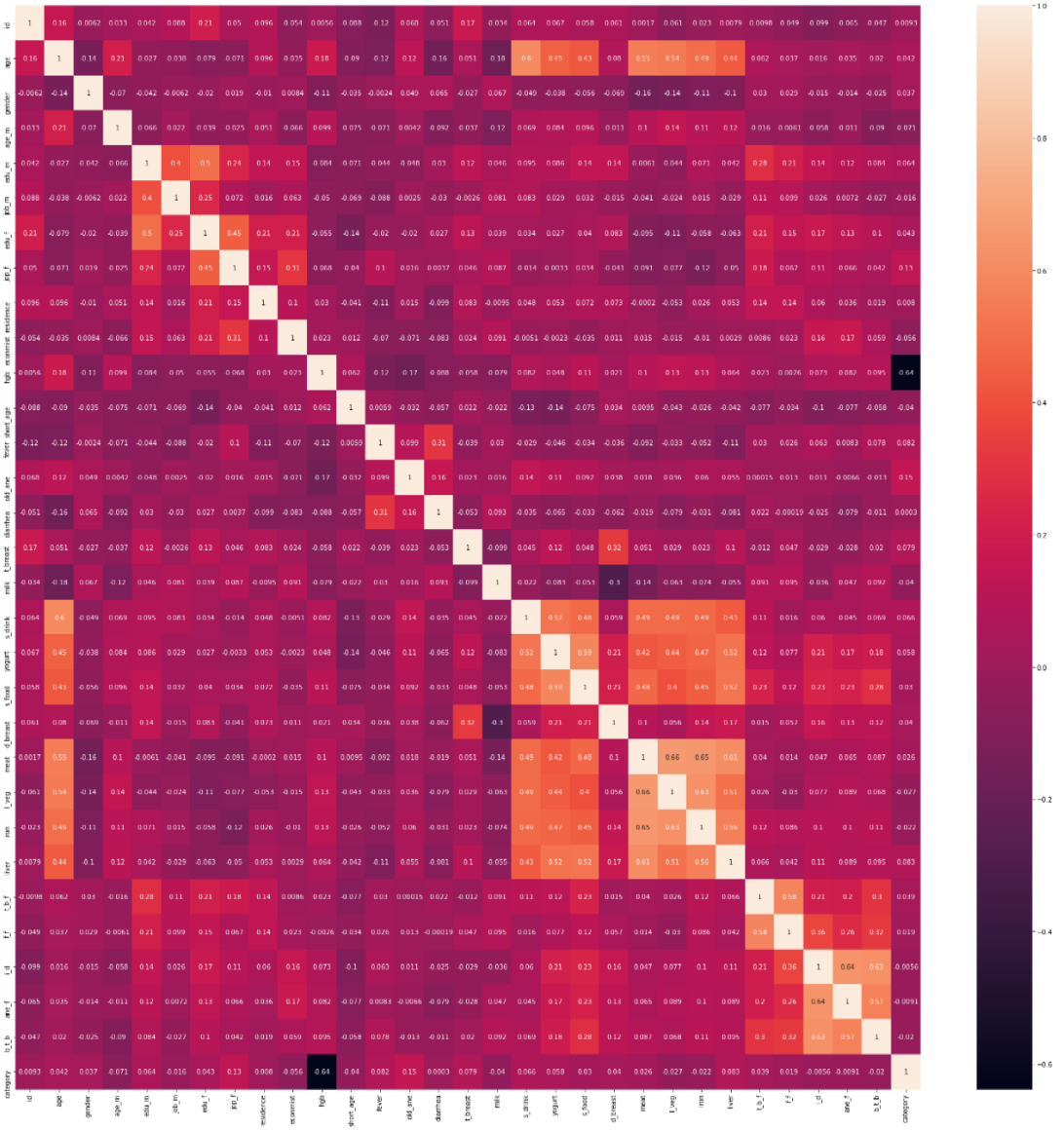

Figure 4.3. Correlation between all feature in dataset.

4.2.4.2. Recursive Feature Elimination

Recursive Feature Elimination (RFE) is a feature selection wrapper method to define features by regularly looking at smaller collections of features. First, the estimator is trained on the initial set of features. Next, the coef_attribute or feature value is used to determine the significance of each feature. The less important features are then trimmed from the original dataset. This process is repeated until the highest accuracy is reached and each function is rated [72].

4.2.4.3. Decision Tree

One of the essential advantages of the decision tree algorithm is the feature importance characteristic that defines each of the input variables according to their importance in determining the class(target). This characteristic is used to identify significant features and be used to select features [73]. Table 4.2 shows the features chosen by all the feature selection techniques used.

Table 4.2. features selected by feature selection techniques.

Feature Selection Techniques	Number of features selected	Features selected
Pearson correlation	8	{mother's age, mother's education level, father's occupational level, fever, previous history of anemia, type of breastfeeding, consume sugary drink, consumption of liver}
Recursive Feature Elimination	14	{child's age, gender, mother's age, mother's education level, father's education level, father's occupational level, fever, previous history of anemia, type of breastfeeding, consumption of foods that are sources of iron, know the optimal time of complementary feeding, know the first complementary food, know the optimum food of supplemental iron, know the optimal time of breastfeeding}
Decision Tree	12	{child's age, gender, mother's education level, father's education level, residence, fever, type of breastfeeding, duration of breastfeeding, know the optimal time of complementary feeding, know the optimum food of supplemental iron, know nutrients related to anemia, know the optimal time of breastfeeding}

4.3. IMPLEMENTING ML ALGORITHMS

In this section was discussed the implementation of all ML techniques used in this study. In addition, some of the parameters of ML techniques were tuned and discussed. The 8 ML techniques are built using a scikit-learn library. It is a powerful library used to implement ML models and pre-processing and model validation phases. In scikit-learn library, there are a group of parameters of ML techniques. These parameters are optional but tuning them improves the performance of the algorithms.

Before implement ML techniques, the Dataset was split into two 80% of the data was used as a training set and 20% used as a test set. Then all ML techniques used in this study were implemented to predict anemia in children in three stages.

In the first stage, ML techniques were implemented on dataset 1. Dataset 1 contains all the features (socio-demographic factors, pediatric medical information, child nutrition practice, and mother's nutritional knowledge) except the hemoglobin level.

In the second stage, ML techniques were implemented on dataset 2. Dataset 2 contains only the hemoglobin level as a feature.

In the third stage, was used feature selection techniques on dataset 1. Then we implement the classification algorithms.

4.3.1. Implementing Decision Tree (DT) Algorithm

Decision tree (CART) algorithm was implemented to predict anemia in children. It is one of the most popular classification algorithms [48]. The DT has a set of optional parameters that are important for the algorithm to work well. For example (max_depth, min_samples_leaf, min_samples_split, etc.), if not set well, can lead to full-grown trees that are likely to be very large and cause overfitting of the algorithm. So, these parameters must be fine-tuned to improve the performance of the decision tree algorithm and reduce overfitting and memory consumption. The DT has been implemented to predict anemia in children in three stages.

In first stage DT was implemented on dataset 1. DT parameters were tuned as {*criterion='gini', splitter='best', max_depth=13, min_samples_split=2, min_samples_leaf=1, random_state=2, max_features='auto', min_weight_fraction_leaf=0.0*}.

where Criterion is a function to measure the quality of a split, *splitter* is a method for selecting the split at each node, *max_depth* is the maximum depth of the tree, *min_samples_split* is a minimum number of samples required to split an internal node, *min_samples_leaf* is a minimum number of samples required to be at a leaf node, *random_state* is using to controls the randomness of the estimator, max_features is number of features to consider when looking for the best split, *min_weight_fraction_leaf* is the smallest weighted fraction of the total number of weights. Figure 4.4. shows the confusion matrix when implementing DT on dataset 1.

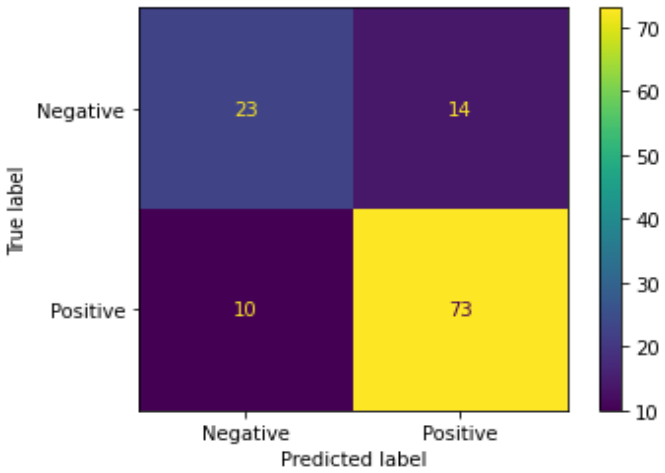

Figure 4.4. Confusion matrix to implement the DT algorithm on dataset 1.

In the second stage, DT was implemented on dataset 2. Again, DT parameters were tuned as the same parameters in stage 1 with a change max depth value of 4. Figure 4.5. shows the confusion matrix when implementing DT on dataset 2.

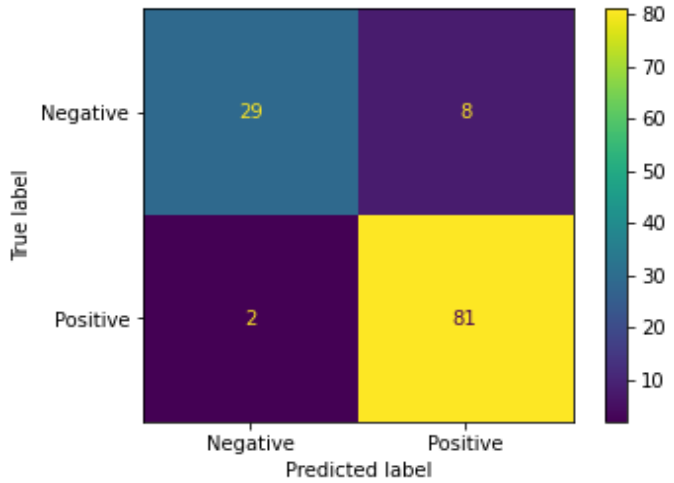

Figure 4.5. Confusion matrix to implement the DT algorithm on dataset 2.

In the third stage, DT was implemented on a sub dataset selected by applying feature selection techniques on dataset 1. DT parameters were tuned as the same parameters in stages 1 && 2 with a change max depth value of 15. Figure 4.6. shows the confusion matrix when implementing DT with feature selection techniques on data set 1.

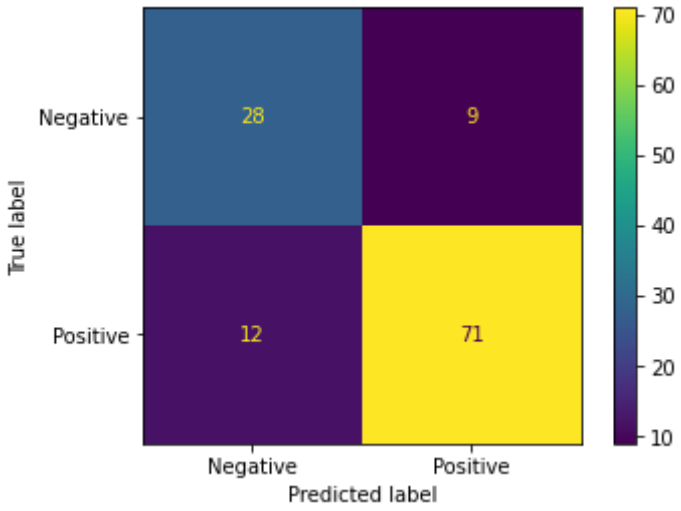

Figure 4.6. Confusion matrix of DT algorithm with features selected.

4.3.2. Implementing Support Vector Machine (SVM) Algorithm

SVM was applied to predict anemia in children. It is one of the most effective classification algorithms. SVM contains a set of optional parameters. The kernel type for the SVM algorithm can be chosen where there are {‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’,

'precomputed'}. The SVM has been implemented to predict anemia in children in three stages.

In first stage SVM was implemented on dataset 1. SVM parameters were tuned as *{**C** =1.0, kernel='poly',degree =3,gamma='scale', coef0=3, shrinking=False, probability=False, decision_function_shape='ovo', random_state=0}*.

where C is a regularization parameter, kernel is using to determine the kernel type of SVM algorithm, *degree* is the degree of the polynomial kernel function, *gamma* is the Kernel coefficient, *coef0* is an independent term in the kernel function, *shrinking* is use the shrinking heuristic or not, *probability* is enable probability estimates or not, *decision_function_shape* is return a one-vs-rest or not, *random_state* is used to control the generation of random numbers to mix data for probability estimates. Figure 4.7. shows the confusion matrix when implementing SVM on dataset 1.

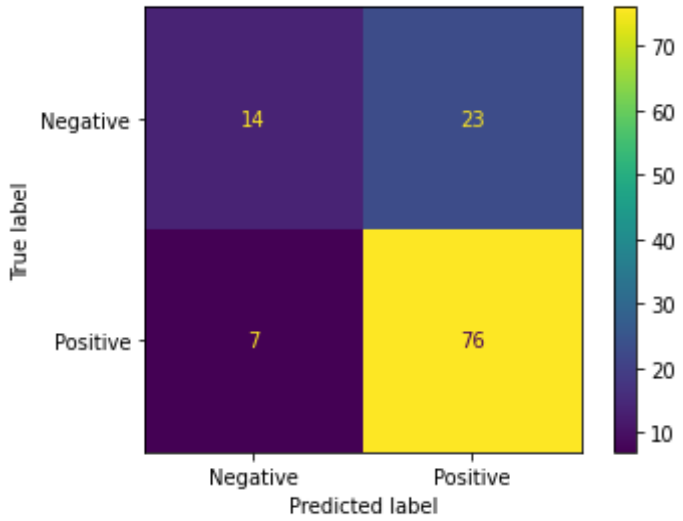

Figure 4.7. Confusion matrix to implement the SVM algorithm on dataset 1.

In the second stage, SVM was implemented on dataset 2. Again, SVM parameters were tuned as the same parameters in stage 1 with a change degree value to be 1. Figure 4.8. shows the confusion matrix when implementing SVM on dataset 2.

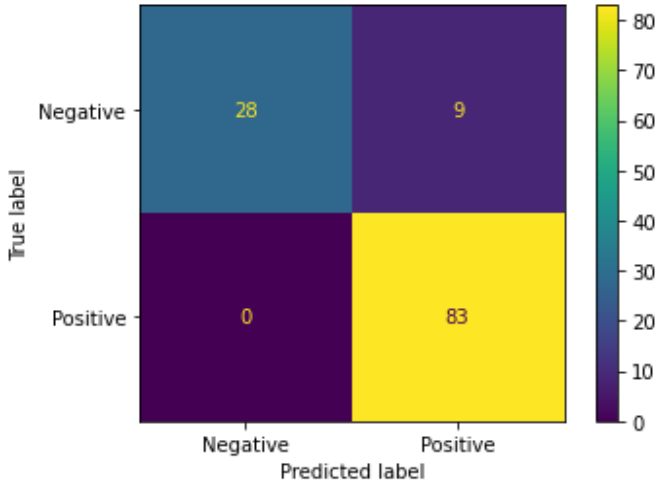

Figure 4.8. Confusion matrix to implement the SVM algorithm on dataset 2.

In the third stage, SVM was implemented on a sub dataset selected by applying feature selection techniques on dataset 1. SVM parameters were tuned as the same parameters in stages 1 && 2 with a change degree value of 11. Figure 4.9. shows the confusion matrix when implementing SVM with feature selection techniques on data set 1.

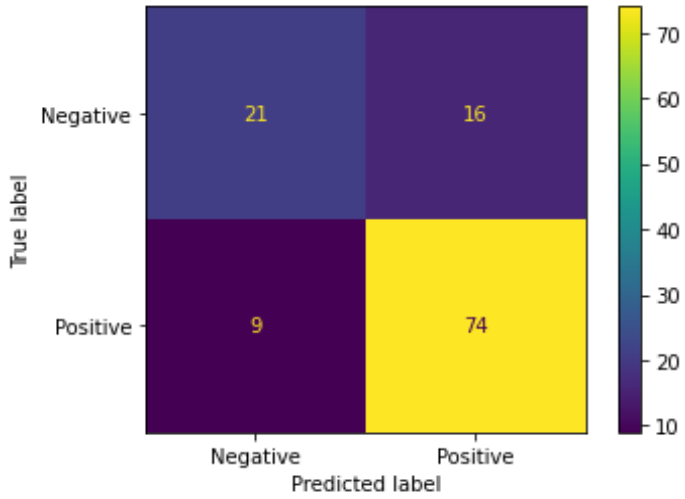

Figure 4.9. Confusion matrix of SVM algorithm with features selected.

4.3.3. Implementing Random Forest (RF) Algorithm

RF algorithm was implemented to predict anemia in children. It is one of the most popular classification algorithms. The RF has a lot of optional parameters that are important for the algorithm to work well. So, it was concentrated on the most important

parameters, such as the number of estimators, maximum tree depth, and random state. Since the goal of random forests is to create many independent trees before selecting the predictions with the most votes. As a result, the number of trees, also known as the number of estimators, and the maximum depth of the tree are good parameters to tune to improve the random forests algorithm's performance. The RF has been implemented to predict anemia in children in three stages.

In the first stage, RF was implemented on dataset 1. RF parameters were tuned as *{n_estimators =100, criterion = 'gini', max_depth=10, min_samples_split =2, random_state=4}*.

where n_estimators is the forest's total number of trees, *criterion* is a function to measure the quality of a split, *max*_depth is maximum depth of the tree, *min_samples_split* is a minimum number of samples required to split an internal node, *random_state* is using to controls the randomness of the bootstrapping of the samples used in tree construction. Figure 4.10. shows the confusion matrix when implementing RF on dataset 1.

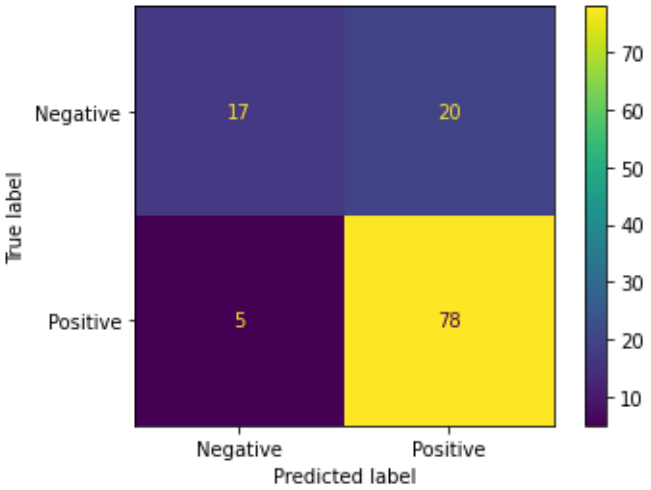

Figure 4.10. Confusion matrix to implement the RF algorithm on dataset 1.

In the second stage, RF was implemented on dataset 2. Again, RF parameters were tuned as the same parameters in stage 1 with a change max depth value to be 4. Figure 4.11. shows the confusion matrix when implementing RF on dataset 2.

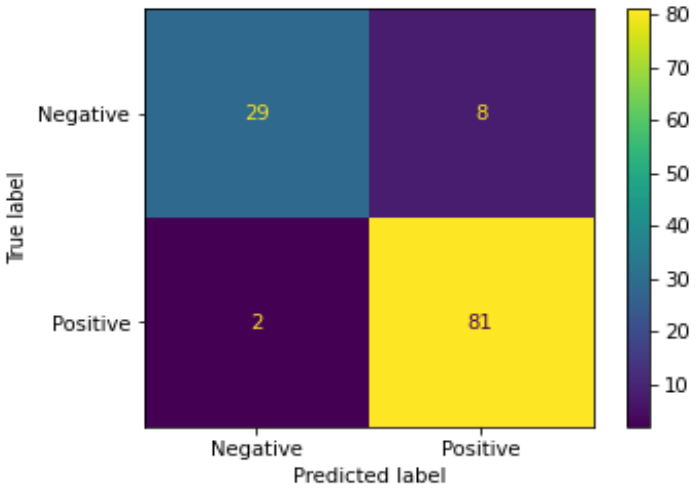

Figure 4.11. Confusion matrix to implement the RF algorithm on dataset 2.

In the third stage, RF was implemented on a sub dataset selected by applying feature selection techniques on dataset 1. Again, RF parameters were tuned as the same parameters in stage 1. Figure 4.12. shows the confusion matrix when implementing RF with feature selection techniques on data set 1.

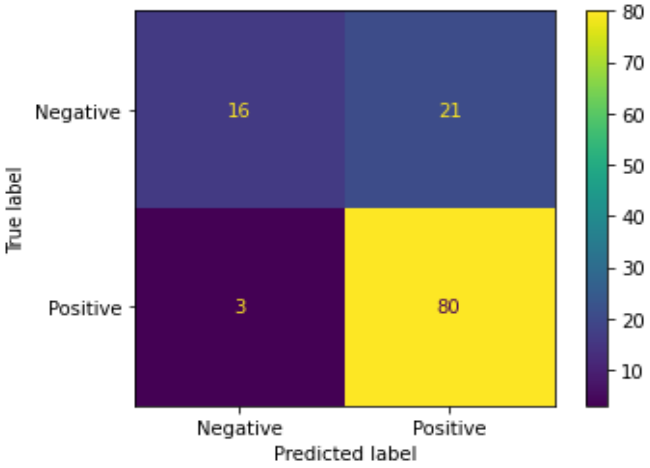

Figure 4.12. Confusion matrix of RF algorithm with features selected.

4.3.4. Implementing Naïve Bayes (NB) Algorithm

NB algorithm was implemented to predict anemia in children. The Bayes' theorem is used to create NB classifiers. NB assumes that one feature in a class has no bearing on the existence of any other feature. The NB has been implemented to predict anemia in children in three stages.

In the first stage, NB was implemented on dataset 1. The settings for the NB parameters were tuned as default settings {*priors=None, var_smoothing=1e-09*}.

where priors are the classes' prior probabilities, *var_smoothing* is the part of the largest variance of all features that have been added to variances for calculation stability. Figure 4.13. shows the confusion matrix when implementing NB on dataset 1.

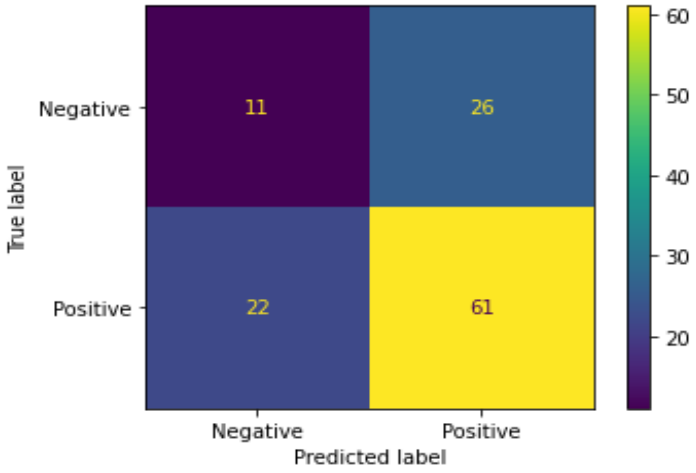

Figure 4.13. Confusion matrix to implement the NB algorithm on dataset 1.

In the second stage, NB was implemented on dataset 2. Again, NB parameters were tuned as the same parameters in stage 1. Figure 4.14. shows the confusion matrix when implementing NB on dataset 2.

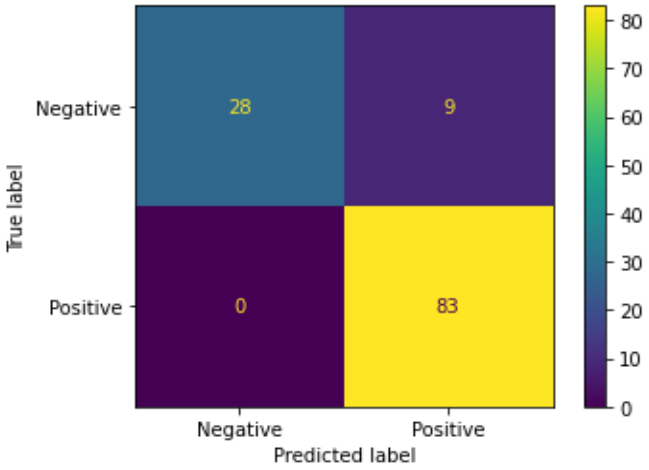

Figure 4.14. Confusion matrix to implement the NB algorithm on dataset 2.

In the third stage, NB was implemented on a sub dataset selected by applying feature selection techniques on dataset 1. Again, NB parameters were tuned as the same parameters in stages 1 & 2. Figure 4.15 shows the confusion matrix when implementing NB with feature selection techniques on data set 1.

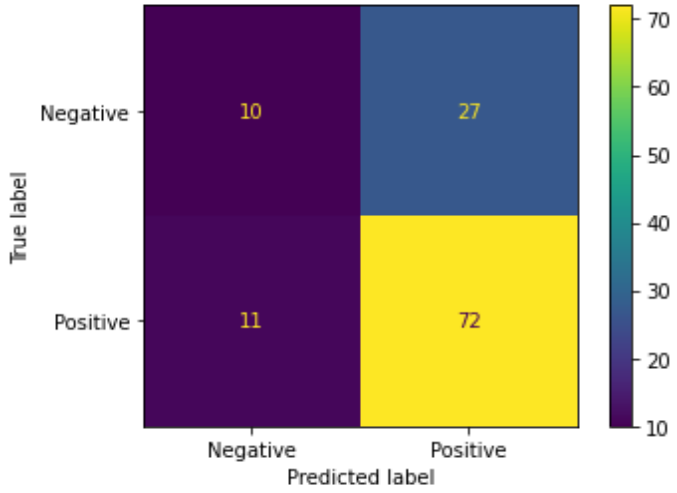

Figure 4.15. Confusion matrix of NB algorithm with features selected.

4.3.5. Implementing Logistic Regression (LR)Algorithm

LR algorithm was implemented to predict anemia in children. LR is one of the most widely used classification methods in the medical field and diagnoses diseases. The LR has a set of optional parameters that are important for the algorithm to work well. The most important of these parameters have been tuned. The LR has been implemented to predict anemia in children in three stages.

In the first stage, LR was implemented on dataset 1. LR parameters were tuned as *{penalty='l2', tol=0.1, C=1.0, fit_intercept=True, class_weight='dict', random_state=0, solver='lbfgs', max_iter=100}.*

where penalty is using to determine the norm used in the penalization, *tol* is tolerance for stopping criteria., *C* is the inverse of regularisation strength; the float must be positive, *fit_intercept* is used to determine if a constant must be added to the decision function, *class_weight* are weights associated with classes, *random_state* is using to

controls the randomness, *solver* is the algorithm used for optimization problems, *max_iter* is the maximum number of iterations the solvers will take to converge. Figure 4.16. shows the confusion matrix when implementing LR on dataset 1.

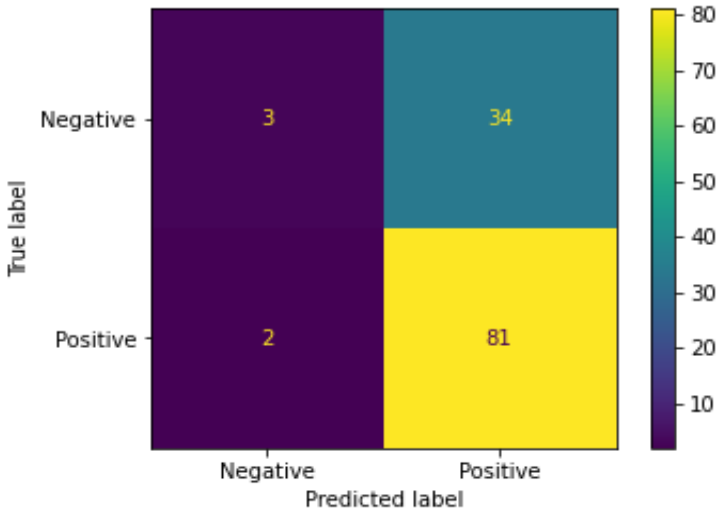

Figure 4.16. Confusion matrix to implement the LR algorithm on dataset 1.

In the second stage, LR was implemented on dataset 2. Again, LR parameters were tuned as the same parameters in stage 1. Figure 4.17. shows the confusion matrix when implementing LR on dataset 2.

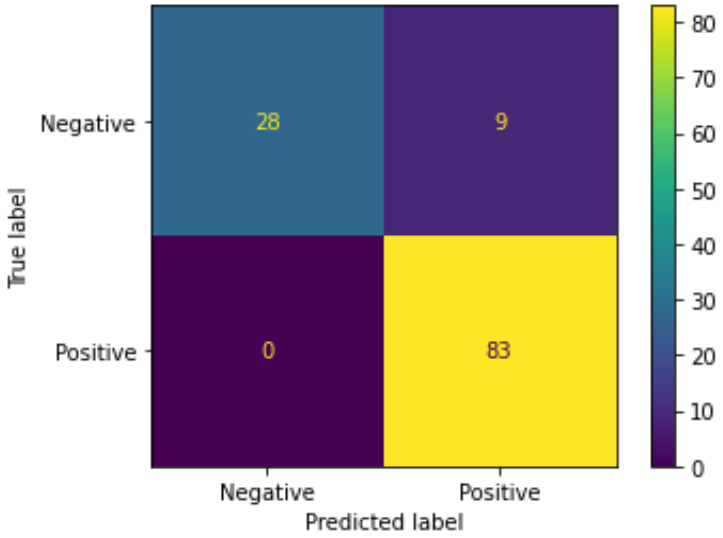

Figure 4.17. Confusion matrix to implement the LR algorithm on dataset 2.

In the third stage, LR was implemented on a sub dataset selected by applying feature selection techniques on dataset 1. Again, LR parameters were tuned as the same parameters in stages 1 && 2. Figure 4.18. shows the confusion matrix when implementing LR with feature selection techniques on data set 1.

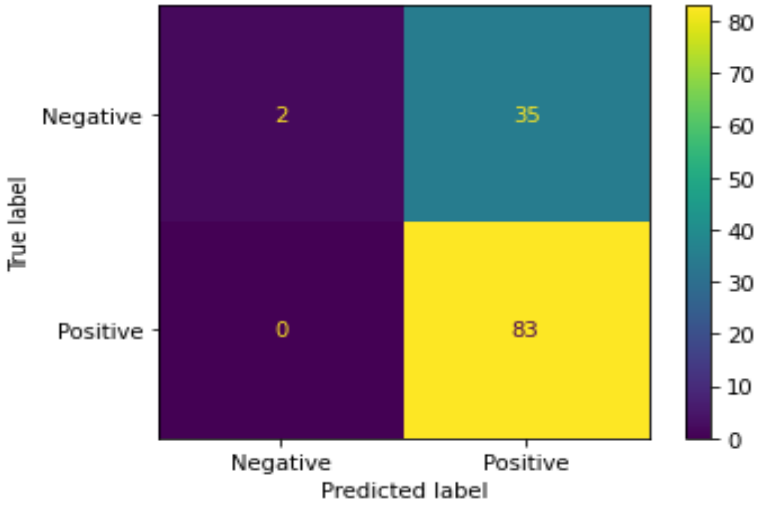

Figure 4.18. Confusion matrix of LR algorithm with features selected.

4.3.6. Implementing Linear Discriminant Analysis (LDA) Algorithm

LDA algorithm was implemented to predict anemia in children. A linear decision boundary classifier created by fitting class conditional densities to data and applying Bayes' rule. The model gives each class a Gaussian density, which assumes that all classes have the same covariance matrix. LDA is usually used to reduce dimensions [63]. LDA has a set of parameters. The most important of them was tuned. The LDA has been implemented to predict anemia in children in three stages.

In the first stage, LDA was implemented on dataset 1. LDA parameters were tuned as *{solver= 'svd', store_covariance=True, n_components=1}*.

Where solver is solver to use, possible values, *store_covariance* is used to computed and stored for the other solvers, *n_components* is the components required for dimensionality reduction. Figure 4.19. shows the confusion matrix when implementing LDA on dataset 1.

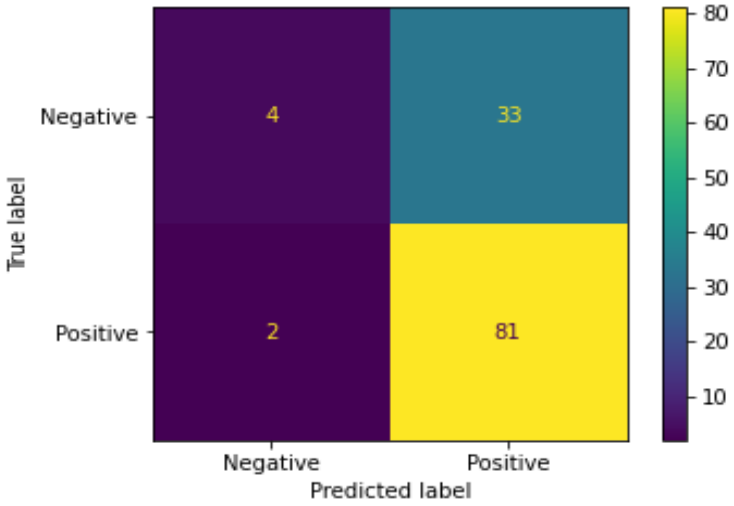

Figure 4.19. Confusion matrix to implement the LDA algorithm on dataset 1.

In the second stage, LDA was implemented on dataset 2. Again, LDA parameters were tuned as the same parameters in stage 1. Figure 4.20 shows the confusion matrix when implementing LDA on dataset 2.

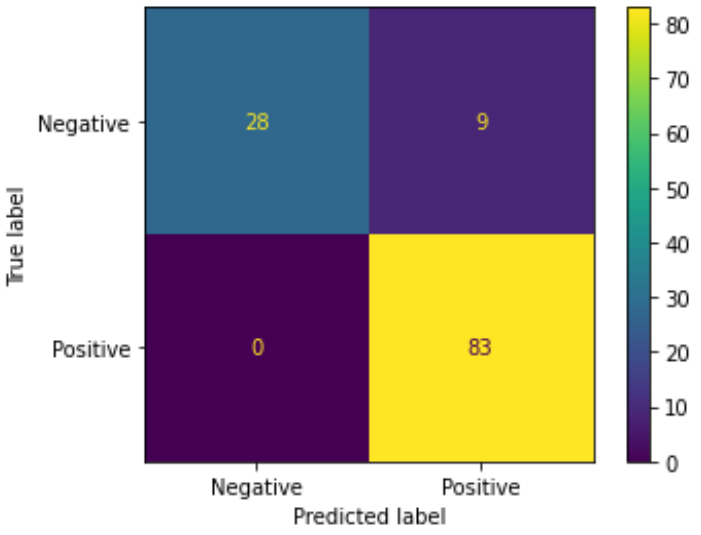

Figure 4.20. Confusion matrix to implement the LDA algorithm on dataset 2.

In the third stage, LDA was implemented on a sub dataset selected by applying feature selection techniques on dataset 1. Again, LDA parameters were tuned as the same parameters in stages 1 & 2. Figure 4.21 shows the confusion matrix when implementing LDA with feature selection techniques on data set 1.

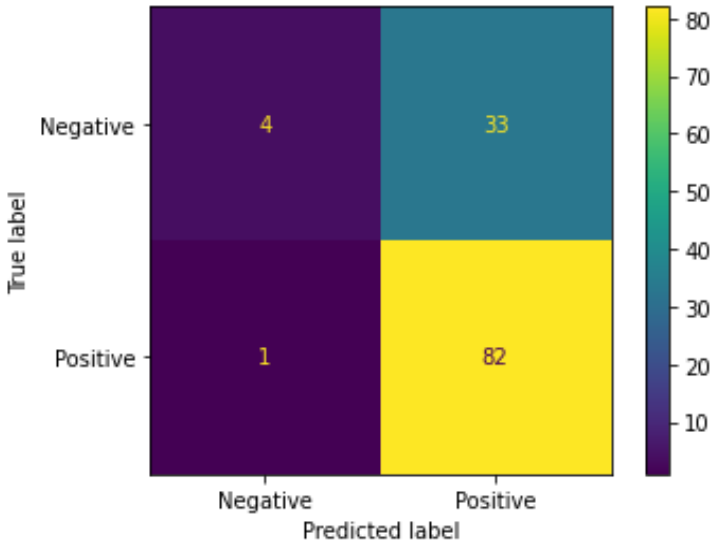

Figure 4.21. Confusion matrix of LDA algorithm with features selected.

4.3.7. Implementing K-Nearest Neighbors (K-NN) Algorithm

KNN algorithm was implemented to predict anemia in children. KNN is one of the simplest classification algorithms that rely heavily on the vote of k-nearest neighbors in its classifications. The KNN has been implemented to predict anemia in children in three stages.

In the first stage, KNN was implemented on dataset 1. KNN parameters were tuned as *{n_neighbors=11, weights='distance', algorithm='auto', p=2}.*

where n_neighbors is the number of neighbors of the class to be classified, *weights* is weight function used in prediction., *algorithm* is the algorithm that was used to calculate the nearest neighbours, *p* is power parameter. Figure 4.22. shows the confusion matrix when implementing KNN on dataset 1.

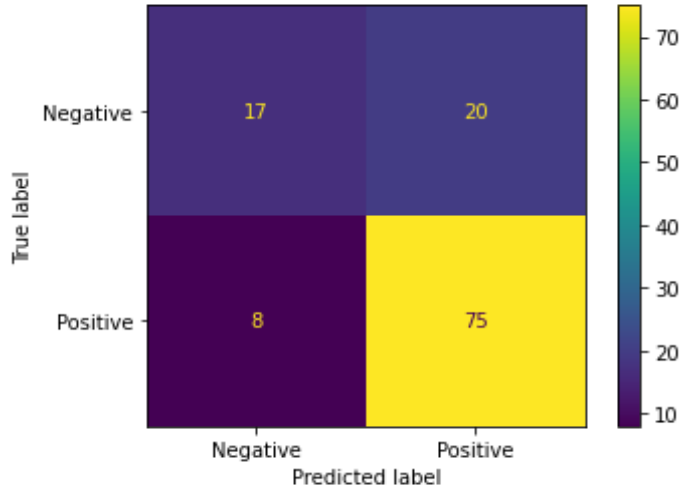

Figure 4.22. Confusion matrix to implement the KNN algorithm on dataset 1.

In the second stage, KNN was implemented on dataset 2. Again, KNN parameters were tuned as the same parameters in stage 1. Figure 4.23. shows the confusion matrix when implementing KNN on dataset 2.

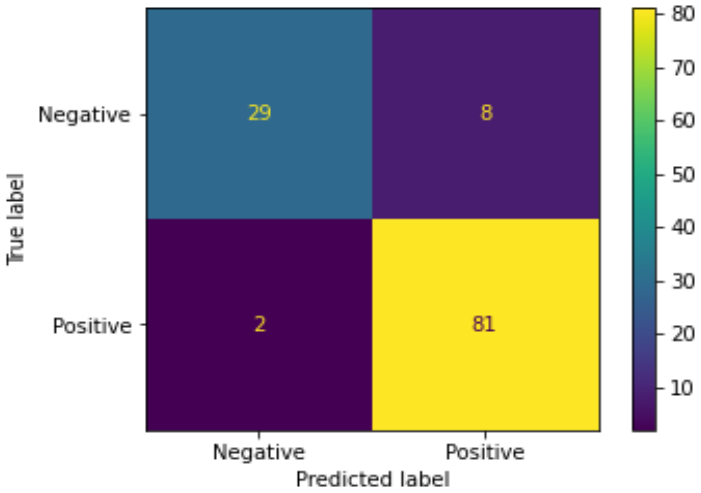

Figure 4.23. Confusion matrix to implement the KNN algorithm on dataset 2.

In the third stage, KNN was implemented on a sub dataset selected using feature selection techniques on dataset 1. KNN parameters were tuned as the same parameters in stages 1 && 2. Figure 4.24. shows the confusion matrix when implementing KNN with feature selection techniques on data set 1.

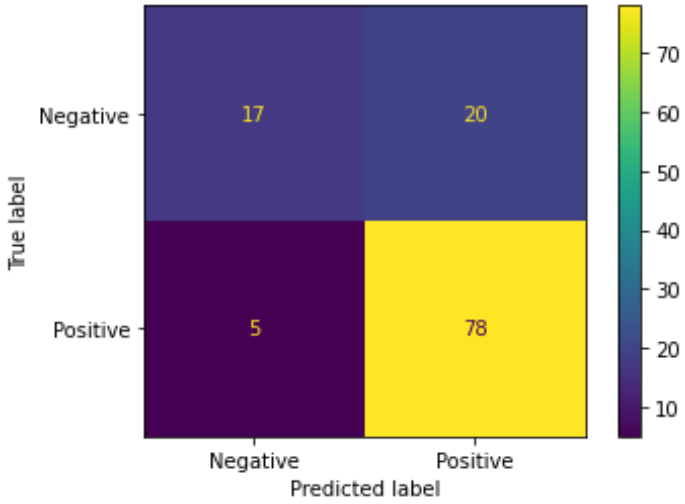

Figure 4.24. Confusion matrix of KNN algorithm with features selected.

4.3.8. Implementing Multilayer Perceptron (MLP) Algorithm

MLP algorithm was implemented to predict anemia in children. Unlike other classification algorithms like SVM and NB, MLP performs the classification task depending on the underlying neural network [66]. The MLP has a lot of optional parameters that are important for the algorithm to work well. So, it was concentrated on the most important of them, such as the number of neurons in the hidden layer, maximum number of iterations, and generate a random number for weights. The MLP has been implemented to predict anemia in children in three stages.

In the first stage, MLP was implemented on dataset 1. MLP parameters were tuned as *{hidden_layer_sizes=100, activation='tanh',alpha=0.1, batch_size=min(200, 600), max_iter=100000}.*

*where hidden_layer*_sizes are number of neurons in the hidden layer, *activation* is activation function for the hidden layer, *alpha* is regularization term, *batch_size* is size of mini batches for stochastic optimizers, *max_iter* is maximum number of iterations. Figure 4.25. shows the confusion matrix when implementing MLP on dataset 1.

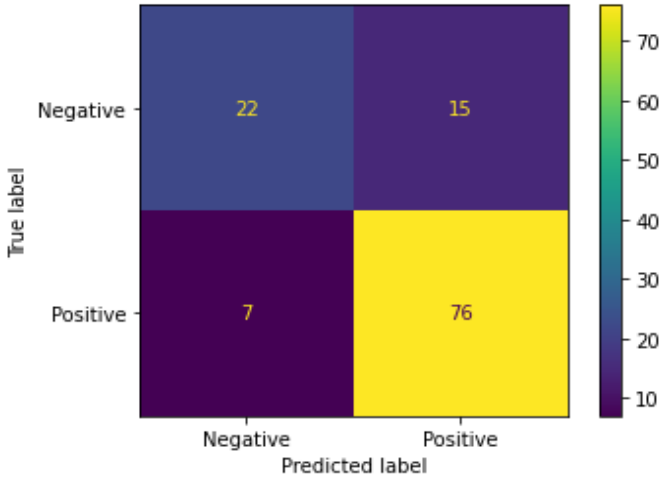

Figure 4.25. Confusion matrix to implement the MLP algorithm on dataset 1.

In the second stage, MLP was implemented on dataset 2. Again, MLP parameters were tuned as the same parameters in stage 1. Figure 4.26. shows the confusion matrix when implementing MLP on dataset 2.

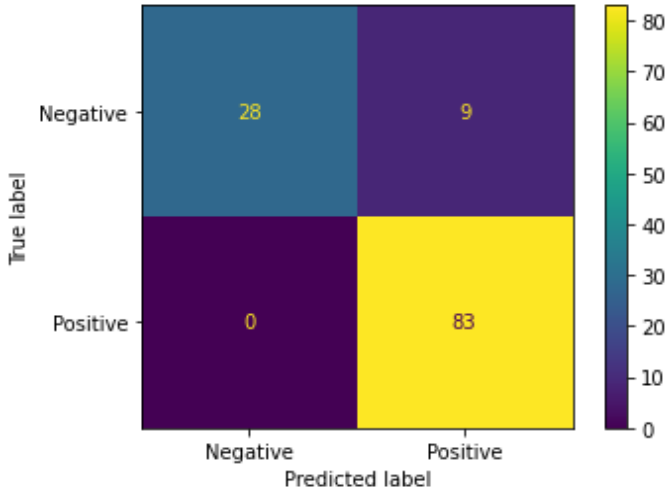

Figure 4.26. Confusion matrix to implement the MLP algorithm on dataset 2.

In the third stage, MLP was implemented on a sub dataset selected by applying feature selection techniques on dataset 1. Again, MLP parameters were tuned as the same parameters in stages 1 && 2. Figure 4.27. shows the confusion matrix when implementing MLP with feature selection techniques on data set 1.

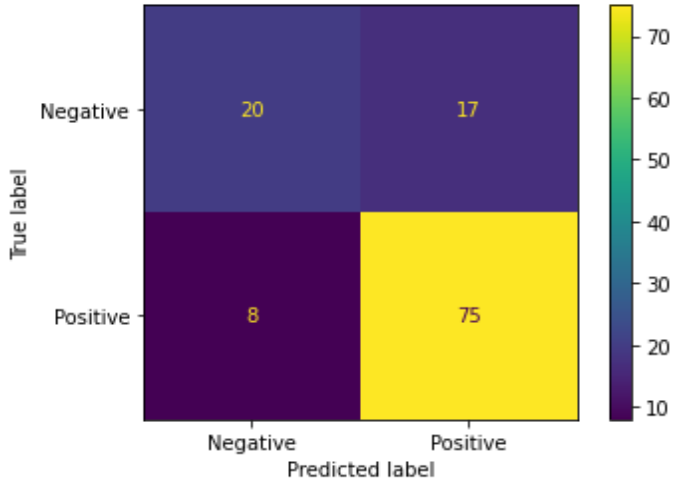

Figure 4.27. Confusion matrix of MLP algorithm with features selected.

4.4. PERFORMANCE MEASUREMENT

To evaluate the performance of the classification methods, a confusion matrix was used to visualize the performance of the techniques. In addition, we use accuracy, precision, sensitivity, specificity and F-Score [28] metrics for evaluation.

4.4.1. Confusion Matrix

The confusion matrix is used to display the performance of the algorithm. The confusion matrix for binary prediction problems consists of two rows and two columns (class 0 and class 1), as shown in Figure 4.28, which categorize the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Anemic samples were considered positive “1”, and not-anemic samples were considered negative “0”.

- **True Positives (TP):** Anemic children who are predicted as anemic.
- **True Negatives (TN):** Not-anemic children who are predicted as not-anemic.
- **False Positives (FP):** Not-anemic children who are predicted as anemic.
- **False Negatives (FN):** Anemic children who are predicted as not-anemic.

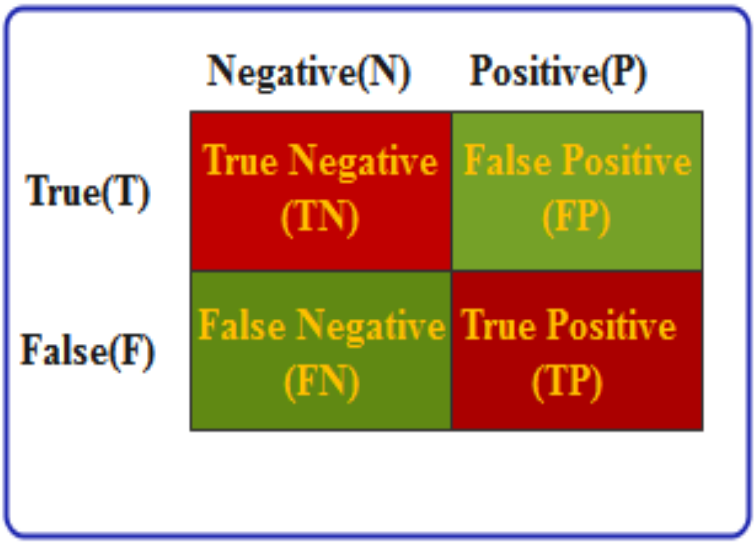

Figure 4.28. Confusion Matrix.

4.4.2. Accuracy

Accuracy of classification the proportion of occurrences correctly classified by the classification learner (anemics classified as anemics and not-anemic classified as not-anemic). Means the ratio of correctly classified samples to the total number of tested samples [27]. The below equation is used to calculate it.

$$\text{Accuracy} = (TN + TP)/(TN + TP + FN + FP) \tag{4.3}$$

4.4.3. Precision

Positive predictive value or precision is the proportion of true positives (anemic classified as anemic). Precision is calculated as the number of correctly positive predicts (anemic classified as anemic) divided by the total number of positively predict (anemic classified as anemic and not-anemic classified as anemic) [27]. The below equation is used to calculate it.

$$\text{Precision} = TP/(TP + FP) \tag{4.4}$$

4.4.4. Sensitivity

Also, recall or true positive is the proportion of positive instances (anemics classified as anemics). Sensitivity is calculated as the number of correctly positive predicts (anemic classified as anemic) divided by the total number of positively predict (anemic classified as anemic and anemic classified as not-anemic) [36]. The below equation is used to calculate it.

Sensitivity = $TP/(TP + FN)$ (4.5)

4.4.5. Specificity

True negative rate or specificity the proportion of negative samples among all negative samples is known as specificity. Specificity is calculated as the number of correctly negative predicts (not-anemic classified as not-anemic) divided by the total number of negatively predict (not-anemic classified as not-anemic and not-anemic classified as anemic) [36]. The below equation is used to calculate it.

Specificity = $TN/(TN + FP)$ (4.6)

4.4.6. F-Score

Also, known as F1 Score ranges from 0 to 1, with 0 being the lowest and 1 being the highest. So F1 is represented as follows: the best value is 1, and the worst value is 0. we can calculate F1 Score using the equation below [27].

F1 Score = $2 * (Precision * Sensitivity/Precision + Sensitivity)$ (4.6)

OR

F1 Score = $2TP/(2TP + FP + FN)$ (4.7)

PART 5

RESULTS & DISCUSSION

5.1. EXPERIMENTS AND RESULTS

The eight ML techniques are built using a scikit-learn library, a powerful library used to implement ML models and pre-processing and model validation phases. The dataset consists of 600 samples with 429 positive samples of anemic children and 171 negative samples of not-anemic children. 80% of the data was used as a training set and 20% used as a test set.

5.1.1. Statistical Analysis

A statistical analysis of the data set used in this study was performed, which contains all the variables of sociodemographic and medical information for children, the practice of feeding the child, and the nutritional knowledge of the mother, as shown in Table 5.1. To find out the prevalence of anemia in children for each variable (risk factors) and determine the association between independent and dependent variables in the dataset. So, with a 95 % confidence interval, binary logistic regression was used to find the correlation between the outcome and its determinants (at 5% significance level) [7]. To determine the significance of each variable, the p-value was calculated using binary logistic regression.

5.1.1.1. Socio-Demographical Characteristics Variables

This section contains all basic social and demographic data such as (Id (the number of samples), age for child, gender, mother age, education level for mother, occupational level for mother, educational level for father, occupation level for father, Place of residence, socio-economic status).

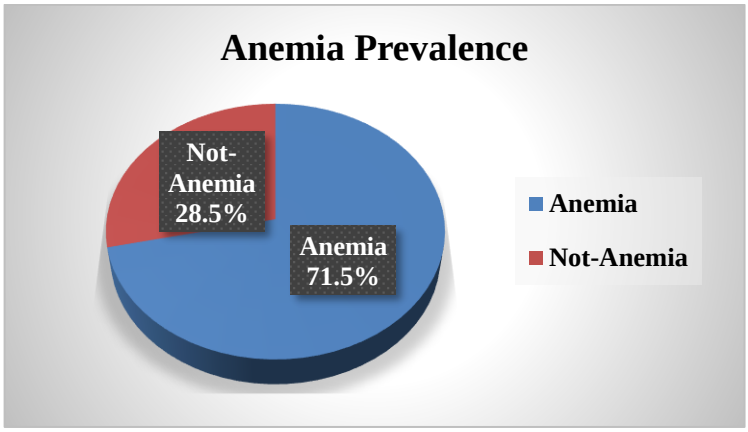

Figure 5.1. Average of anemia prevalence.

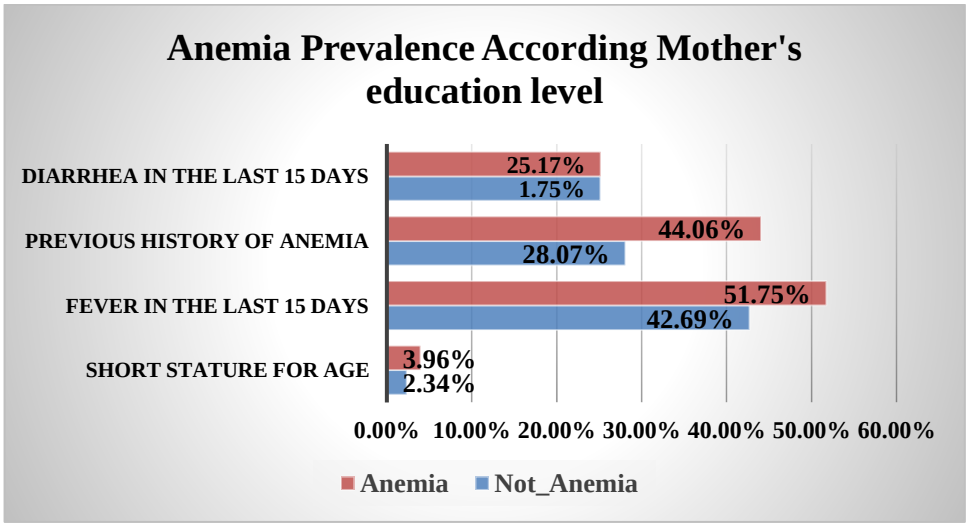

Figure 5.2. Anemia prevalence depending on mother's education level.

.

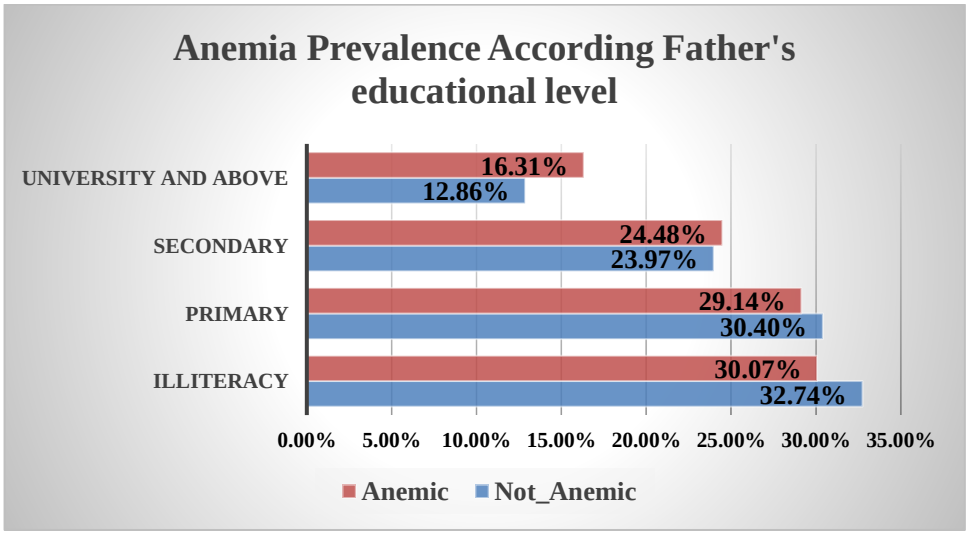

Figure 5.3. Anemia prevalence depending on father's educational level.

5.1.1.2. Clinical Status

This section contains health data for the children participating in this study, such as (Short stature for age, Fever in the last 15 days, Previous history of anemia, and Diarrhea in the last 15 days).

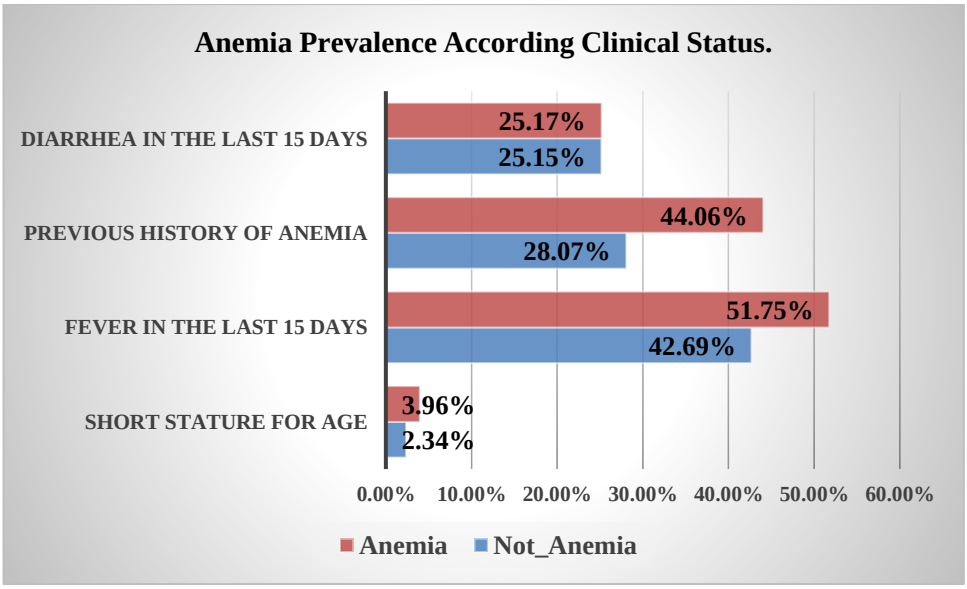

Figure 5.4. Anemia prevalence depending on clinical status.

5.1.1.3. Nutritional Status

This section contains data on nutrition practice for children participating in this study, such as (Type of breastfeeding, consume milk powder, consume the sugary drink, consume yogurt, consume solid/ semisolid food, duration of breastfeeding, consumption of meat, consumption of dark-green leafy vegetables, consumption of foods that are sources of iron and consumption of liver).

5.1.1.4. Mother Knowledge

This section contains a group of questions that define a mother's nutrition knowledge, such as (Know the optimal time of complementary feeding, know the first complementary food, know the optimum food of supplemental iron, know nutrients related to anemia and know the optimal time of breastfeeding).

Table 5.1. Analyses of selected features and anemia status of the children.

All Features	Not-Anemic		Anemic		p-value
	F	%	F	%	
Socio-demographical characteristics variables					
Id	171	28.5%	429	71.5%	0.682
Age groups for child					<0.05
6 months to 2 years	88	51.46%	166	38.69%	
2 -6 years	83	48.53%	263	61.31%	
Gender					0.174
Male	85	49.70%	231	53.85%	
Female	86	50.29%	198	46.15%	
Mother's age					0.681
< 30 years	81	47.36%	237	55.24%	
≥ 30 years	90	52.63%	192	44.76%	
Mother's educational level					<0.05
Illiteracy	97	56.72%	206	48.02%	
Primary	41	23.97%	133	31.00%	
Secondary	30	17.54%	75	17.48%	
University and above	3	1.75%	15	3.50%	
Mother's occupation					0.416
Employee	11	6.43%	24	5.59%	
Un employee	160	93.56%	405	94.41%	
Father's educational level					<0.05
Illiteracy	56	32.74%	129	30.07%	
Primary	52	30.40%	125	29.14%	
Secondary	41	23.97%	105	24.48%	
University and above	22	12.86%	70	16.31%	
Father's occupation					<0.05
Employee	70	40.93%	238	55.48%	
Un employee	101	59.06%	191	44.52%	
Residence area					0.800
Urban	121	70.76%	307	71.56%	
Rural	50	29.23%	122	28.44%	
Socio economic Status					0.412
Low	18	10.52%	62	14.45%	
Moderate	150	87.71%	362	82.38%	
High	3	1.75%	5	1.17%	
Clinical Status					
Short stature for age					0.521
Yes	4	2.34%	17	3.96%	
No	167	97.66%	412	96.04%	
Fever in the last 15 days					0.331
Yes	73	42.69%	222	51.75%	
No	98	57.31%	207	48.25%	
Previous history of anemia					0.266
Yes	48	28.07%	189	44.06%	
No	123	71.93%	240	55.94%	
Diarrhea in the last 15 days					0.689

Yes	43	25.15%	108	25.17%	
No	128	74.85	321	74.83%	
Feeding practice					
Type breast feeding					0.756
Normal	51	29.82%	142	33.10%	
abnormal	85	49.71%	176	41.03%	
mixed	35	20.47	111	25.87%	
Duration of breastfeeding					0.795
normal	83	48.54%	227	52.91%	
abnormal	88	51.46	202	47.09%	
Consume milk powder					<0.05
Yes	126	73.68%	299	69.70%	
No	45	26.32%	130	30.30%	
Consume sugary drink					0.173
Yes	80	46.78%	232	54.08%	
No	91	53.22	197	45.92%	
Consume yoghurt					<0.05
Yes	56	32.75%	167	38.93%	
No	115	67.25	262	61.07%	
Consume solid/ semisolid food					0.669
Yes	54	31.58%	149	34.73%	
No	117	68.42%	280	65.27%	
Consumption of meat					0.737
Yes	74	43.27%	198	46.15%	
No	97	56.73%	231	53.85%	
Consumption of dark-green leafy vegetables					<0.05
Yes	90	52.63%	213	49.65%	
No	81	47.37%	216	50.35%	
Consumption of foods that are sources of iron (meat + beans)					0.287
Yes	81	47.37%	193	44.99%	
No	90	52.63%	236	55.01%	
Consumption of liver					<0.05
Yes	38	22.22%	131	30.54%	
No	133	77.78%	298	69.46%	
Mother's Nutrition knowledge					
Is able identify the optimum timing of complementary feeding					0.130
Know	32	18.71%	108	25.17%	
Do not know	118	69.01%	269	62.71%	
Do not sure	21	12.28%	52	12.12%	
Is able identify to the first complementary food which should be consumed by infants					0.331
Know	36	21.05%	112	26.11%	
Do not know	109	63.74%	256	59.67%	
Do not sure	26	15.20%	61	14.22%	
Has known the optimum food of supplementary iron					0.251
Know	29	16.96%	62	14.45%	
Do not know	107	62.57%	276	64.34%	
Do not sure	35	20.47%	91	21.21%	
Is able identify nutrient relate to anemia					0.706
Know	22	12.87%	53	12.3%	

Do not know	117	68.42%	298	69.46%	
Do not sure	32	18.71%	78	18.18%	
Is able identify the optimum timing of breastfeeding					0.241
Know	39	22.81%	100	23.31%	
Do not know	93	54.39%	240	55.94%	
Do not sure	39	22.80%	89	20.75%	

5.1.2. Experimental Results on Dataset 1

Anemia in children was predicted using Dataset 1, which contains all the features (socio-demographic factors, pediatric medical information, child nutrition practice and mother's nutritional knowledge) except the hemoglobin level.

Table 5.2. shows the results of the algorithms. MLP achieved the best accuracy of 81.67%. DT ranks second with an accuracy of 80.00%.

Table 5.2. Performance comparison of ML algorithms on Dataset 1.

Algorithms	Accuracy	precision	Sensitivity	Specificity	F1 score
DT	80.00%	83.91%	87.95%	62.16%	85.88%
SVM	75.00%	76.77%	91.57%	37.84%	83.52%
RF	79.17%	79.59%	93.98%	45.95%	86.19%
NB	60.00%	70.11%	73.49%	29.73%	71.76%
LR	70.00%	70.43%	97.59%	8.11%	81.82%
LDA	70.83%	71.05%	97.59%	10.81%	82.23%
KNN	76.67%	78.95%	90.36%	45.95%	84.27%
MLP	81.67%	83.51%	91.57%	59.46%	87.36%

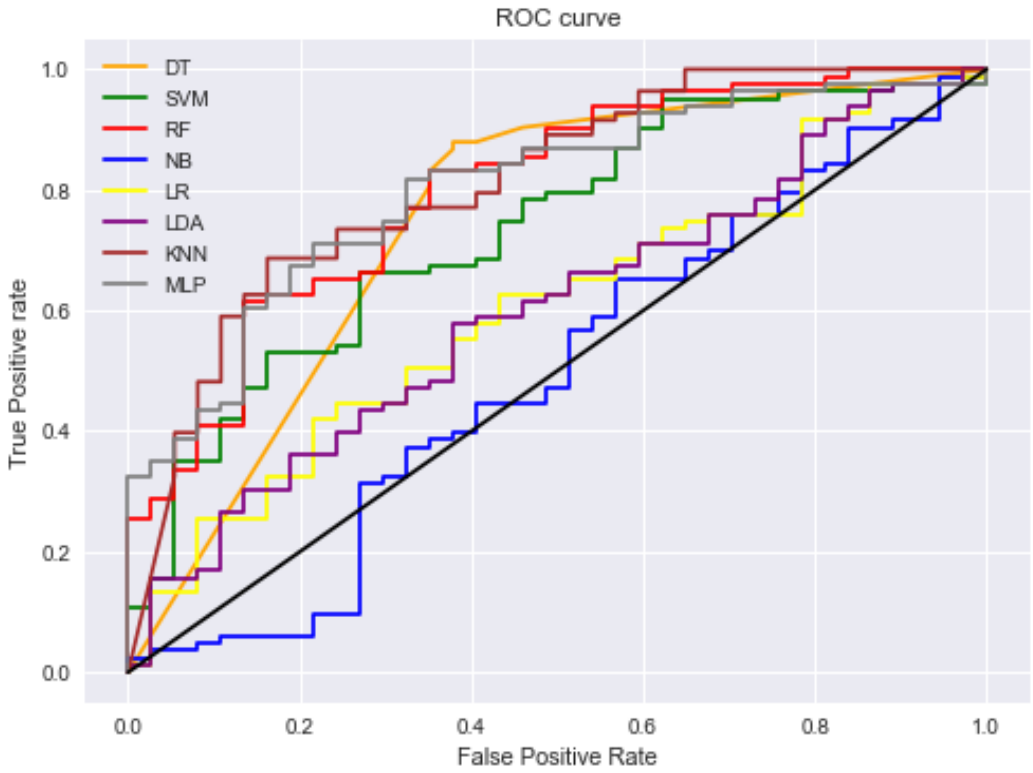

Figure 5.5. Roc curve for performing ML techniques on dataset 1.

5.1.3. Experimental Results on Dataset 2

In this section, anemia in children was predicted by using hemoglobin levels only. SVM, NB, LR, LDA and MLP algorithms achieved the highest accuracy of 92.50%. Results are shown in Table 5.3.

Table 5.3. Performance comparison of ML algorithms on Dataset 2.

Algorithms	Accuracy	precision	Sensitivity	Specificity	F1 score
DT	91.67%	91.01%	97.59%	78.38%	94.19%
SVM	92.50%	90.22%	100.0%	75.67%	94.86%
RF	91.67%	91.01%	97.59%	78.38%	94.19%
NB	92.50%	90.22%	100.0%	75.68%	94.86%
LR	92.50%	90.22%	100.0%	75.68%	94.86%
LDA	92.50%	90.22%	100.0%	75.68%	94.86%
KNN	91.67%	91.01%	97.59%	78.38%	94.19%
MLP	92.50%	90.22%	100.0%	75.68%	94.86%

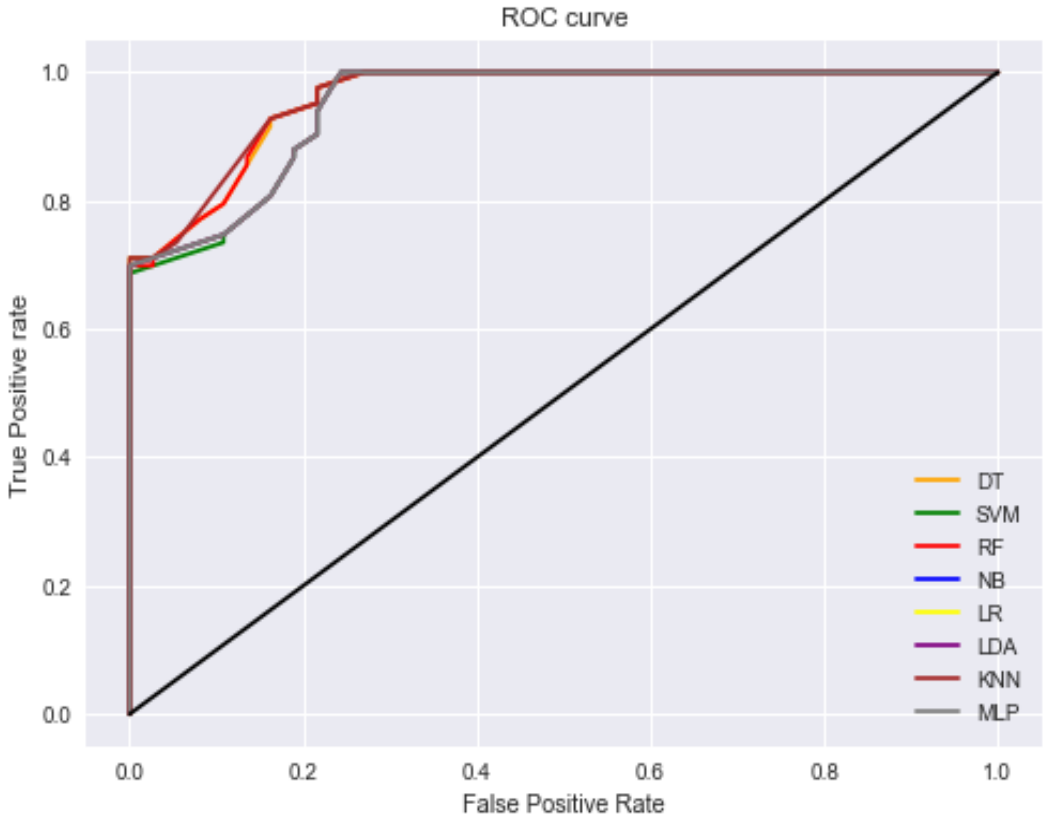

Figure 5.6. Roc curve for performing ML techniques on the dataset 2.

5.1.4. Experimental Results on Dataset 1 Using Feature Selection Techniques

Three different feature selection techniques (Pearson correlation, Recursive Feature Elimination and Decision Tree) were used on Dataset 1 to determine the most appropriate method for better accuracy on ML techniques.

As shown in Table 5.4, remarkable improvement was observed for DT, SVM, KNN, RF, NB LR, and LDN. DT achieved the best result using feature selection techniques with an accuracy of 82.50%, while MLP cannot improve their results.

Table 5.4. Performance comparison of ML algorithms with features selected.

ML Techniques	Feature Selection Techniques	Accuracy	precision	Sensitivity	Specificity	F1 score
DT	Decision Tree	82.50%	88.75%	85.54%	75.67%	87.12%
SVM	Recursive Feature Elimination	79.17%	82.22%	89.16%	56.76%	85.55%
RF	Recursive Feature Elimination	80.00%	79.21%	96.39%	43.24%	86.96%
NB	Recursive Feature Elimination	68.33%	72.72%	86.75%	27.03%	79.12%
LR	Pearson correlation	70.83%	70.34%	100.0%	5.40%	82.59%
LDA	Recursive Feature Elimination	71.67%	71.30%	98.80%	10.81%	82.83%
KNN	Recursive Feature Elimination	79.17%	79.59%	93.98%	45.95%	86.19%
MLP	Recursive Feature Elimination	79.17%	81.52%	90.36 %	54.05 %	85.71%

5.2. DISCUSSION

The ML techniques have achieved promising results in diagnosing anemia and exploring the knowledge of social factors associated with it. This study proved that socio-demographic factors could be considered good predictors of anemia in children. It can be concluded that a clear correlation between the educational level of the father and mother, place of residence and the appearance of anemia, in addition to the importance of the mother's nutritional knowledge and her ability to identify complementary feeding. Also, it can be said that both the type of breastfeeding and the period of normal breastfeeding affects the risk of anemia.

According to the accuracy, the performance of the eight classification techniques was compared, and MLP is the best method without using feature selection techniques as shown in Figure 5.7. When feature selection techniques are used on Dataset 1, DT is the best method as shown in Figure 5.8, RF and MLP are very competitive techniques.

When we analyze the results of the other studies [17,28,29] in the literature among the classification techniques used in the study, SVM has the best accuracy in [17], RF has the best accuracy in [28,29]. Furthermore, the deep learning model in [17] has the best accuracy.

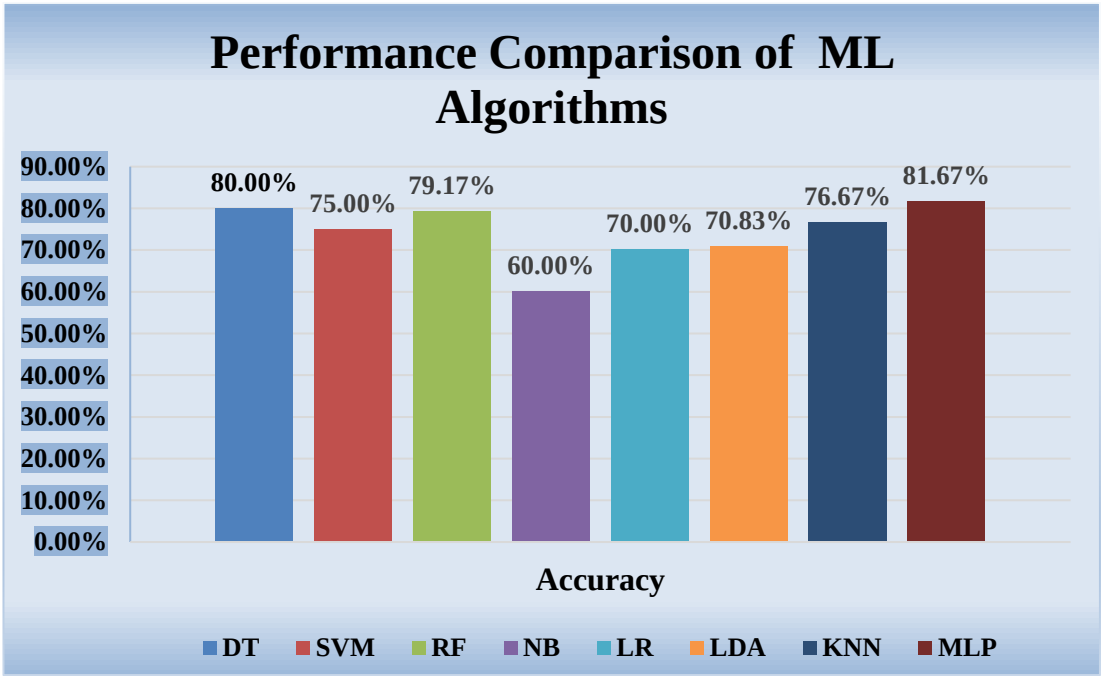

Figure 5.7. Accuracy of ML techniques when implemented on data set 1.

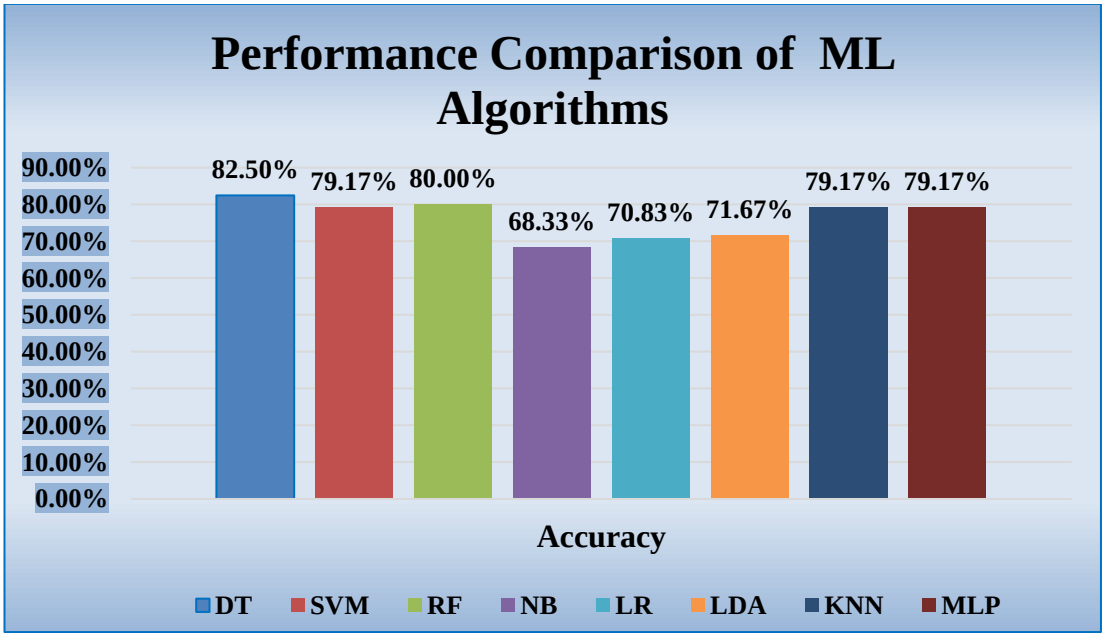

Figure 5.8. Accuracy of ML techniques with selected features.

PART 6

CONCLUSION AND FUTURE WORKS

This study compares eight ML techniques (DT, SVM, RF, NB, LR, LDA, KNN, MLP) to predict anemia in children using social factors as features. These techniques successfully diagnose anemia thorough analysis and deep understanding of data and achieved good results in predicting anemia. We tested the proposed methods in three stages. In the first stage, we use Dataset 1, which contains social factors, and the results indicate that MLP algorithm achieved the highest accuracy of 81.67%. The second stage uses only the hemoglobin level as a feature, and the results show that SVM, NB, LR, LDA and MLP algorithms achieved the highest accuracy of 92.50%. In the last stage, we try to make classification using feature selection methods, and the results indicate that DT outperforms the rest of the algorithms with an accuracy of 82.50%. The results confirm the importance of using social factors to predict anemia in children, which are considered factors associated with disease occurrence. In addition, it is important to choose the appropriate feature selecting technique for features used by the model. As future work, we recommend improving the performance of ML techniques and the use of deep learning techniques to predict anemia in children. Also, different types of classes can be used (Not-anemic, Mild, Moderate, Severe) to predict the anemia in more detail.

REFERENCES

1. Kawo, K. N., Asfaw, Z. G., and 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. Gautam, S., Min, H., Kim, H., and Jeong, H. S., "Determining factors for the prevalence of anemia in women of reproductive age in Nepal: Evidence from recent national survey data", ***PLoS ONE***, 14 (6): 1–17 (2019).

3. Janz, T. G., Johnson, R. L., and Rubenstein, S. D., "Anemia in the emergency department: evaluation and treatment.", ***Emergency Medicine Practice***, 15 (11): (2013).

4. Ewusie, J. E., Ahiadeke, C., Beyene, J., and Hamid, J. S., "Prevalence of anemia among under-5 children in the Ghanaian population: Estimates from the Ghana demographic and health survey", ***BMC Public Health***, 14 (1): 1–9 (2014).

5. Al-Alimi, A. A., Bashanfer, S., and Morish, M. A., "Prevalence of Iron Deficiency Anemia among University Students in Hodeida Province, Yemen", ***Anemia***, 2018: (2018).

6. Department, A. G. A. G., "Iron Deficiency Anaemia in Pregnancy: Developed Versus Developing Countries - European Medical Journal", ***EMJ Hematol***, (August): 101–109 (2018).

7. Patel, K. K., Vijay, J., Mangal, A., Mangal, D. K., and Gupta, S. D., "Burden of anaemia among children aged 6–59 months and its associated risk factors in India – Are there gender differences?", ***Children And Youth Services Review***, 122 (January): 105918 (2021).

8. Seid Adem, O., "Iron Deficiency Aneamia is Moderate Public Health Problem among School Going Adolescent Girls in Berahle District, Afar, Northeast Ethiopia", ***Journal Of Food And Nutrition Sciences***, 3 (1): 10 (2015).

9. Gebreweld, A., Ali, N., Ali, R., and Fisha, T., "Prevalence of anemia and its associated factors among children under five years of age attending at Guguftu health center, South Wollo, Northeast Ethiopia", ***PLoS ONE***, 14 (7): 1–13 (2019).

10. Goswmai, S. and Das, K. K., "Socio-economic and demographic determinants of childhood anemia", ***Jornal De Pediatria (Versão Em Português)***, 91 (5): 471–477 (2015).

11. . K., Sr., M., Ahuja, S., and Nagaraj, N., "Prevalence and risk factors of anemia

in under five-year-old children in children's hospital", ***International Journal Of Contemporary Pediatrics***, 5 (2): 499 (2018).

12. Huang, Z., Jiang, F. X., Li, J., Jiang, D., Xiao, T. G., and Zeng, J. H., "Prevalence and risk factors of anemia among children aged 6-23 months in Huaihua, Hunan Province", ***BMC Public Health***, 18 (1): 1–11 (2018).

13. Tezera, R., Sahile, Z., Yilma, D., Misganaw, E., and Mulu, E., "Prevalence of anemia among school-age children in Ethiopia: A systematic review and meta-analysis", ***Systematic Reviews***, 7 (1): 1–7 (2018).

14. Mattiello, V., Schmugge, M., Hengartner, H., von der Weid, N., and Renella, R., "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 (4): 527–545 (2020).

15. Dukhi, N., Sewpaul, R., Sekgala, M. D., and Awe, O. O., "Artificial intelligence approach for analyzing anaemia prevalence in children and adolescents in brics countries: A review", ***Current Research In Nutrition And Food Science***, 9 (1): 1–10 (2021).

16. Meena, K., Tayal, D. K., Gupta, V., and Fatima, A., "Using classification techniques for statistical analysis of Anemia", ***Artificial Intelligence In Medicine***, 94 (February 2018): 138–152 (2019).

17. Sow, B., Mukhtar, H., Ahmad, H. F., and Suguri, H., "Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques", ***Informatics For Health And Social Care***, 45 (3): 229–241 (2020).

18. Witten, I. H. and Frank, E., "Credibility: Evaluating What's Been Learned", Data Mining: Practical Machine Learning Tools and Techniques, 150 (2005).

19. Aswad, S. A. and Sonuc, E., "Classification of VPN Network Traffic Flow Using Time Related Features on Apache Spark", ***4th International Symposium On Multidisciplinary Studies And Innovative Technologies, ISMSIT 2020 - Proceedings***, (2020).

20. Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., and Sakr, S., "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project", ***PLoS ONE***, 12 (7): 1–15 (2017).

21. Internet: MILLER, D. P. L., "Iron Metabolism Part I: Sources, Transport, Testing - Dr. Philip Lee Miller", **https://blog.antiaging.com/iron-essential-for-oxygenation/** (2021).

22. Balarajan, Y., Ramakrishnan, U., Özaltin, E., Shankar, A. H., and Subramanian, S. V., "Anaemia in low-income and middle-income countries", ***The Lancet***, 378

(9809): 2123–2135 (2011).

23. Porwit, A., McCullough, J., and Erber, W. N., "Blood and Bone Marrow Pathology E-Book", ***Elsevier Health Sciences***, (2011).

24. Killip, S., Bennett, J. M., and Chambers, M. D., "Iron deficiency anemia", ***American Family Physician***, 75 (5): 671–678 (2007).

25. Howard, Martin R and Hamilton, P. J., "Haematology E-Book: An Illustrated Colour Text - Martin R. Howard, Peter J Hamilton - Google Books", ***Elsevier Health Sciences***, (2013).

26. WHO, "Assessment of iodine deficiency disorders and monitoring their elimination", ***Guide For Programme Managers***, Third edit: 1–108 (2007).

27. Ayon, S. I., Islam, M. M., and Hossain, M. R., "Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques", ***IETE Journal Of Research***, 0 (0): 1–20 (2020).

28. Anand, P., Gupta, R., and Sharma, A., "Prediction of Anaemia among children using Machine Learning Algorithms", (June): 469–480 (2020).

29. Khan, J. R., Chowdhury, S., Islam, H., and Raheem, E., "Machine Learning Algorithms To Predict The Childhood Anemia In Bangladesh", ***Journal Of Data Science***, 17 (1): 195–218 (2021).

30. Dithy, M. . and KrishnaPriya, V., "Predicting Anemia in Pregnant Women By Using Gausnominal", 118 (20): 3343–3349 (2018).

31. Karagül Yıldız, T., Yurtay, N., and Öneç, B., "Classifying anemia types using artificial learning methods", ***Engineering Science And Technology, An International Journal***, 24 (1): 50–70 (2021).

32. Mohammed, M. S., Ahmad, A. A., and Sari, M., "Analysis of anemia using data mining techniques with risk factors specification", ***2020 International Conference For Emerging Technology, INCET 2020***, 1–5 (2020).

33. Sanap, S. A., Nagori, M., and Kshirsagar, V., "Classification of Anemia Using Data Mining Techniques BT - Swarm, Evolutionary, and Memetic Computing", 113–121 (2011).

34. Dithy, M. D. and Priya, D. V. K., "Anemia Screening in Pregnant Women by Using Vect Neighbour Classification Algorithm", ***Journal Of Advanced Research In Dynamic And Control Systems***, Volume 11 (04-Special Issue): 1894–1905 (2019).

35. Dithy, M. D. and Krishnapriya, V., "Anemia selection in pregnant women by using random prediction (Rp) classification algorithm", ***International Journal Of Recent Technology And Engineering***, 8 (2): 2623–2630 (2019).

36. Dwivedi, A. K., "Performance evaluation of different machine learning techniques for prediction of heart disease", ***Neural Computing And Applications***, 29 (10): 685–693 (2018).

37. Nkasu, M. M., "Investigation of the Effects of Critical Success Factors on Enterprise Resource Planning (ERP) Systems Implementation in the United Arab Emirates", Smart Innovation, Systems and Technologies, 611–623 (2020).

38. Kunwar, V., Chandel, K., Sabitha, A. S., and Bansal, A., "Chronic Kidney Disease analysis using data mining classification techniques", ***Proceedings Of The 2016 6th International Conference - Cloud System And Big Data Engineering, Confluence 2016***, 300–305 (2016).

39. Markos, Z., "Predicting Under Nutrition Status of Under-Five Children Using Data Mining Techniques: The Case of 2011 Ethiopian Demographic and Health Survey", ***Journal Of Health & Medical Informatics***, 5 (2): (2014).

40. Jordan, M. I. and Mitchell, T. M., "Machine learning: Trends, perspectives, and prospects", 349 (6245): (2015).

41. Kodratoff, Y., "Introduction to Machine Learning", ***Elsevier***, (2014).

42. Simeone, O., "A Brief Introduction to Machine Learning for Engineers", ***ArXiv Preprint ArXiv:1709.02840***, (2017).

43. "Types of Machine Learning | MLK - Machine Learning Knowledge", **https://machinelearningknowledge.ai/types-of-machine-learning/** (2021).

44. Mohan, S., Thirumalai, C., and Srivastava, G., "Effective Heart Disease Prediction using Hybrid Machine Learning Techniques", ***IEEE Access***, PP: 1 (2019).

45. Edition, S., "Data Mining and Knowledge Discovery Handbook", .

46. "1.10. Decision Trees — Scikit-Learn 0.24.2 Documentation", **https://scikit-learn.org/stable/modules/tree.html#mathematical-formulation** (2021).

47. "Machine Learning Decision Tree Classification Algorithm - Javatpoint", **https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm** (2021).

48. Mohsenipour, A. A., "Decision Tree-Based Diagnosis of Coronary Artery Disease: CART Model", ***Computer Methods And Programs In Biomedicine***, 105400 (2020).

49. Cortes, C. and Vapnik, V., "Support-Vector Networks", 297: 273–297 (1995).

50. Yang, Y., "THE RESEARCH OF THE FAST SVM CLASSIFIER METHOD", (1): 121–124 .

51. Boulesteix, A., Janitza, S., and Kruppa, J., "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics", 2 (December): 493–507 (2012).

52. Khalilia, M., Chakraborty, S., and Popescu, M., "Predicting disease risks from highly imbalanced data using random forest", (2011).

53. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., "Classification and Regression Trees", (1984).

54. Internet: Jagannath, V., "File:Random Forest Diagram Complete.Png - Wikipedia", **https://en.wikipedia.org/wiki/File:Random_forest_diagram_complete.png** (2021).

55. Series, I. O. P. C. and Science, M., "Analysis of classification and Naïve Bayes algorithm k-nearest neighbor in data mining", (2020).

56. Vembandasamy, K., Sasipriya, R., and Deepa, E., "Heart Diseases Detection Using Naive Bayes Algorithm", 2 (9): 441–444 (2015).

57. "Naive-Bayes.Png (3095×1549)", **https://thatware.co/wp-content/uploads/2020/04/naive-bayes.png** (2021).

58. Kemppainen, L. M., Kemppainen, T. T., Reippainen, J. A., Salmenniemi, S. T., and Vuolanto, P. H., "Use of complementary and alternative medicine in Europe : Health-related and sociodemographic determinants", (February): 1–8 (2017).

59. "Logistic Regression in Machine Learning - Javatpoint", **https://www.javatpoint.com/logistic-regression-in-machine-learning** (2021).

60. Chao, C., Yu, Y., and Cheng, B., "Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine , Logistic Regression and Decision Tree", 1–7 (2014).

61. Tharwat, A., Gaber, T., Ibrahim, A., and Ella, A., "Linear discriminant analysis : A detailed tutorial", 30: 169–190 (2017).

62. Xanthopoulos, P., Pardalos, P. M., and Trafalis, T. B., "Linear Discriminant Analysis", ***Springer, New York, NY***, 27–33 (2013).

63. Ricciardi, C., Valente, A. S., Edmund, K., Cantoni, V., and Santini, S., "Linear discriminant analysis and principal component analysis to predict coronary artery disease", (2020).

64. Taylor, P., Altman, N. S., and Altman, N. S., "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression An Introduction to Kernel and

Nearest-Neighbor Nonparametric Regression", (December 2014): 37–41 (2012).

65. Cai, Z., Gu, J., Wen, C., Zhao, D., Huang, C., Huang, H., Tong, C., Li, J., and Chen, H., "An Intelligent Parkinson ' s Disease Diagnostic System Based on a Chaotic Bacterial Foraging Optimization Enhanced Fuzzy KNN Approach", 2018: (2018).

66. Hastie, Trevor and Tibshirani, Robert and Friedman, J., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", ***Springer Science \& Business Media***, 2009file:///D:/Anrmia of children/Work/Anemic_not .

67. Sharifzadeh, F. and Akbarizadeh, G., "Ship Classification in SAR Images Using a New Hybrid CNN – MLP Classifier", ***Journal Of The Indian Society Of Remote Sensing***, 0123456789: (2018).

68. Krizhevsky, B. A., Sutskever, I., and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks", (2012).

69. Singh, D. and Singh, B., "Investigating the impact of data normalization on classification performance", ***Applied Soft Computing Journal***, 105524 (2019).

70. Das, A. K., Sengupta, S., and Bhattacharyya, S., "SC", ***Applied Soft Computing Journal***, (2018).

71. Saidi, R., Bouaguel, W., and Essoussi, N., "On the Genetic Algorithm and Pearson Correlation Coefficient", ***Springer International Publishing***, .

72. Guyon, I., "Gene Selection for Cancer Classification", 389–422 (2002).

73. Sugumaran, V. Ã., Muralidharan, V., and Ramachandran, K. I., "Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing", 21: 930–942 (2007).

RESUME

Qusay Luay SAIHOOD graduated first and elementary education in Baghdad-Iraq. He completed high school education at (Barwana High School) in Al Anbar governorate, then, he obtained bachelor's degree from University of Tikrit/College of Computer sciences and Mathematics department of Computer sciences in 2018. After graduation he worked with UNDP as field monitors. To complete M.Sc. education, he moved to Karabuk/Turkey in 2019. He started his master education at the department of computer engineering in Karabuk University.