

S	
E1	
E2	
For Office Use Only	



**Masters Project Final Report**  
**(MCS)**  
**2019**

Project Title	Determine the domestic hematology reference range for healthy adults in Sri Lanka
Student Name	K.A.H. Semini
Registration No. & Index No.	Registration Number: 2017/MCS/073 Index Number: 17440739
Supervisor's Name	Dr. H.A. Caldera

For Office Use ONLY

## Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: K.A.H.Semini

Registration Number: 2017/MCS/073

Index Number: 17440739

---

Signature:

Date:

his is to certify that this thesis is based on the work of Mrs. K.A.H. Semini under my supervision. The thesis has been prepared according to the format stipulated and is of an acceptable standard.

Certified by:

Supervisor Name: Dr. H.A.Caldera

---

Signature:

Date:



# **Determine the domestic hematology reference range for healthy adults in Sri Lanka**

**A dissertation submitted for the Degree of Master of  
Computer Science**

**K. A. H. Semini  
University of Colombo School of Computing  
2020**



## **Acknowledgment**

First and for most, I would like to give my glory and praise to Dr (Mr) H A Caldera, Senior Lecturer and My Research Supervisor for his invaluable cares and support throughout the research. I'm grateful to appreciate him who has taken all the trouble with me while I was doing this research. Especially his valuable and prompt advice, constructive corrections and insightful comments, suggestions and encouragements are highly appreciated.

Next, my special gratitude goes to the Dr (Mr) M. H. B. Ariyaratne, Medical Officer as well as the Senior Registrar in Health Information at the Ministry of Health and Indigenous Medicine and also my Co-supervisor, who gave this research topic for me. Especially his ideas, guidance, motivations are highly appreciated and once again I thanked him to his big support throughout the whole research period.

Finally, I would like to thank my husband, parents, and very best friends for encouraging me to finalize this research as soon as possible. Moreover, I highly appreciate everybody who supports me in various types to complete my research.

# Abstract

Hematology science is a subsection in medical science. It is one of the very useful things to measure the healthiness of a person. There is a standard hematology referential range for each hematology attribute. This standard hematology referential range is built many years ago by considering the Gaussian population. All countries use this standard referential range to measure the healthiness of the people. But lately some countries identify that the referential ranges can be changed according to various reasons and using the standard referential range to measure healthiness is not accurate. Therefore, many countries started researches to find a local hematology referential range for their own countries and successfully established them. Through these researches they able to found other hidden patterns related to hematology.

Sri Lanka is not a Gaussian country. Hence, using the same standard hematology referential range maybe not accurate for our country. Therefore, finding a local referential range for the White Blood Cell count is the goal of this research. The compatibility of the standard hematology referential range of white Blood Cell count with Sri Lanka was checked through this research.

The dataset used for the research contained 600 records related to the full blood count reports. Next, the dataset was separated as a training dataset and testing dataset. The dataset was analyzed using the WEKA data mining tool. Machine learning algorithms are used to build the model by using the training dataset. Several models are built by using classification algorithms and one model was selected by considering its accuracy.

The selected model was evaluated by using the testing dataset. The local referential range for White Blood Cell count is found by using the evaluation. There is a difference between the standard White blood cell referential range and the local white blood cell referential range. Therefore, Applying the standard white blood cell referential range for Sri Lankans may be not accurate.

# Content

Acknowledgement.....	iv
Abstract.....	v
Content.....	vi
List of Figures .....	x
List of Tables.....	xi
Introduction.....	1
1.1 Introduction.....	1
1.2 Problem.....	2
1.3 Problem Domain .....	2
1.3.1 Machine Learning .....	3
1.3.2 Classification and Clustering.....	3
1.3.3 Medical Science and Hematology science .....	4
1.4 Motivation.....	4
1.5 Exact Computer Science Problem.....	5
1.6 Research Contribution .....	5
1.6.1 Aim .....	5
1.6.2 Objective of Study .....	6
1.7 Scope .....	6
1.8 Evaluation.....	6
1.9 Structure of the Dissertation .....	7
Literature Review .....	8
2.1 Introduction.....	8
2.2 Similar Approaches Based on statistical Analysis.....	8
2.3 Similar Approaches Based on Machine Learning.....	10
2.4 Summary.....	13

Technology Adapted.....	14
3.1 Introduction.....	14
3.2 Software Tools .....	14
3.2.1 Data Mining Tools .....	14
3.2.2 Statistical analysis tools .....	18
3.3 Statistical Techniques.....	22
3.3.1 Mann-Whitney U test.....	22
3.3.2 Chi-square test.....	23
3.3.3 Kruskal-Wallis test .....	24
3.3.4 Kolmogorov– Smirnov test .....	25
3.3.5 Shapiro–Wilk test .....	25
3.3.6 t-test.....	26
3.4 Machine Learning Classification Algorithms.....	26
3.4.1 Random Forest classification Algorithm .....	29
3.4.2 Decision Tree Algorithm.....	29
3.4.3 Naïve Bayes Algorithm.....	30
3.5 Features used to measure the accuracy of the model .....	30
3.6 Summary.....	33
Methodology .....	34
4.1 Introduction.....	34
4.2 Abstract View .....	34
4.3 Generating the Dataset .....	35
4.3.1 Selection of the sample .....	35
.....	35
4.3.2 The ethical considerations .....	36
4.3.3 Prepare the Data Set.....	37
4.3.4 Generate Health Code/health State for the data set .....	38

4.4 Prepare training dataset and testing dataset .....	39
4.5 Pre-processing of dataset .....	40
4.6 Build classification data model .....	42
4.9 Statistical analysis .....	50
4.9.1 Descriptive statistic .....	50
4.9.2 Check Normality by using Data Visualization .....	52
4.9.3 Applying Statistical tests into the dataset .....	53
4.10 Summary .....	55
Results and Evaluation .....	53
5.1 Introduction .....	53
5.2 Test and evaluate the built model .....	53
5.3 Find the corresponding WBC referential range .....	55
5.4 Results of statistical-based data analysis .....	56
5.5 Summary .....	58
Chapter 06 .....	59
Conclusion Further work .....	59
6.1 Introduction .....	59
6.2 Overall Achievements .....	60
6.3 Limitations of the findings .....	61
6.4 Achievements of the objective .....	61
6.5 Further Works .....	62
6.2 Summary .....	64
References .....	65
Appendix A .....	69
Appendix B .....	70
Appendix C .....	71
Appendix D .....	72



Appendix E .....	73
Appendix F.....	74

# List of Figures

Figure 1 : Data Mining Process .....	15
Figure 2 : Classification algorithm .....	27
Figure 3: Steps of Classification algorithm .....	28
Figure 4 : Selection of the sample.....	35
Figure 5: Sample of received data set .....	37
Figure 6: Re-prepare data set.....	38
Figure 7: Standard Hematology ranges which are consider to generate health code .....	38
Figure 8 : Training Dataset and Testing Dataset .....	40
Figure 9: Data visualization in Weka pre-processing tab .....	41
Figure 10: Threshold curve for unhealthy instances in the model built by using the Random Forest algorithm.....	46
Figure 11: Threshold curve for healthy instances in the model built by using the Random Forest algorithm.....	47
Figure 12: Threshold curve for unhealthy instances in the model built by using the Random Tree algorithm.....	47
Figure 13: Threshold curve for healthy instances in the model built by using the Random Tree algorithm.....	48
Figure 14: frequency visualize from histogram with normal curve.....	52
Figure 15: Q-Q plot with diagonal line .....	53
Figure 16 : Frquency graphs belong to haemoglobin .....	54
Figure 17 : Frequency graphs belong to hemoglobin with the normal curve.....	54
Figure 20: model testing – accuracy result.....	54
Figure 21: WBC value.....	55
Figure 18 : Summary of Hemoglobin dataset.....	56
Figure 19: Result of Mann-Whiney U test applying into Hemoglobin value .....	57

# List of Tables

Table 1: Standard Hematology Referential Ranges .....	1
Table 2: Data Mining Vs. Data Analysis .....	16
Table 3: Data mining Tools with their features .....	17
Table 4 : Statistical Analysis tools with their features.....	20
Table 5: Accuracy features of models -I .....	43
Table 6: Accuracy features of models -II.....	44
Table 7: Confusion Matrix .....	49

## Introduction

### 1.1 Introduction

The reference range or reference interval is a value that is considered to measure the healthiness of a person in Health-related fields. Reference ranges for blood tests (these values blood-related reference ranges are also called “Hematology Reference Range”) and urine tests are some examples of important reference ranges in the medical sector [1]. Only the hematology reference ranges are considered in this research. These hematology reference ranges are very important for monitoring pathophysiological changes after an infection or disease, to detect diseases such as Dengue fever, HIV, cancer, etc., for the administration of drugs in therapeutic, clinical interventions, vaccine studies, etc. [2]. There is a set of values for hematology reference ranges that are accepted worldwide. It is called “Standard hematology reference ranges”. Below values are examples for some Standard hematology reference ranges [3].

<b>Hematology Reference Ranges</b>	<b>Normal Adult</b>	<b>Males</b>	<b>Females</b>	<b>Units</b>
Hemoglobin (HB)		130-180	130-180	g/L
White Cell Count (WBC)		4-11	4-11	$10^9/L$
Platelet Count (PLT)		150-450	150-450	$10^9/L$
Red Blood Count (RBC)		4.5-6.5	3.8-5.8	$10^{12}/L$
Mean Cell Volume (MCV)		80-100	80-100	80-100
Packed Cell Volume (PCV)/Hematocrits (HCT)		0.40-0.52	0.37-0.47	L/L
Mean Cell Hemoglobin (MCH)		27-32	27-32	g
Mean Cell hemoglobin Concentration (MCHC)		320-360	320-360	g/L
Neutrophil Count		2.0-7.5	2.0-7.5	$10^9/L$
Lymphocyte Count		1.5-4.5	1.5-4.5	$10^9/L$
Eosinophil value		0-0.4	0-0.4	$10^9/L$
Monocyte Count		0.2-0.8	0.2-0.8	$10^9/L$

Table 1: Standard Hematology Referential Ranges

These standard hematology reference range values were determined many years ago by doing some researches for the Caucasian populations. A Caucasian population is a group of people who are originated from Europe and are also commonly known as “white” or “white-skinned” people [4] [5].

## **1.2 Problem**

Gradually, people realized that the Standard hematology reference range can vary due to many reasons such as age, gender, genetics, attitudes, lifestyle, ethnic origin, dietary habits, geographical location, climate, environmental factors, etc. [6] [2] [7]. Hence the Clinical and Laboratory Standards Institute (CLSI) recommended that a domestic hematology reference range should be established for each region, because many different factors which are mentioned above, have an impact on the hematological parameters in different populations [8]. As a result, hematology reference ranges per each country have been focused to be established [9] [6]. Eastern India, China, Sudan, Malawi, Togo, Nigeria, and many African countries have successfully done many kinds of research and established the domestic hematology reference ranges for their countries which contain different values rather than the standard hematology reference range [6] [4] [2] [7] [9].

When considering Sri Lanka, the population is not a Caucasian population. Hence the standard hematology reference range might not be applicable for Sri Lanka due to the difference in dietary habits, geographical location, climate, environmental factors, etc. relative to the population, for which the standard hematology reference ranges are determined. But a domestic hematology reference ranges for Sri Lanka has not been determined yet. This is the problem that is going to be addressed with this research.

## **1.3 Problem Domain**

This research is related to the domain of medical science and the solution approach is related to a few major areas in modern computer science such as Data mining and Machine learning.

### **1.3.1 Machine Learning**

Machine learning (ML) is a sub-branch that comes under the vast tree of Artificial Intelligence (AI). It is called the brain science of modern-world computer science. In simple terms, ML is the study of the pattern description and prediction with statistical algorithms. These algorithms can act automatically with no interference from a human.

There are 3 common phases in these algorithms. They are,

- Implementation
- Training
- Testing

We can directly find some information through the data sets but if we train these data sets through some appropriate algorithms using ML, then we can find some unseen information, relationships that we couldn't find earlier. So, these algorithms act as some filters that can identify and extract valuable information from data. ML exerts some initial basic instruction or a set of previous experiences which have been collected earlier, to derive and obtain some unseen information from new data. Machine Language has several overlaps with statistics and basics of AI, hence ML is not a field that stands by itself. ML-based solutions are highly used in the vast area of application domains in the current world. face detection, fraud detection, online recommendations, elevator scheduling, Social Media Services, Online Customer Support, etc.

### **1.3.2 Classification and Clustering**

There are two basic categories of training algorithms based on the behavior and evolving themselves to predict the classes. Supervised learning and unsupervised learning is these two categories. In supervised learning algorithms, they are trained by some training data set whose target values/labels are known, while in unsupervised learning algorithms those target values are not known. Classification algorithms are categorized as supervised learning approaches, since as they exploit training data to initialize the algorithms for automatic categorization while using clustering most of the time to

do the data grouping according to a measure of distance. Unsupervised learning algorithms are used to find classes of similar data sets automatically.

### **1.3.3 Medical Science and Hematology science**

Medical science investigates, evaluates, and explains how the human body functions as a system. Medical science can be divided into several domain areas such as physiology, anatomy, hematology, and pathology concerning the number of subject areas such as biochemistry, microbiology, molecular biology, and genetics, etc. This research is focused on Hematology science.

Hematology science can be introduced as a study of blood, blood-forming organs, and diseased related to blood. Hematology includes the treatment of numerous blood disorders and malignancies such as leukemia, hemophilia, lymphoma, etc. Hematology Science has introduced the referential value range for males and females as a result of many types of research. People can do hematology tests such as complete blood count, etc. and can detect clotting problems, anemia, immune system disorders, blood cancers, and countless infections by analyzing these referential values.

### **1.4 Motivation**

Sri Lanka population is not a Caucasian population. So, using the hematology reference range which was introduced by considering the Caucasian population is not best applicable for our country. Most doctors in the medical sector face this problem. When doctors are considering a blood report, sometimes they know and identify that the patient is healthy even the blood referential value is not between the ranges by their experiences. So, this research opinion is categorized as “something should need to do” by the doctors who are met. Also, the Clinical and Laboratory Standards Institute (CLSI) recommended that a domestic hematology reference range should be established for each region too [8]. Most of the researchers who worked with hematology referential ranges, introduced domestic hematology referential range for their countries.

## **1.5 Exact Computer Science Problem**

In previous subsections are showed that many other countries have introduced domestic hematology referential range for their countries. So, the standard hematology referential range which is used currently in Sri Lanka should need to verify that it is suitable for our country or not.

When considering the medical sector in Sri Lanka, simply it can divide into two as government medical sector and the private medical sector. Both of them use hematology reports to detect patients' healthiness. But the problem is, that the detect criteria (here the standard hematology referential range) can be considered to our country or not. So, checking this problem through a suitable data set will be our computer science problem. The main problem can divide into subproblems as follows.

- **Sample:** It is difficult to consider the whole population in Sri Lanka to collect hematology referential values. So, it is needed to find a suitable sample of the population.
- **Data Set:** To find a suitable data set is next subproblem. A large and accurate data set will be given the best outcome from this research.
- **Analysis:** Finding the most suitable data analysis method will be another major problem in this research. This will be also helped to develop a referential data model too.

## **1.6 Research Contribution**

### **1.6.1 Aim**

Determine a domestic hematology reference range for Sri Lankan adults is the aim of this research. Accuracy and appropriateness of using the standard hematology reference range for Sri Lankans can be evaluated through this domestic hematology reference range, which is to be determined. This can be done by comparing the standard hematology reference range which



is being used at present in Sri Lanka with the domestic hematology reference range of Sri Lanka which will be found at the end of this research.

### **1.6.2 Objective of Study**

To achieve the above-mentioned aim, some objectives must be covered. These objectives are listed below.

- Critical studying of the area of blood reference ranges /blood test categories/ reasons for changing blood reference ranges, etc.
- A critical survey on technologies that can be applied for the project.
- Gather the Suitable data set to accomplish the task.
- Process the data set and build the data model.
- Evaluate the model.
- Do statistical analysis and describe outcomes.
- Produce an efficient research document.

### **1.7 Scope**

The considered population is adults (above 21 years old) in Sri Lanka. The blood reports which will be used as data set for this study are generated for a sample within this population. The considered sample may contain healthy and unhealthy people. White Blood Count (WBC) hematology reference range will check at the end of this research. And describe using standard hematology referential range of WBC for Sri Lanka is compatible or not.

### **1.8 Evaluation**

The built data model will be evaluated in this step. And check the accuracy level by looking at the outcome of the evaluation.

## **1.9 Structure of the Dissertation**

After presenting an in-detail description and explanation about the basic domain and scope of the study in the 1<sup>st</sup> Chapter, the next chapters of this thesis contain descriptions of the research in much deeper as mentioned below.

The second chapter covers details about the literature review of the basic domain, similar approaches that were referred for this study. The current knowledge and new methods concerning the research.

The third chapter covers the details about the technologies adapted for this research. software technologies, Statistical analysis technologies, data mining, and machine learning technologies such as classification techniques. will be discussed in this section.

The fourth chapter explains the methodology adopted to achieve the objectives of the research. How the data set is generated and which actions were taken for the process of feature extraction and how they are integrated with new ML algorithms, are described in this chapter.

The fifth chapter will be described the evaluation and the results which are obtained by this study. It explains the improvements required and the accuracy of the outputs of this research.

In the final chapter, the conclusion and the further works of the research are described by evaluating the whole research effort.

# Literature Review

## 2.1 Introduction

This chapter gives a clear idea about what other countries have done to find domestic hematology referential range for their countries via a literature review. Most researchers have successfully introduced domestic hematology referential range for their countries which is different from standard hematology referential range.

## 2.2 Similar Approaches Based on statistical Analysis

A domestic hematology reference range has been determined for Malawi people successfully. Malawians are not included in the Caucasian population. The research was conducted to determine white blood cell (WBC) count, mean corpuscular volume (MCV), hemoglobin (Hb), hematocrit (Hct), platelet count of healthy Malawians from birth to adulthood. The blood donors were identified as healthy or unhealthy before the research started by using a questionnaire, health reports/ records, body mass index (BMI), etc. Hb, WBC, platelet count, MCV was determined by using HMX hematological analyzer using the sample collected in the EDTA tubes from blood donors. Mann-Whitney test was used to decide the significance of the observations statically in blood report data collection. Hb, Hct [2]<sup>[10]</sup>. The considered sample out of the population is 660 (344 female and 316 male) with twelve different age groups. This may be not sufficient to make good decisions.

Another study was done to determine an age-specific domestic hematology reference range for healthy males in Eastern India. The blood donors included in the male adult population (ages between 20-59) of West Bengal, Bihar, Assam, Orissa, and other states of Eastern India were selected by conducting a rigorous screening through interviews and hemoglobin measurements. Collected blood samples were analyzed using an automated hematology analyzer (Wipro LabLife). Age group-specific variation of hematological parameters was evaluated by one-way analysis of variance (ANOVA) for independent samples, and variation

was shown by box-and-whisker plots. Statistical difference between the obtained mean and international data for each parameter was compared by the Chi-square test. This population exhibited lower platelet (PLT) and hemoglobin (HGB) counts as compared to the standard reference values. But this inequality was statistically significant only for the count of platelets. However, the digression from the standard hematology reference range of data was clinically important, except for the white blood cell (WBC) counts and red blood cell (RBC) counts [6]. The sample (528 blood samples) which is taken to conduct the research is small, hence the data set is not sufficient to make a good decision.

In another research, a domestic hematology reference range for White Blood Cells Count was determined in Sudan by considering 1076 healthy Sudanese adults from both sexes, with the age range of 20 – 60 years. Blood donors who are suffering from chronic diseases (cardiac diseases, TB, asthma, thyroid disorders, diabetes mellitus, hypertension, renal failure, liver diseases, etc.), recent acute diseases (malaria, typhoid fever, etc.), recent surgery, drug abuse, pregnancy, lactation, and heavy smokers were excluded. A Sysmex KX-21 automated hematology analyzer was used for measuring WBCs and differential counts. To determine the inequalities between groups for continuous and abnormally distributed variables, the medians were compared using the Mann-Whitney U test. The result showed that WBCs count was positively correlated with elevated BMI value and the hematology reference values of WBCs count in adult Sudanese are lower than the standard one [8]. The sample data set is bigger than the other researches. (Here 1076 blood samples were taken)

Local hematological reference values were determined for healthy adults in Togo in 2008. They had used a Sysmex SF-3000 automated hematology analyzer to perform a whole-blood analysis of hematological parameters. The standard deviation, median and mean values were calculated for each of those parameters. To compare parameters according to gender, the Kruskal-Wallis test was used for two groups. It is a known fact that the platelet count of black people is globally less t[4].

The research was done to determine hematological reference values in Turkey, especially of the people living near the sea level. For blood analysis of hematological parameters, they used the Sysmex XT-2000i automated hematology analyzer. Complete blood count (CBC) parameters were examined for normal distribution via histograms, kurtosis, t-values (Skewness /SE Skewness), Q-Q normality, and Box plots, and then the significance levels were

determined from the Kolmogorov-Smirnov and Shapiro–Wilk tests of normality. Certain differences were observed when compared to the previously used and established values of Turkey, particularly in platelets and [10].

A domestic hematology reference range was determined in Nigeria in 2014. A complete blood count (CBC) and a differential were performed on the blood samples using Sysmex KX-21N. Sysmex KX-21N is an automated 3-segment [7] differential hematology analyzer. Reference ranges were calculated using nonparametric methods. The median values were calculated and reference values were determined at 2.5th and 95th percentiles. standard deviations, mean and median values were computed for each of the clinical chemistry and hematological parameters of the study subjects. Parametric student's t-test was exploited to determine significant differences between non-pregnant females and males; as well as pregnant and non-pregnant females. Researchers found slightly higher hematology ranges when compared them with the hematology reference ranges in the USA and suggested establishing reference levels for local populations and additional researches to validate these interesting findings [7].

In China, reference ranges were calculated for their population in different geographical areas. Three separate research groups in China independently studied to discover the relation between hemoglobin reference ranges and the geographical area in China of adult females and males. They also considered the effect of topographical locations in infants. In Saudi Arabia, they have established reference ranges for their population especially for adolescent boys and infants [9].

Similarly, several countries have been done many types of research to find the accuracy of applying standard blood reference range for the people of their country and to establish or suggest a domestic blood reference range for their region.

### **2.3 Similar Approaches Based on Machine Learning**

Most of the domestic hematology ranges are found based on statistical analysis. However, Machine Learning algorithms are also used for researches based on hematology. Most of them are used to detect hematology diseases. For predicting the disease according to the blood analysis reports, precise patterns that can be used to identify the diseases accurately should be

identified. Machine learning is the field that is responsible for constructing appropriate models for predicting an output based on previous data.

In one research a dataset was generated by analyzing blood samples and 28 related attributes were included in it. This dataset contained 4 major classes related to four different blood diseases as Thrombocytopenia (the lack of platelets), Leukocytosis (a boost of white cells above the normal range), Anemia (decrease in the number of red blood cells or hemoglobin in the blood) and Normal (all parameters values are normal). Cross-Validation is a type of statistical method of evaluating and comparing learning classifiers by partitioning data into two classes. Classifiers such as Naive Bayes, Multilayer Perceptron, Regression analysis, etc. were used to build the model [11]. Moreover, this research has shown the possibility of having each disease in the current state of health.

In other research is mentioned that all the diseases originate from or causes differences on a cellular and molecular level, and these changes are almost always detectable directly or indirectly by analyzing the changes of blood parameter values. These changes in the values can be large, and physicians can observe them by checking for blood parameter values that are not residing within the normal ranges. Machine learning models can recognize disease-related blood laboratory patterns that are beyond the scope of current medical knowledge. That will result in higher and better diagnostic accuracy compared to traditional quantitative interpretations based on the discovered reference ranges of blood parameters. Using a machine learning-based methodology in blood laboratory-based diagnosis would lead to a significant [12].

Artificial Intelligence is also used for analysis and predictions in Hematology. AI-based applications are helped to diagnose specific hematology diseases such as anemia, thalassemia, and leukemia. These are done based on neural networks trained with data from peripheral blood analysis. Also, applications of AI in medicine include devices applied to clinical diagnosis in neurology and cardiopulmonary diseases, as well as the use of expert or knowledge-based systems in routine clinical use for diagnosis, therapeutic management, and prognostic evaluation. Biological applications include genome sequencing or DNA gene expression microarrays, modeling gene networks, analysis, and clustering of gene expression data, pattern recognition in DNA and proteins, protein structure prediction. [13]

Also, Machine Learning and Artificial Intelligence can be used in automated blood film reporting. The steps are summarized below in another research paper. The first step is to digitize all blood films, which is already technically possible. Generate a vast library of normal and abnormal blood films needed for automation in this step. An intelligent system will also cross-check clinical records, biochemistry results, and pharmacy notes as the second step. Then it can be used to pattern recognition to decide whether human interference is required or not. Then create models based on large data sets and it can be an aid to prediction and risk stratification. Also, these models can be integrated with the many data points routinely collected. These can be improved the prediction regarding how laboratory parameters will develop such as para-protein levels and white blood counts in monoclonal gammopathy of undetermined significance (MGUS) or chronic myeloid leukemia (CML). [14]

In other research, there were built two models to predict a hematology disease. One predictive model used all the available blood test parameters and the other used only a reduced set that is usually measured upon patient admittance. Both models were produced good results and were obtained prediction accuracies of 0.88 and 0.86 when considering the list of five most likely diseases and 0.59 and 0.57 when considering only the most likely disease. The models have been recommended for practitioners. These models indicated blood test results with more information than physicians generally recognized. [15]

Another research was conducted to compare the existing methods such as traditional methods, AI-based diagnostic systems, etc. for discriminating anemia (IDA) and  $\beta$ -thalassemia trait ( $\beta$ -TT). The main goal of this study was to reduce the diagnosis time and cost for  $\beta$ -TT and IDA subjects by increasing their discrimination precision through the analysis of Complete blood count (CBC) indices. The data set contained 750 CBC tests which were obtained from the blood specimens of 390 males aged 20-35 and 360 females aged 17-32 years. The Pattern Based Index Selection (PBIS) was used to eliminate the redundant CBC indices from the input set leads to a more efficient system. The proposed new system through this research is based on a dynamic harmony search (DHS). [16]

## 2.4 Summary

Considering the above-mentioned studies, we can observe that many countries have already introduced a domestic hematology referential range for their countries. age, gender, genetics, attitudes, lifestyles, ethnic origin, dietary habits, geographical location, climate, environmental factors...etc. can be caused to this difference hematology range rather than getting a similar one for standard hematology referential range. And also, to address a large number of blood reports will be gained the efficiency of the research.

Most of them are done by using statistical analysis. But machine learning is also have used for various kinds of hematology analyses. The accuracy of both scenarios depends on the size and accuracy of the data set.

When considering the data set, it is not easy to find quality data. Because the data-id related to human records, so it is sensitive data. Hence the ethical consideration is highly expected.



# Technology Adapted

### 3.1 Introduction

This chapter describes the technologies used in this research. The technologies can be divided into many categories such as software technologies, Statistical analysis technologies, data mining techniques, machine learning technologies, etc. Taking an overall idea and being familiar with these techniques will be helped to understand and develop a methodology to succeed in this research.

### 3.2 Software Tools

Data can be stored manually in books or virtually in a database or else it can be store as a text file too. First of all, when the data set is found then we need to store it properly. Roughly go through the data set and arrange the data set as you want. Microsoft Excel is the best tool to do this simply. It helps you to arrange the data set and get a basic idea about it by applying various formulas.

Then for the real analysis part, there are several software that can be used in this kind of researches. Mainly it can divide into two as data mining tools and statistical analysis tools.

#### 3.2.1 Data Mining Tools

Data Science is getting a popular concept in modern days. So, scientists, researchers, and various kinds of sectors analyze data to take many benefits. Data analysis is a process that is applying tasks such as cleaning, transforming, and modeling onto the data and then discover useful information. And also, this analysis gives many advantages. A few of them are listed below.

- Ability to make faster, more informed decisions, predictions by using historical data.
- It can be identified with new trends.
- Best way to a deeper understanding of relationships between attributes.
- It can be used to identify upcoming risks.
- It can be helped to reduce costs and increase profit.

When considering Data Science, Data mining is also a popular topic. Data Mining is a subset of Data Analysis. A comparison of Data Mining and Data Analysis will be shown in *Table 1* [17]. Data mining helps to find hidden, valuable, and fully or partially useful patterns by considering large data sets. Data Mining can be discovered unsuspected and relationships which are unknown in previous stages by examining the data set. Also, Machine learning (ML), statistics, Artificial Intelligence (AI), and database technologies are used in Data Mining. *Figure 1* will be shown about the data mining process.

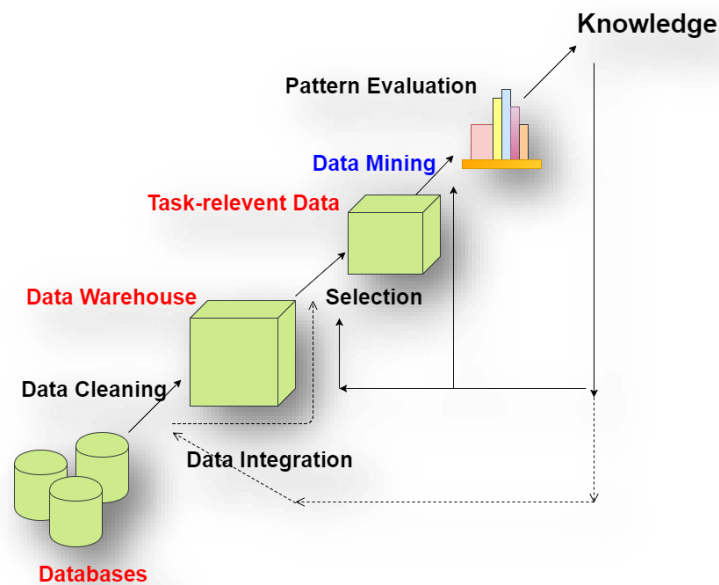



Figure 1 : Data Mining Process

<b>Data Mining</b>	<b>Data Analysis</b>
The process will help to find hidden patterns from large datasets.	The process will prepare raw data to determine useful decisions.
Machine learning, statistics, AI, and database technology are used.	Computer science, Information Technology, statistics, mathematics, subject knowledge, AI/Machine Learning are used.
Knowledge discovery from the data set.	Data Mining is a subset of data analysis.
The pattern is the output of data mining.	Hypothesis or insight on the data is the output of Data Analysis.
The hypothesis does not need to identify the patterns of the data.	The hypothesis needs and tests it.
Use Mathematical and scientific methods.	Use business intelligence and analytics models.
Mostly studies structured data.	Studies both structured, semi-structured, or unstructured data

Table 2: Data Mining Vs. Data Analysis

Several software tools can be used for Data Mining and the rest of this section will be talking about them [18].

<b>Name of the Software</b>	<b>Features</b>
Rapidminer 	<ul style="list-style-type: none"> <li>• Open Source Software.</li> <li>• It can be used for deep learning, text mining, machine learning &amp; predictive analysis.</li> <li>• Provide template-based frameworks and it helped to deliver the result with fewer errors quickly.</li> <li>• Have features to do workflow designing, prototype designing, data validation, build predictive data models, predictive analysis with Hadoop</li> </ul>

<p>Orange</p> 	<ul style="list-style-type: none"> <li>• Open Source Software.</li> <li>• Provide component-based interfaces.</li> <li>• Have features to do data visualization, built predictive models using steps among pre-processing to an evaluation of various algorithms.</li> </ul>
<p>Weka</p> 	<ul style="list-style-type: none"> <li>• Free Software.</li> <li>• Provide tools for data visualization and analysis using machine learning algorithms.</li> <li>• Have features for data mining, processing, visualization, regression, etc.</li> <li>• It can be used to data in a flat-file to build assumptions of data analysis.</li> </ul>
<p>KNIME</p> 	<ul style="list-style-type: none"> <li>• Open Source Software.</li> <li>• Use data pipeline technology.</li> <li>• Commonly use this tool for financial data analysis and business intelligence.</li> <li>• Provide fewer steps to pre-process the data for analysis and visualization.</li> </ul>
<p>Sisense</p> 	<ul style="list-style-type: none"> <li>• Licensed Software.</li> <li>• It can be used to work with small as well as large organizations.</li> <li>• Provide features to work with multiple data sources to build a common repository and generate reports.</li> </ul>
<p>Oracle Data Mining</p> 	<ul style="list-style-type: none"> <li>• Proprietary License.</li> <li>• Commonly use this tool for data classification, prediction, regression, and special analysis.</li> </ul>

Table 3: Data mining Tools with their features


Not only are these tools but also there are several data mining tools. When choosing a tool to apply your data mining techniques, you need to consider the availability of the software (open source, free, licensed... etc.), the performance of the tool, easy to use, compatibility/suitability with your task... etc.





So, by considering the above things, Weka is better to achieve the goal because it can be quickly switched between algorithms and train them on a portion of the dataset then compare the results without having to write much code. And also, it is free software.

### 3.2.2 Statistical analysis tools

Statistical analysis is used to analyze data to find patterns, trends, make predictions. The area of the Statistics is very huge and it contains various statistical methods to apply to data to get results. These statistical methods have a theory and most of them have specific equations. Applying the values which are taken from your dataset, into these equations helps to go to the final result. This can be done manually. But nowadays many tools already have an implementation for most of the common statistical methods. Using these statistical tools minimizes the error which can happen in manual calculations.

There are several statistical tools that people use to analyze data. *Table 4* will give you knowledge about these tools. [19]

Name of the Statistical Tool	Features
SPSS(IBM)  	<ul style="list-style-type: none"> <li>• SPSS is a commercial software.</li> <li>• The most widely used statistical software package within human behavior research.</li> <li>• Easy to compile descriptive statistics, parametric, and non-parametric analyses.</li> <li>• Easily generate graphical depictions of results through the graphical user interface (GUI).</li> <li>• Includes the option to create scripts to automate analysis, or to carry out more advanced statistical processing.</li> </ul>

<p>R (R Foundation for Statistical Computing)</p> 	<ul style="list-style-type: none"> <li>• R is free statistical software.</li> <li>• Widely used across both human behavior research and in other fields.</li> <li>• Toolboxes are available for a range of applications, which can simplify various aspects of data processing.</li> <li>• Has a steep learning curve, requiring a certain degree of coding.</li> </ul>
<p>MATLAB (The Mathworks)</p> 	<ul style="list-style-type: none"> <li>• MATLAB is commercial software.</li> <li>• This is an analytical platform and programming language that is widely used by engineers and scientists.</li> <li>• The learning path is steep than R software.</li> <li>• A plentiful number of toolboxes are available.</li> </ul>
<p>Microsoft Excel</p> 	<ul style="list-style-type: none"> <li>• Microsoft Excel is commercial software but it collaborates for free with an online version.</li> <li>• Offer a wide variety of tools for data visualization and simple statistics.</li> <li>• Simple to generate summary metrics and customizable graphics and figures.</li> </ul>
<p>SAS (Statistical Analysis Software)</p> 	<ul style="list-style-type: none"> <li>• SAS is open-source software.</li> <li>• It offers options to use either the GUI or to create scripts for more advanced analyses.</li> <li>• Widely used in business, healthcare, and human behavior research.</li> <li>• Possible to carry out advanced analyses and produce publication-worthy graphs and charts.</li> </ul>

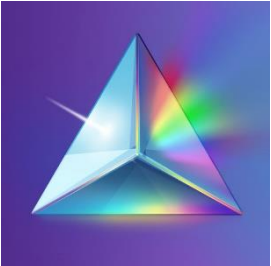

<p>GraphPad Prism</p> 	<ul style="list-style-type: none"> <li>• GraphPad Prism is a commercial software.</li> <li>• Primarily used within statistics related to biology, but offers a range of capabilities that can be used across various fields.</li> <li>• Scripting options are available to automate analyses or carry out more complex statistical calculations as similar to the SPSS.</li> </ul>
<p>Minitab</p> 	<ul style="list-style-type: none"> <li>• Minitab is commercial software.</li> <li>• Offers a range of both basic and fairly advanced statistical tools for data analysis.</li> <li>• Commands can be executed through both the GUI and script as similar to the GraphPad.</li> </ul>

Table 4: Statistical Analysis tools with their features

Selecting a tool depends on a range of factors, including your research question, knowledge of statistics, and experience of coding. So, you need to consider the following things when selecting a statistical tool. [20]

- Most of these tools support common statistical methods. But sometimes it can be needed some specific methods for your analysis which is not common. So, the first task is identifying exactly what methods you need to apply in your research. Then find what are the tools that offer those statistical methods. It will save you time.
- Some tools might give good results, but they can't understand how and why those results were reached. It courses to rebuild the statistical methods and the model to understand the model and the system better. But it might not be user friendly. For example, Microsoft Excel does not give such insights with their results. So, the user needs to spend more time to understand the result even he uses a tool ta analyze. So, the way of the result representation, what are the insights give with the result will be needed to consider when you select a statistical software.

- When you select a statistical tool, you need to consider the popularity of the tool. Popularity means that the tool is a common one. Hence more people try it, got results, examine the results, and probably reported problems if occurred, discussions for various errors, problems might be published in online forums, has sufficient reviews and feedbacks about the tool. So, using such a kind tool will be easy as the user has enough resources to handle the tool if any problem occurs. So, when selecting a tool, try to select a tool that releases a few years ago (it means do not select the newest tool), also the tool must be maintained by a reputable organization and check the availability of forums, blogs, and literature about the tool.
- Check whether the tool has comprehensive and helpful official documentation. It will be helped to install, configurations, applying a statistical method on the dataset.
- Every software release to do a specific task. And sometimes some features add optionally. For example, suppose a software “Y” releases to do X type tasks. So, if you have to do tasks similar to the X type then software “Y” is the best option for you.
- Suppose that you need to get data from the database to the statistical analysis tool. In such a case if your selected statistical tool does not support databases then you need to export the data or do a similar scenario. Also, suppose if you want to publish the results of the analysis in a web application or if you want to build an application based on the result. Similarly, if your selected tool does not support developing an application then it will be hard to work with several software. So, before you select your statistical tool, you need to give attention to this point too.
- When you buy a commercial statistical tool, you need to give your attention to the feature list of each category. It will be helpful to select the exact tool for you. Some commercial software has categories such as student, academic, and business...etc. The offered features are different in each category. Using the student category tool for the business purpose may be illegal and also you will be miss high-level features that are offered in the business category. So, you need to look at the features that offer, policies, legal backgrounds before and after the production when you choose your tool.



- If you familiar with any statistical tool in your academic period or else it is easy to work with it as you know about it. And also, if the supporting programming language for the tool is also familiar to you will be an added advantage too. But sometimes there may be other tools which can be used to do the same analysis very easily. So, you need to consider this point too before you select your tool.

So by considering the above things, the SPSS tool is selected for the statistical analysis part. It supported to techniques mentioned in literature reviews, user friendly, generating graphs is easy.

### 3.3 Statistical Techniques

In Chapter 2, we talk about some statistical tests which are used to analyze the hematology referential values under statistical-based similar approaches. Those tests are discussed under this topic.

#### 3.3.1 Mann-Whitney U test

This test can be categorized under the non-parametric test. Mann-Whitney U test uses to compare two independent groups. Also, this can be used to test the null hypothesis of two groups which is taken from the same population. Here the important point is that both two groups have a similar shape/curve in graphs. And the two groups do not lie in the normal distribution.

Suppose R is the sum of ranks in the sample, and n is the number of items in the sample. Then one of the following formulas can be used to calculate Mann Whitney U Test. If the sample is small then these formulas can directly use. But if the sample is large then applying this formula manually will be coursed to occur errors. So, it is easy to use statistical software which offers to the calculation of U statistic such as SPSS when the sample is large.

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} \quad \text{or} \quad U_2 = R_2 - \frac{n_2(n_2 + 1)}{2}$$

Following all four assumptions or a number of them should need to hold by your dataset to apply this test. [21]

1. Both groups do not lie in normal distribution and should have the same shape/curve. For example, both have bell-shaped and skewed left.
2. Both groups do not hold a relationship between them. This means that the groups should be independent.
3. The output variable which is also called an independent variable should have only two independent options. For example, gender can select as an output variable and it contains only two independent options as male or female. A variable that contains “yes or no” also select as the independent variable.
4. The dependent variable must be computed by an ordinal scale or a continuous scale.

### 3.3.2 Chi-square test

The Chi-square test is regularly used to check patterns of a set of frequencies. There are two types of Chi-square tests as the goodness of fit test by using multinomial tables and independence tests by using contingency tables.

The Chi-square goodness of fit test is used to find patterns mainly. But the most common Chi-square test is the second one, the Chi-square independence test. It is used to identify the dependencies between two classification variables. Hence, many surveys have used this test to analyze their dataset [22]. The common chi-square formula is given below.

$$x^2 = \sum \frac{(O - E)^2}{E}$$

Here  $x^2$  is the test statistics, O is the observed value and E is the expected value. If there is no relationship between the observed value and the expected value among the dataset, then  $x^2$  is a single number. But if the observed value and the expected value are the same, then  $x^2$  is equal to zero. So, if you get a value for  $x^2$  which is almost near to the zero, then there is a high correlation among your data.

This test is also offered by several statistical tools such as SPSS, MS Excel and it will be helpful to analyze without errors which can appear in manual calculations.

### 3.3.3 Kruskal-Wallis test

This also calls as a nonparametric test. This test can be used instead of a one-way ANOVA test. Kruskal-Wallis test uses to find the difference between the medians of two or more than two samples. The data which is used for the analysis does not need to distribute normally. Hence the frequency curve does not need to symmetric, it can be skewed left or right.

The test statistic of this test is called the “H statistic”. The rank of the data values is used to take this H statistic instead of using actual data values. Let assume there are ‘c’ number of samples and ‘n’ is the sum of all sample sizes. ‘T<sub>j</sub>’ denotes the sum of ranks in the j<sup>th</sup> sample and ‘n<sub>j</sub>’ denotes the size of the j<sup>th</sup> sample. So, the below formula helps to calculate the H statistic manually.

$$H = \left[ \frac{12}{n(n+1)} \sum_{j=1}^c \frac{T_j^2}{n_j} \right] - 3(n+1)$$

Several statistical tools such as SPSS, Minitab offer this test to analyze the dataset, so it helps to prevent errors which are occurred in the manual calculation.

This also has some assumptions which need to satisfy before you apply this test to your dataset.  
[23]

1. The independent variable should have two or more independent options.
2. The dependent variable must be computed by ordinal, ratio, or interval scale.
3. Similar to the Mann-Whitney U test, here also both groups do not hold a relationship between them. This means that the groups should be independent.

- Both groups do not lie in normal distribution and should have the same shape/curve.

### 3.3.4 Kolmogorov– Smirnov test

This test is also a non-parametric statistical test. The test is introduced based on the empirical distribution. There are two types of Kolmogorov-Smirnov test as the one-sample Kolmogorov-Smirnov test and the independent-sample Kolmogorov-Smirnov test. The first one is used to test whether a variable follows a particular distribution or not in a population. Here the particular distribution does not mean normal distribution. It can be any continuous distribution and could not apply with discrete distributions. The second one is used to test whether a variable has an identical distribution or not in 2 populations [24] [25].

Here the test statistic is called “D statistic”. Let assume there is ‘N’ number of ordered data points as  $Y_1, Y_2, Y_3 \dots Y_N$ . The points are ordered from small to large. Then the basic Kolmogorov-Smirnov formula can write as follows. ‘F’ is the cumulative distribution of the distribution which is going to be tested.

$$D = \max_{1 \leq i \leq N} \left( F(Y_i) - \frac{i-1}{N}, \frac{i}{N} - F(Y_i) \right)$$

Statistical tools such as SPSS also offer this test and it prevents errors that can happen in manually.

### 3.3.5 Shapiro–Wilk test

This test can be categorized as a semiparametric test. The test is used to find whether a random sample is related to normal distribution or not. Here the test statistic is called ‘W statistic’ and it can calculate by using the following formula.

$$W = \frac{\left( \sum_{i=1}^n a_i x_{(i)} \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Let assume that we have 'n' size random sample. Then 'x' denotes the ordered random sample value and 'i' denotes its position. 'a<sub>i</sub>' denotes a constant value which is calculated by considering mean, variance, and covariance of the sample. If you get a small value for the W statistic then it means that the sample is not distributed normally. Several statistical tools such as SPSS, Minitab, R, Excel, MATLAB, SAS also support this test [26] [27].

### **3.3.6 t-test**

This is also called the Student's T-test and it is a parametric test. The T-test can be used to check the significant differences between the two samples. The T-test can be applied if the dataset follows the normal distribution. There are three types of T-test that people use to analyze their data commonly [28] [29].

1. Independent samples T-test: This can be used to compare the means of two samples.
2. Paired sample T-test: This test can be used to compare means inside one sample but a different period.
3. One sample T-test: This test can be used to compare the means in one sample according to the known mean.

## **3.4 Machine Learning Classification Algorithms**

Classification is the most important part of the data analysis. It can be applied to both structured and unstructured datasets. Simply classification technique categorized our dataset into classes that we have given by considering other data patterns. And then we give another dataset and check how much data can categorize correctly (*Figure 2*). Some terms are used usually when considering machine learning classification algorithms.

Classifier: There are several algorithms related to classification and when considered individually, each algorithm is a classifier.

Classification Model: When you apply a classification algorithm for your dataset, it separates your data point into given classes. This is the model that you create using a classification algorithm. After you supply the test data set into this model then it predicts the data points' class. This is the purpose of building a model.

There are two types of classification algorithms that are mostly used for data analysis.

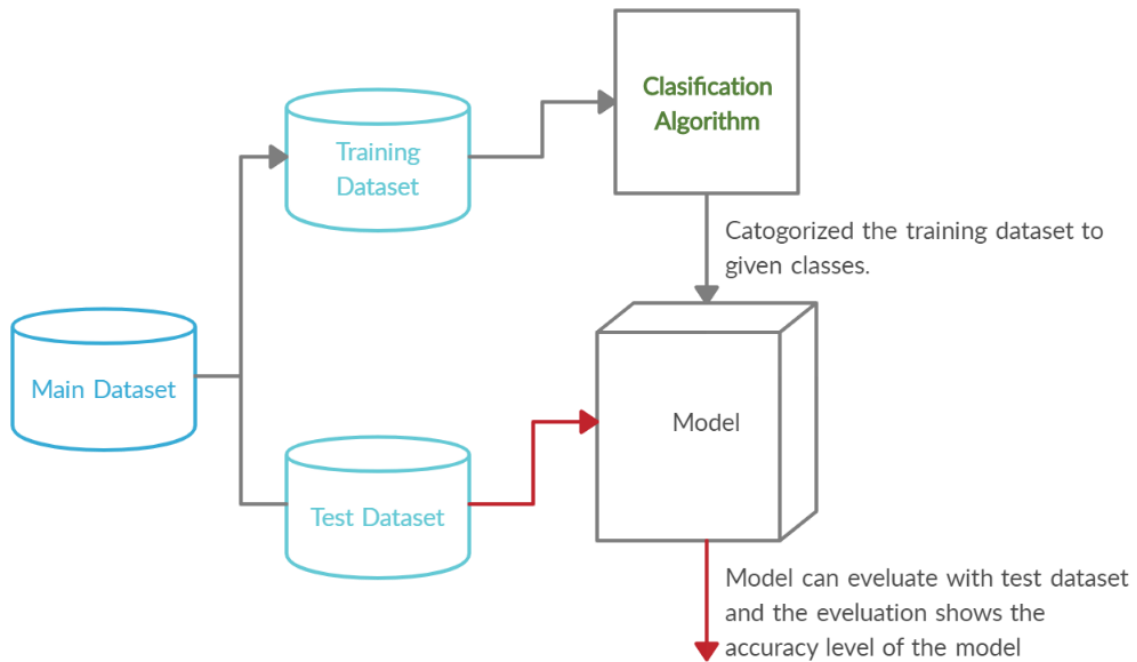


Figure 2 : Classification algorithm

1. **Binary Classification:** There are only two given classes in this classification. For example, there can be two classes with yes/no, male/female, positive/negative...etc.
2. **Multi-class Classification:** There are available more than two given classes. For example, there can be tree classes with sunny/windy/rainy. But in each time one data point belongs to only one class. one data point cannot belong to several classes.

There are 4 main steps that you need to follow when applying a classification algorithm. (Figure 3)

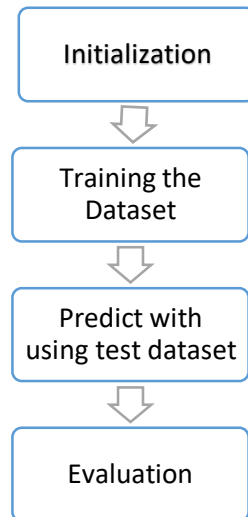


Figure 3: Steps of Classification algorithm

#### Step 1: Initialization

You need to select your algorithm and do the basic configurations according to your requirements in this step.

#### Step 2: Training the Dataset

Train your dataset by using the algorithm which is initialized in step 1. You can build a data model end of this step.

#### Step 3: Predict with using the test dataset

The accuracy of the model can be measure in this step.

#### Step 4: Evaluation

The accuracy of the prediction can be evaluated in this step.

There are several classifier algorithms available for data analysis and most of these algorithms are offered by many data mining tools. Some most popular algorithms are listed below.

- Logistic regression algorithm
- Naive Bayes algorithm
- SVG (Support vector machines) algorithm

- Least squares support vector machines
- Kernel estimation algorithm
- k-nearest neighbor algorithm
- Decision tree algorithm
- Random forest algorithm
- J48 algorithm

A few of them are briefly explained below [30].

### **3.4.1 Random Forest classification Algorithm**

This is a supervised learning algorithm categorized under the decision trees. First, the training data set is converted to the subsamples. Here each subsample's size is similar to the original training data set size. But the place of the data point is changed. Next, it builds several decision trees on the training dataset and takes the predictions from each tree. The final solution is delivered by considering the average of each prediction.

The main advantage of using a Random Forest algorithm is controlling the overfitting of the data set with predictions. Also, this is accurate than other algorithms which are categorized as decision trees. But the process of generating prediction is slower. And if you write code for this algorithm then it will become difficult and complex. But the number of data mining tools such as Weka, provide this algorithm without coding.

### **3.4.2 Decision Tree Algorithm**

This is also a supervised learning algorithm and helps to build models for prediction purposes. When you apply this algorithm to your dataset, first it categorizes your dataset in various ways. Next, it defines the rule set by considering categories and those rule sets help to prediction.

There are several advantages to the decision tree. Mainly this algorithm is easy to understand and can be visualized easily. It does not require heavy data preparation before applying this algorithm. Also, you can use this algorithm to analyze numerical data as well as categorical data.



### 3.4.3 Naïve Bayes Algorithm

The Bayes Theorem is the base of this algorithm. When you apply this algorithm to your dataset, it assumes that the features of a given class are independent of other features in the same class. The naïve Bayes algorithm is very popular among the data analysts and this is the perfect algorithm to analyze documents.

Fast is the main advantage of this algorithm. Also, this does not require large training data set to build the model. But the naïve Bayes algorithm considers as a bad estimator, so it will be a disadvantage.

### 3.5 Features used to measure the accuracy of the model

Once you create a model you need to clarify the accuracy of your model. You can decide the accuracy of the model you have built by looking at the number of features. This section will take about those things based on the weka output.

Some key components can be used to get an idea about your model. Some of them are listed below.

- **Correctly Classified Instances and Incorrectly Classified Instances:**  
This shows how many instances are correctly classified by the model and how many instances are incorrectly classified by the model. it gives numerical value as well as the percentage. So, if it shows a higher value for the Incorrectly Classified Instances than Correctly Classified Instances then the model you have built is not good. And also, if it shows Correctly Classified Instances as 100% then it also not a better outcome as your data over-fitting to the model. so normally it is better to have a value between 80 to 100 for Correctly Classified Instances.
- **Confusion Matrix:**  
Confusion matrix also generates according to the Correctly Classified Instances and Incorrectly Classified Instances. The matrix size depends on the number of options in

the output class. For example, if the output class has only two options then it generates 2 by 2 matrix.

<i>option_1</i>	<i>option_2</i>	
<i>a</i>	<i>b</i>	<i>option_1</i>
<i>c</i>	<i>d</i>	<i>option_2</i>

In here,

$a + d$  = value of Correctly Classified Instances

$b + c$  = value of Incorrectly Classified Instances

- **Kappa Statistic:**

This is also a good measurement to check the accuracy of your model. Simply it shows the accuracy of classifying into the correct class when you consider any random data point. This value is generated by matching expected accuracy with observed accuracy. The following formula uses to calculate the kappa statistic.

$$\text{kappa statistic} = \frac{\text{observed accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}}$$

The following lines will show you how to get an idea by looking at the kappa static value.

- Kappa statistic < 0 means there is no agreement with accuracy.
- $0 < \text{kappa statistic} < 0.20$  means that the accuracy is slight.
- $0.21 < \text{kappa statistic} < 0.40$  means that the accuracy is fair.
- $0.41 < \text{kappa statistic} < 0.60$  means that the accuracy is moderate.
- $0.61 < \text{kappa statistic} < 0.80$  means that the accuracy is substantial.
- $0.81 < \text{kappa statistic} < 1$  means that the accuracy is perfect.

- TP rate :

This means True Positive Rate and it is also a numerical value. It gives how many instances are correctly classified into the classes. The following formula uses to calculate the TP rate.

$$\text{TP rate} = \frac{\text{True positive instances}}{\text{Total number of instances}}$$

- FT rate:

Opposite of the TP rate. This means False Positive Rate and it is also a numerical value. It gives how many instances are incorrectly classified into the classes. The following formula uses to calculate the FP rate.

$$\text{FP rate} = \frac{\text{False positive instances}}{\text{Total number of instances}}$$

- Precision:

This talks about how many selected items are relevant to the given class. It shows a proportion of instances that are truly inside the class. The value can take by using the following formula.

$$\textit{precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

- Recall:

This talks about how many relevant items are selected in the given class. It shows the proportion of instances that are classified inside the class. The value can take by using the following formula.

$$\textit{Recall} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False negative}}$$

- F-measure:

This is a value that is taken by considering precision and recall. It shows the connection between the low false positives and the low false negatives. So, it is better to get a value near 1 for this and if you take a value near 0 then the model is quite bad. Following formula uses to calculate F-measure.  $2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$

$$F - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- ROC area:

The meaning of this is the ‘Receiver Operator Characteristic’ area. It is better to receive a value of more than 0.5. The curve can be visualized by using the ‘visualize threshold curve’ option.

### 3.6 Summary

A surround knowledge about the technologies which can be used in this project is covered from this chapter. This chapter talks about data mining and statistical tools and compare them and give an overview of them. Next presented various statistical tests with explanations. It will be helped to choose the right test for the data analysis. Then talk about machine learning classification. How the classification works, what are the steps, what are the algorithms, advantages, and disadvantages of those algorithms are talk under this section. Finally, talk about how to check the accuracy of the model by using various features.

# Methodology

### 4.1 Introduction

This chapter describes the research methodology of the dissertation. The previous three chapters are the basis of this chapter. So, this will present how the technologies and the approaches are combined with the success in the research. Also, the knowledge which is earned from the Literature Review chapter is the most important thing in this chapter.

How to find the dataset, ethical consideration of the dataset, how to prepare the data set, how to build a model from the data set, how to analyze the data set statistically, how to work with association rules...etc. will be covered under this chapter.

### 4.2 Abstract View

The research will be done by using data mining and machine learning concepts. An abstract view of methodology as follows.

1. Find a suitable data set.
  - Data will be collected from Sri Lanka laboratories. Mainly it should contain values according to various hematological parameters such as hemoglobin level, platelet count, white blood cell count, etc., and information such as age, gender, test type, the report issued to date, etc.
  - Personal data such as name, the address will be not collected to prevent ethical and privacy issues.
2. Generate a healthy state for the data set.
3. Divide the data set into two sets as a training set and test set.
4. Prepare data set for processing of a machine learning algorithm
5. Find a suitable machine learning algorithm and build a data model by using the training data set. The model will be trained to identify healthiness or unhealthiness.

6. Test the model using the testing data set.
7. Applying clustering techniques to find the ranges that the healthy values lying.
8. Do Statistical data analysis.

### 4.3 Generating the Dataset

This was the major task of this research. It can be divided into the subtasks,

- The selection of the sample
- Ethical considerations related tasks
- Prepare the Data Set

These sub-tasks are described in the next sections.

#### 4.3.1 Selection of the sample

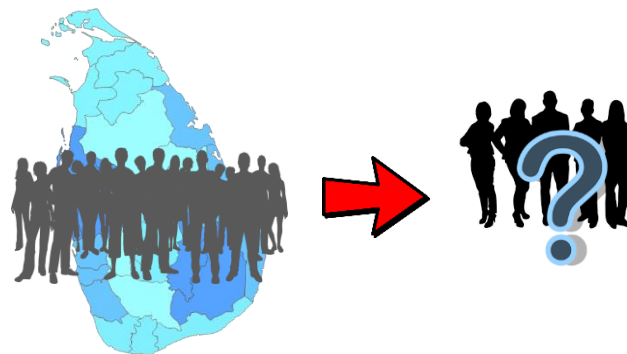


Figure 4 : Selection of the sample

This section explains how to select the sample for the research. Sri Lanka has over 21 million population with people in different age ranges. So, it is very hard to consider the whole population for this research. Hence selecting an appropriate sample is mandatory.

According to the scope, the research considers only the adults in Sri Lanka. A person whose age is over 21 can be categorized as an adult. So being an adult is the first criterion to select the sample.

There are several medical organizations such as hospitals, private medical centers in Sri Lanka. Let us considered few leading medical organizations for our research. Then we can find blood reports data from these organizations.

Next, let us consider the medical domain. The whole project will be run on a data set that is taken from the hematology reports. There are two methods to take hematology reports as follows.

- Interview the selected sample out of the population and select suitable people for blood donation. Generate blood reports after the blood donation and take the data to create a relevant data set.
- Take past blood reports which are stored in hospital, blood bank...etc. and collect relevant data from these reports and create a data set.

It will take considerable cost when considering the first method. Most researchers who have done the same research had selected this method to collect data set. The count of the donors is not more than 1000. But when considering the second method, the cost is lower than the first method and can create a large data set to collect data through past blood reports.

#### **4.3.2 The ethical considerations**

Every research should need to clear ethical background according to their research area. Most research that works with human/animal data should need to get an ethical clearance report before collecting the data set. So, ethical approval is mandatory for this research because this research also works with human data. Therefore it is needed to take ethical approval before starting to collect data.

The analyzed data set holds reliability it can be considered as a trusted resource. The data set is stored as excel files that did not contain private information such as patient name, address. So, the data set to hold an ethical clearance.

### 4.3.3 Prepare the Data Set

There are various types of test which are used in hematology science such as full blood count test, C-reactive protein (CRP) test, liver function tests, thyroid function test...etc. Here we consider only full blood count (FBC) test reports as it is the most common test and it shows the main attributes of blood.

When considering the received data set, it contained details regarding various kinds of blood tests. And a single patient's data didn't store in a single row.

No	Sex	Age	Date of Birth	Approved At	Investigation	Test	Value
49	Female	1 years and 11	04 February 2018	04 May 2019	FBC	Platelet count value	277,000
50	Female	1 years and 11	04 February 2018	04 May 2019	FBC	Test Comments	
51	Female	1 years and 11	04 February 2018	04 May 2019	CRP (C-Reactive Protein)	CRP (C-Reactive Protein) value	1.3
52	Female	1 years and 11	04 February 2018	04 May 2019	CRP (C-Reactive Protein)	Test Comments	
53	Female	17 years.	04 May 2002	04 May 2019	Dengue (NS1) Antigen Ag	Dengue Antigen Ag value	Negative
54	Female	17 years.	04 May 2002	04 May 2019	Dengue (NS1) Antigen Ag	Test Comments	
55	Female	40 years.	04 May 1979	04 May 2019	TSH	TSH value	12.14
56	Female	40 years.	04 May 1979	04 May 2019	TSH	Test Comments	
57	Female	38 years.	04 May 1981	04 May 2019	FBC	Total Leucocyte Count (WBC) val	9,300
58	Female	38 years.	04 May 1981	04 May 2019	FBC	Neutrophils value	61
59	Female	38 years.	04 May 1981	04 May 2019	FBC	Lymphocytes value	31
60	Female	38 years.	04 May 1981	04 May 2019	FBC	Monocytes value	03
61	Female	38 years.	04 May 1981	04 May 2019	FBC	Eosinophils value	05
62	Female	38 years.	04 May 1981	04 May 2019	FBC	Basophils value	00
63	Female	38 years.	04 May 1981	04 May 2019	FBC	Erythrocyte (RBC) Count value	4.66
64	Female	38 years.	04 May 1981	04 May 2019	FBC	Haemoglobin (Hb) value	9.8
65	Female	38 years.	04 May 1981	04 May 2019	FBC	Packed Cell Volume (PCV) value	30.5
66	Female	38 years.	04 May 1981	04 May 2019	FBC	MCV (Mean Corpuscular Volume)	65.5
67	Female	38 years.	04 May 1981	04 May 2019	FBC	MCH (Mean Corpuscular Hb) val	21.0
68	Female	38 years.	04 May 1981	04 May 2019	FBC	MCHC value	32.1
69	Female	38 years.	04 May 1981	04 May 2019	FBC	Platelet count value	362,000
70	Female	38 years.	04 May 1981	04 May 2019	FBC	Test Comments	
71	Male	32 years.	04 May 1987	05 May 2019	Creatinine Serum	Creatinine Serum	1.4
72	Male	32 years.	04 May 1987	05 May 2019	Creatinine Serum	Test Comments	
73	Female	26 years.	04 May 1993	04 May 2019	UFR	APPEARANCE	Slightly Turbid
74	Female	26 years.	04 May 1993	04 May 2019	UFR	COLOUR	Yellow
75	Female	26 years.	04 May 1993	04 May 2019	UFR	SPE. GRAVITY	1.010

Figure 5: Sample of the received data set

A data row was generated for a single hematology attribute. So the main task was to identify a single person's FBC record and restore it as a patient record. *Figure 5* blocked area belongs to a single person's FBC record. Select the number of these kinds of blocks and re-prepare an excel file as *Figure 6*.



Patient ID	Age	Gender	Sample taken Date	Investigation	Total Leucocyte Count (WBC) value	Neutrophils value	Lymphocytes value	Monocytes value	Eosinophils value	Basophils value	Erythrocyte (RBC) Count value	Haemoglobin (Hb) value	Packed Cell Volume (PCV) value	MCV (Mean Corpuscular Volume) value	MCH (Mean Corpuscular Hb) value	MCHC value	Platelet count value	Health Code	Health State
1	67	F	2018.06.25	FBC	6500.00	67.00	29.00	3.00	1.00	0.00	4.75	12.40	38.20	80.40	26.10	32.50	307000.00	0	UH
2	79	F	2018.06.25	FBC	8800.00	52.00	1.00	1.00	1.00	0.00	4.36	8.80	28.10	64.40	20.20	31.30	167000.00	0	UH
3	82	F	2018.06.25	FBC	8400.00	46.00	45.00	2.00	7.00	0.00	3.97	11.40	32.30	81.40	28.70	35.30	381000.00	0	UH
4	55	F	2018.06.25	FBC	6800.00	52.00	42.00	3.00	3.00	0.00	4.33	12.50	36.90	85.20	28.90	33.90	227000.00	1	H
5	81	M	2018.06.25	FBC	14.00	90.00	6.00	2.00	2.00	0.00	4.07	12.10	36.40	89.40	29.70	33.20	273000.00	0	UH
6	28	F	2018.06.25	FBC	7000.00	62.00	35.00	1.00	2.00	0.00	2.57	5.70	18.90	73.50	22.20	30.20	364000.00	0	UH
7	47	F	2018.06.25	FBC	12300.00	37.00	59.00	2.00	2.00	0.00	4.52	13.90	40.40	89.40	30.80	34.40	253000.00	0	UH
8	69	F	2018.06.26	FBC	5700.00	59.00	38.00	1.00	2.00	0.00	3.66	10.90	33.60	91.80	29.80	32.40	236000.00	0	UH
9	56	M	2018.06.26	FBC	12700.00	67.00	30.00	1.00	2.00	0.00	5.39	15.80	44.80	83.10	29.30	35.30	232000.00	0	UH
10	34	F	2018.06.26	FBC	7900.00	58.00	37.00	3.00	2.00	0.00	4.76	14.20	41.50	87.20	29.80	34.20	309000.00	1	H

Figure 6: Re-prepare data set

Here the patient ID, sample taken date, and investigation fields are removed in the data analysis phase as those are unwanted attributes for the analysis.

#### 4.3.4 Generate Health Code/health State for the data set

There will be needed a class attribute for the data mining analysis. Therefore, health code and health state columns are generated by using the standard hematology referential ranges.

Here, the main task is finding the validity of WBC's standard hematology referential range for Sri Lanka. Hence the health state is generated by considering standard referential ranges which

Standard Hematology Reference Range		
Normal Value Range	min value	max value
Total Leucocyte Count (WBC) value	4000	11000
Neutrophils value	40	75
Lymphocytes value	10	45
Monocytes value	0	10
Eosinophils value	1	6
Basophils value	0	0.1
Erythrocyte (RBC) Count value	3.5	5.5
Haemoglobin (Hb) value	12	17.5
Packed Cell Volume (PCV) value	36	50
MCV (Mean Corpuscular Volume) value	80	96
MCH (Mean Corpuscular Hb) value	27	32
MCHC value	31.5	34.5
Platelet count value	150000	450000

Figure 7: Standard Hematology ranges which are consider to generate health code

are figured out in *Figure 7* but without considering the WBC's standard hematology referential value. A healthy state has two possibilities as 'H' for healthy people and 'UH' for unhealthy people. Health code depends on this health state and manually add 0 and 1 for 'UH' and 'H' respectively.

WBC is a special attribute in the blood. It depends on several other hematology attributes. So we can do assumptions based on these criteria. Let say X is a person. His all attributes which are considered in FBC reports (but except WBC as we go to get an idea about it) lie between the standard hematology referential ranges. Then X has a high probability to be healthy.

604 records are taken from the initial data for our task. The following points can be useful for further analysis.

- There are 477 people in the data set whose WBC value lies between the standard hematology referential range.
- There are 110 people in the data set whose generated health code equal=1.
- 90 people in the data set are actual healthy by considering all attributes' standard hematology ranges (With considering WBC hematology referential range too).

#### **4.4 Prepare training dataset and testing dataset**

Dividing the dataset into training and testing data will be useful when applying classification algorithms. So, it can be done as *Figure 8* shown below. The important thing is training dataset will be 2/3 of the total dataset and the testing dataset will be 1/3 of the total dataset. All of these two datasets need to contain actual healthy people, people whose health code =0, and health code = 1.

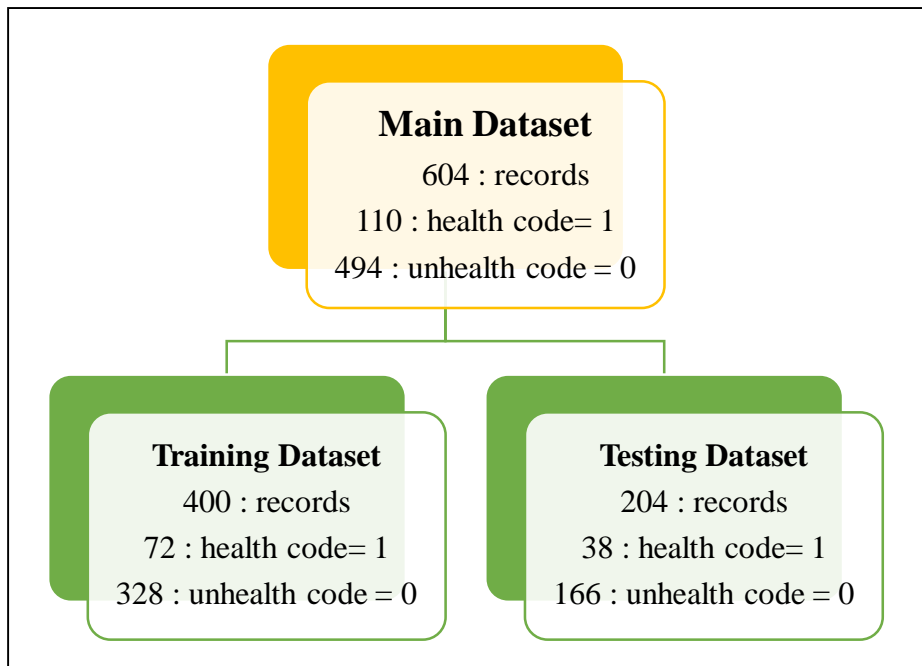


Figure 8: Training Dataset and Testing Dataset

Here the health code = 1 group contains both actual healthy people (by considering all standard hematology referential ranges) and the people who become healthy without considering the WBC's standard hematology referential value.

#### 4.5 Pre-processing of the dataset

Weka tool is selected and used for data mining purposes. It allowed pre-processing datasets easily. Afterload, the training dataset into Weka, preprocessing is the first task that needs to do. Here the main things that that I have done in this section.

1. First, check whether there are unique attributes or not in the dataset. If so, remove these unique attributes by using the remove option. There is a field called 'patient id' and remove this column from the dataset.
2. Next check whether there are any unwanted fields in the dataset. If so, it also needs to remove. Hence 'sample taken date', 'investigation', and 'health state' fields are removed from the dataset.

3. Most algorithms can apply to nominal type data fields. So next converted all data type to nominal.
4. Some fields can be grouped. Find those type fields and put them into the bin. In here age field is put into bins. Not only age but also other attributes also can put into the bin.
5. Set the class attribute. In here health code was the class attribute.

There is another useful feature inside this tab which can be used to visualize the data according to the columns. It can be used to take an overall idea about the shape of each data field by checking these graphs before putting the data into the bin. This will be helped to select the statistical test.

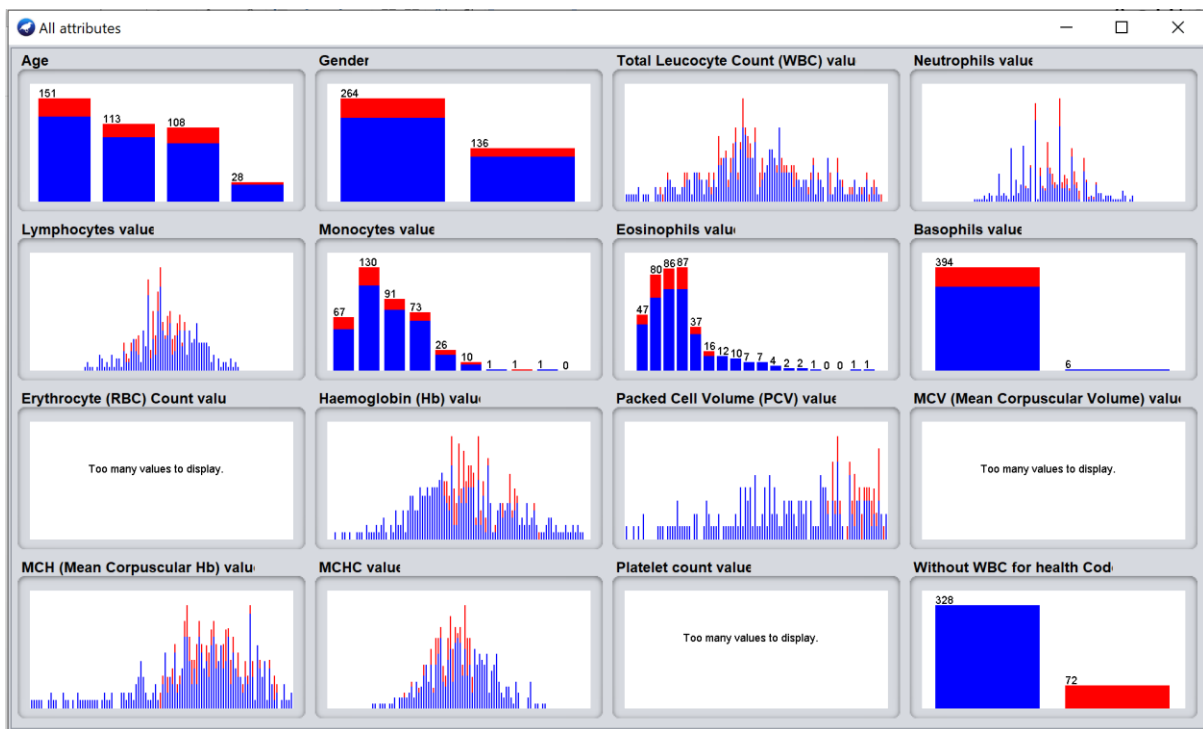


Figure 9: Data visualization in Weka pre-processing tab

Details such as how the values are spread in each attribute, how many options have in each attribute, what are shapes those data visualize can take by looking at these graphs. For example, we can get a rough idea about which graphs hold bell curve and which graph is s bell-shaped with skewed left or right.

#### 4.6 Build classification data model

Once you prepare your dataset then you can go to apply a classification algorithm. But it is needed to find which classification algorithm is suited for analyzing your dataset and achieve the final goal. In Chapter 3, we talk about machine learning classification algorithms. So, this is the place we apply those technologies.

Trees types algorithms are better to apply to build a model when considering all algorithm types. There are several tree algorithms in the WEKA tool as below.

- Random Forest
- Random Tree
- REP Tree
- J48
- Decision Stump
- Hoeffding Tree

Let's apply these algorithms to the training dataset to find the best outcome to build the model. The following table shows the feature values which can be needed to check the accuracy of the model.

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Kappa Statistic
RandomForest-using the training dataset	366	34	0.6884
RandomForest-using 10 folds cross-validation	331	69	0.3563
Random Tree- using the training dataset	366	34	0.6648
Random Tree- using 10 folds cross-validation	326	74	0.2705
REP Tree – using the training dataset	351	49	0.5139
REP Tree – using 10 folds cross-validation	329	71	0.216
LMT – using the training dataset	347	53	0.5437

LMT – using 10 folds cross-validation	332	68	0.3541
J48 – using the training dataset	357	43	0.584
J48– using 10 folds cross-validation	330	70	0.3099
Hoeffding Tree- using 10 folds cross-validation	327	73	0.1456

Table 5: Accuracy features of models -I

Table 5 shows correctly classified instances, incorrectly classified instances, and kappa statistics which are found while building the models. When considering these details, the following things can be summarized.

- The model built by the Random Forest algorithm using the training dataset and The model built by the Random Tree using the training dataset shows the highest number of correctly classifier instances.
- The model built by the Random Forest algorithm using the training dataset and The model built by the Random Tree using the training dataset shows the lowest number of incorrectly classifier instances.
- So, the model built by the Random Forest algorithm shows better performance according to the above two points.
- let's consider the Kappa Statistic. The model built by the Random Forest algorithm's kappa statistic value is higher than the model built by the Random Tree algorithm's kappa statistic value.

So, According to Table 5 contains details, the model built by the Random Forest algorithm using the training dataset is better than the other models. Appendix A contained full details of, the model built by the Random Forest algorithm using the training dataset.

Let's look at the other details which can be used to find the accuracy of the model.

Algorithm	TP Rate		FP Rate		Precision		Recall		F-measure		ROC Area	
	healthy	unhealthy	healthy	unhealthy	healthy	unhealthy	healthy	unhealthy	healthy	unhealthy	healthy	unhealthy
RandomForest-using the training dataset	0.667	0.970	0.030	0.333	0.828	0.930	0.667	0.970	0.738	0.949	0.971	0.971
RandomForest- using 10 folds cross-validation	0.403	0.921	0.079	0.597	0.527	0.875	0.403	0.921	0.457	0.897	0.854	0.854
Random Tree- using the training dataset	0.583	0.988	0.012	0.417	0.913	0.915	0.583	0.988	0.712	0.915	0.972	0.972
Random Tree- using 10 folds cross-validation	0.306	0.927	0.073	0.694	0.478	0.859	0.306	0.927	0.373	0.891	0.739	0.739
REP Tree – using the training dataset	0.472	0.966	0.034	0.528	0.756	0.893	0.472	0.966	0.581	0.928	0.917	0.917
REP Tree – using 10 folds cross-validation	0.208	0.957	0.043	0.792	0.517	0.846	0.208	0.957	0.297	0.898	0.844	0.844
LMT – using the training dataset	0.611	0.924	0.076	0.389	0.638	0.915	0.611	0.924	0.624	0.920	0.925	0.925
LMT – using 10 folds cross-validation	0.389	0.927	0.073	0.611	0.538	0.874	0.389	0.927	0.452	0.899	0.885	0.885
J48 – using the training dataset	0.542	0.970	0.030	0.458	0.796	0.906	0.542	0.970	0.645	0.937	0.931	0.931
J48– using 10 folds cross-validation	0.333	0.933	0.067	0.667	0.522	0.864	0.333	0.933	0.407	0.897	0.851	0.851
Hoeffding Tree- using 10 folds cross-validation	0.139	0.966	0.034	0.861	0.476	0.836	0.139	0.966	0.215	0.897	0.670	0.670

Table 6: Accuracy features of models -II

Table 6 show another set of features which can use to check the accuracy of the built model. It shows the TP rate, FP rate, Precision, Recall, F measure, and ROC area which are explained in Chapter 3. Let's look at the value set and search what can we extract from these details.

- TP rate tells the probability of an instance can correctly be classified. So when considering the model built by using the Random Forest algorithm with the training dataset has the highest TP value for both healthy and unhealthy classes. The weighted average of the TP value related to this model is 0.915.
- The model built by using Random Tree with a training dataset is also showed the same weighted average of the TP value. But When consider the TP rate for classified an instance as healthy is higher in the model built by using the Random Forest algorithm.
- When considering the FP rate, the model built by using the Random Tree algorithm has the smallest FP value for classified an instance as healthy. But when considering the weighted average of the FP value, the model built by using the RandomForest algorithm has the smallest value is 0.279.
- Precision value means how many selected items are relevant to the given class. So getting a high value for this is better. So, it is better to see the weighted average to get an idea about precision. The weighted average value for the model built by using the Random Tree algorithm shows the highest value as 0.915. And the weighted average value for the model built by using the Random Forest algorithm is 0.911.
- When considering the Recall value, the model built by using the Random Tree algorithm showed the highest weighted average value for the Recall as 0.915 and the model built by using the Random Forest algorithm showed the weighted average value is 0.815.
- F-measure value is the most important value when considering the accuracy of the model. When considering the model built by using the Random Forest algorithm, the F-measure value for healthy instances is 0.738 and the F-measure value for unhealthy instances is 0.949. The weighted average F-measure value is 0.911. It is pretty good. All values are near to value 1. When considering the model built by using the Random



Tree algorithm, the F-measure value for healthy instances is 0.712 and the F-measure value for unhealthy instances is 0.950. The weighted average F-measure value is 0.907. All the values are lower than the values shown in the model built by using the Random Forest algorithm.

- Another important thing we need to consider is the ROC area. When considering the model built by using the Random Forest algorithm, the ROC area value for both healthy instances and unhealthy instances is 0.971. The weighted average ROC area value is also 0.971. When considering the model built by using the Random Tree algorithm, the ROC area value for both healthy instances and unhealthy instances is 0.972. The weighted average ROC area value is also 0.972. Hence both algorithms value greater than 0.5, the models are pretty good. When considering the ROC area, looking at the threshold curves is also better before getting an idea.

The threshold curves belong to the model built by using the Random Forest algorithm are given below.

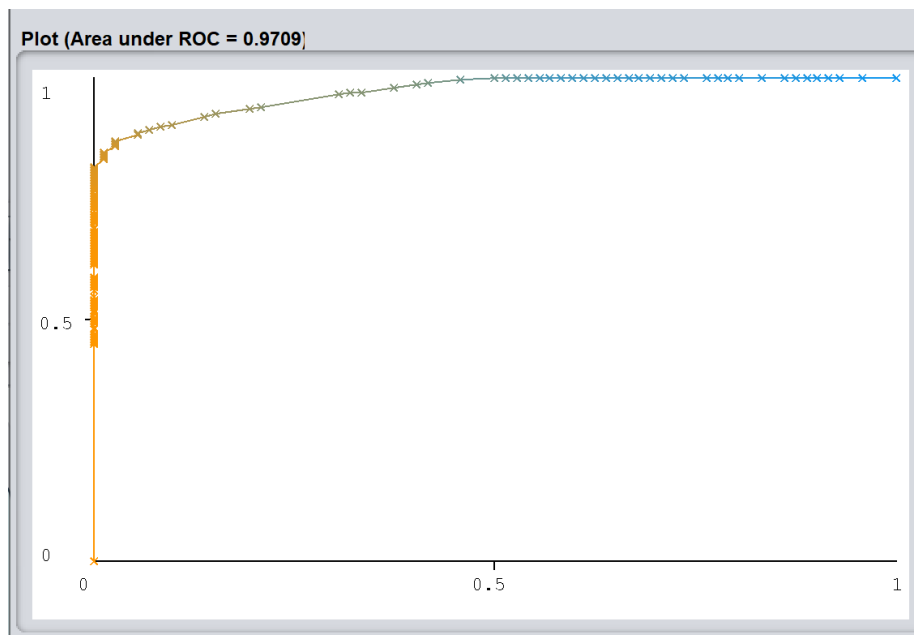


Figure 10: Threshold curve for unhealthy instances in the model built by using the Random Forest algorithm

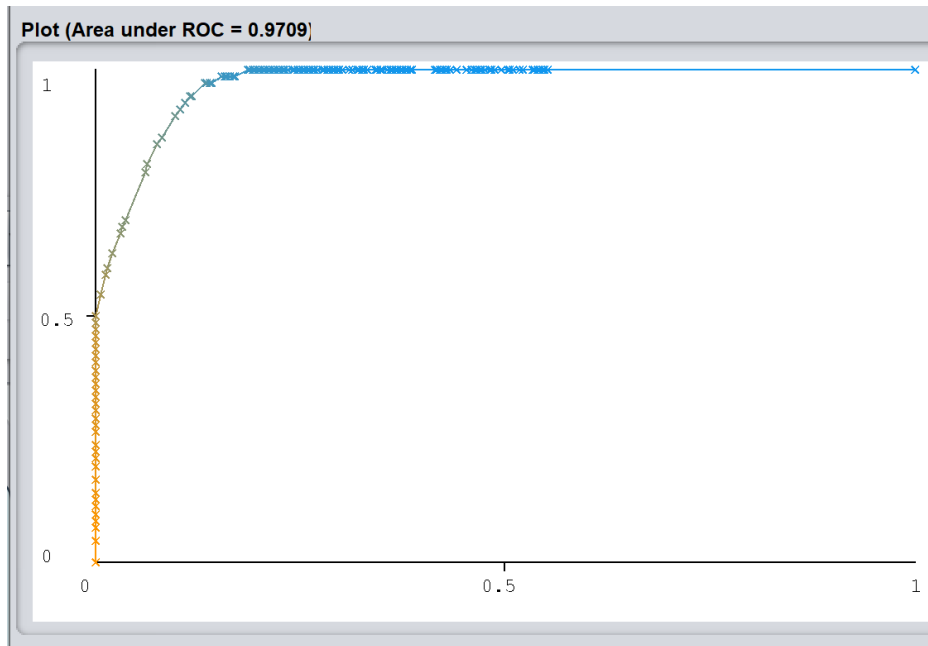


Figure 11: Threshold curve for healthy instances in the model built by using the Random Forest algorithm

The threshold curves belong to the model built by using the Random Tree algorithm are given below.

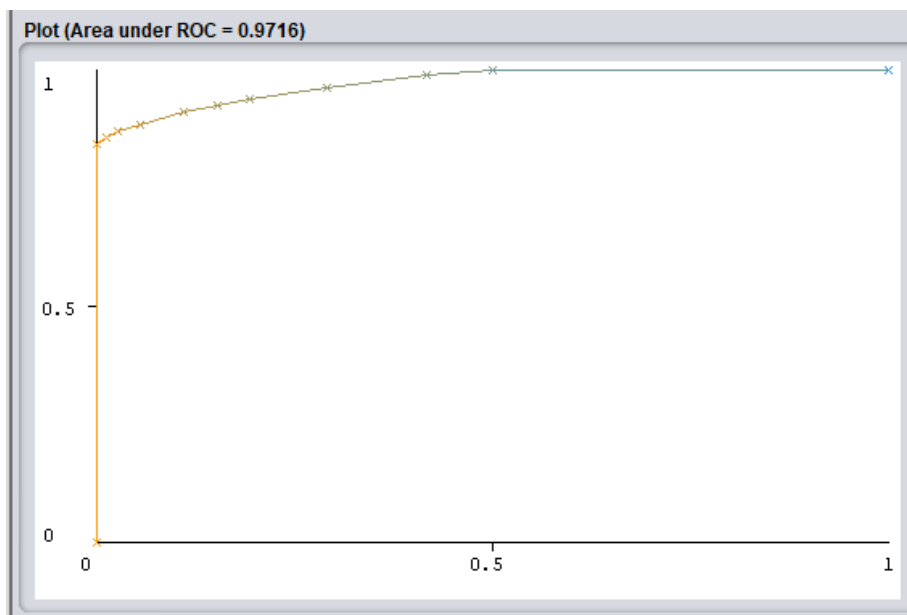


Figure 12: Threshold curve for unhealthy instances in the model built by using the Random Tree algorithm

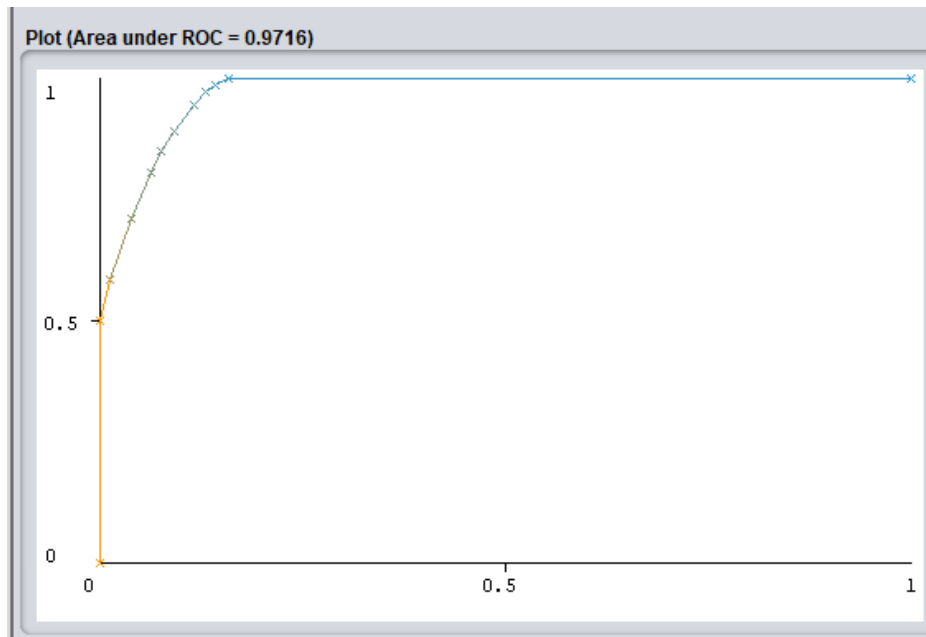


Figure 13: Threshold curve for healthy instances in the model built by using the Random Tree algorithm

All the curves are considered above is going near the X-axis. But the instances of the curves which are shown in Figures 10 and 11, pretty gather with X-axis.

- The Confusion Matrix for each model is shown in Table 7. By looking at the value of correctly classified instances, incorrectly classified instances, we know the percentage and the total number classified correctly or incorrectly. But the value belongs to incorrectly classified instances that do not show how many instances incorrectly classified as healthy and how many instances incorrectly classified as unhealthy. But this confusion matrix shows that detail very well.

When considering the confusion matrix belong to the model built by using the Random Forest algorithm, 48 healthy instances are classified as healthy and 24 healthy instances are classified as unhealthy. Also, 318 unhealthy instances are classified as unhealthy and 24 unhealthy instances are classified as healthy.

When considering the confusion matrix belong to the model built by using the Random Tree algorithm, 42 healthy instances are classified as healthy and 30 healthy instances are

classified as unhealthy. Also, 324 unhealthy instances are classified as unhealthy and 4 unhealthy instances are classified as healthy.

Confusion Matrix	
<b>RandomForest-using the training dataset</b>  <pre> a  b  &lt;-- classified as 318 10   a = 0 24  48   b = 1 </pre>	<b>RandomForest-using 10 folds cross-validation</b>  <pre> a  b  &lt;-- classified as 302 26   a = 0 43  29   b = 1 </pre>
<b>Random Tree- using the training dataset</b>  <pre> a  b  &lt;-- classified as 324  4   a = 0 30  42   b = 1 </pre>	<b>Random Tree- using 10 folds cross-validation</b>  <pre> a  b  &lt;-- classified as 304 24   a = 0 50  22   b = 1 </pre>
<b>REP Tree – using the training dataset</b>  <pre> a  b  &lt;-- classified as 317 11   a = 0 38  34   b = 1 </pre>	<b>REP Tree – using 10 folds cross-validation</b>  <pre> a  b  &lt;-- classified as 314 14   a = 0 57  15   b = 1 </pre>
<b>LMT – using the training dataset</b>  <pre> a  b  &lt;-- classified as 303 25   a = 0 28  44   b = 1 </pre>	<b>LMT – using 10 folds cross-validation</b>  <pre> a  b  &lt;-- classified as 304 24   a = 0 44  28   b = 1 </pre>
<b>J48 – using the training dataset</b>  <pre> a  b  &lt;-- classified as 318 10   a = 0 33  39   b = 1 </pre>	<b>J48– using 10 folds cross-validation</b>  <pre> a  b  &lt;-- classified as 306 22   a = 0 48  24   b = 1 </pre>

Table 7: Confusion Matrix

So, the model built by using the Random Forest algorithm classified the highest number of healthy instances as healthy correctly.

The model built by using the Random Forest algorithm can be select as the best model by considering all the above things.

## **4.9 Statistical analysis**

Statistical analysis is done by using the SPSS statistical tool. The analysis is done by using the original whole dataset. It contains actual healthy and unhealthy records by considering the standard hematology referential range.

### **4.9.1 Descriptive statistic**

Table 8 shows the descriptive analysis for healthy people in the dataset. It shows the minimum and maximum value of the data set, mean value, standard deviation value, variance value, skewness value, kurtosis value. Therefore following things details can be extracted from the table.

- When considering the total leukocyte value, the minimum value is  $4100\text{mm}^3$  and the maximum value is  $11000\text{mm}^3$ . The standard referential range is between  $4000\text{mm}^3$  and  $11000\text{mm}^3$ .
- When considering the neutrophil value, the minimum value is 46% and the maximum value is 75%. The standard referential range is between 40% and 75%.
- The Monocytes value is between 1% and 8% in this analysis. The standard referential range is between 0% and 10%.
- The lymphocyte value is between 20% and 45% in this analysis. The standard referential range is between 10% and 45%.
- The basophil value is always 0 in this analysis. The standard referential range is between 0% and 0.1%

**Descriptive Statistics**

	N Statistic	Range Statistic	Minimum Statistic	Maximum Statistic	Mean		Std. Deviation Statistic	Variance Statistic	Skewness		Kurtosis	
					Statistic	Std. Error			Statistic	Std. Error		
Total Leucocyte Count (WBC) value	89	6900	4100	11000	7775.17	177.586	1675.347	2806786.619	.026	.255	-.508	.506
Neutrophils value	89	29	46	75	59.30	.702	6.625	43.896	.382	.255	-.562	.506
Lymphocytes value	89	25	20	45	34.74	.646	6.095	37.148	-.426	.255	-.273	.506
Monocytes value	89	7	1	8	2.88	.162	1.529	2.337	.896	.255	.531	.506
Eosinophils value	89	5	1	6	3.19	.138	1.305	1.702	.202	.255	-.607	.506
Basophils value	89	0	0	0	.00	.000	.000	.000	.	.	.	.
Erythrocyte (RBC) Count value	89	1.460000000	3.930000000	5.390000000	4.507528090	.0355653597	.3355229328	.113	.611	.255	-.063	.506
Haemoglobin (Hb) value	89	3.6	12.0	15.6	13.227	.0945	.8916	.795	.858	.255	-.150	.506
Packed Cell Volume (PCV) value	89	9.400000000	36.00000000	45.40000000	39.28876404	.2647702099	2.497837165	6.239	.745	.255	-.432	.506
MCV (mean Corpuscular Volume) value	89	14.00000000	80.40000000	94.40000000	87.29775281	.3723561123	3.512800538	12.340	.281	.255	-.859	.506
MCH (Mean Corpuscular Hb) value	89	4.8	27.1	31.9	29.382	.1270	1.1983	1.436	.227	.255	-.552	.506
MCHC value	89	2.800000000	31.70000000	34.50000000	33.65955056	.0652980732	.6160207901	.379	-.796	.255	.323	.506
Platelet count value	89	257000	171000	428000	275707.87	6568.123	61963.553	3839481869	.389	.255	-.366	.506
Valid N (listwise)	89											

Table 8: Descriptive statistics of the dataset healthy people

It seems like all values exceeded its upper limit. That is why most of them's maximum value is similar to the upper boundary of the standard referential range. But when considering the minimum value, it does not similar to the lower boundary most of the time.

#### 4.9.2 Check Normality by using Data Visualization

Check whether the attributes are normally distributed or not is very important when applying statistical tests. Some statistical tests can only be used data that follows the normal distribution. And also some statistic test can be applied data which are not normally distributed.

Normality can be checked by various methods. The easiest method is looking at the graphs and finds the normality.

1. One method is looking at the frequency chats generate with histogram and the normal curve. If the graph holds the symmetric bell shape then the attribute is normally distributed. For example, see figure 14.
2. Also, the Normal Q-Q plot helps to determine the normality. If the data points lie close to the diagonal line, then the data is normally distributed. For example, see figure 15.

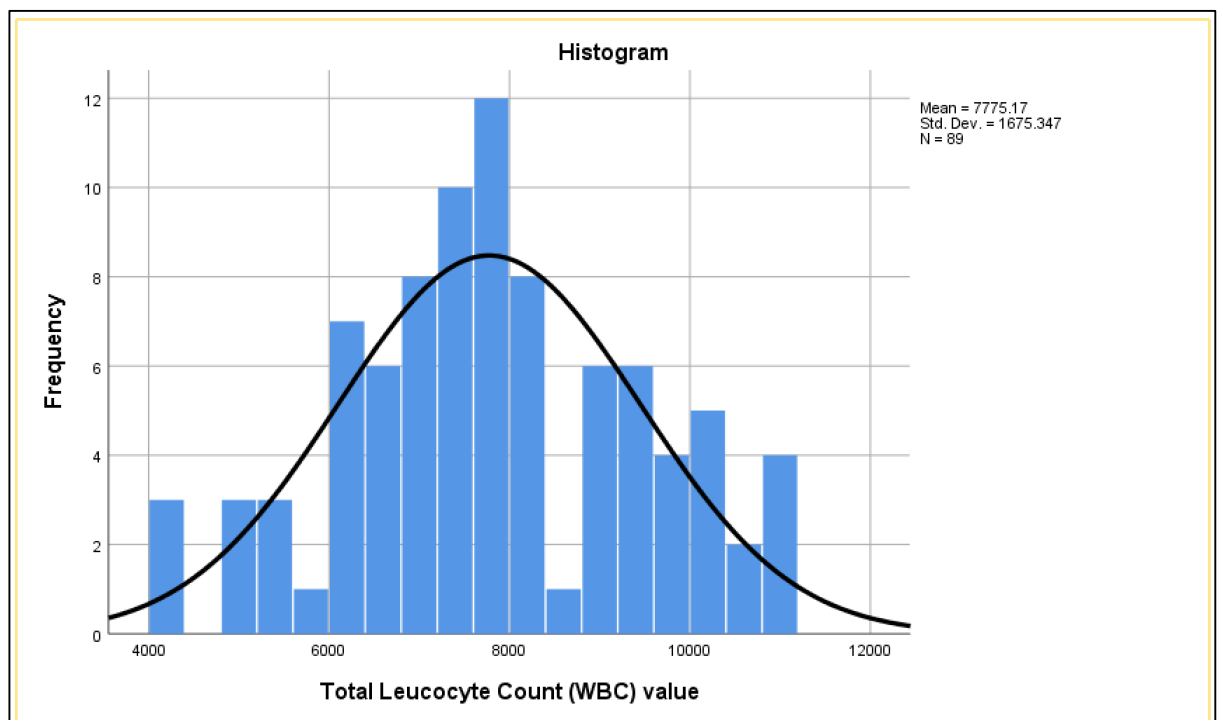


Figure 14: frequency visualize from histogram with normal curve.

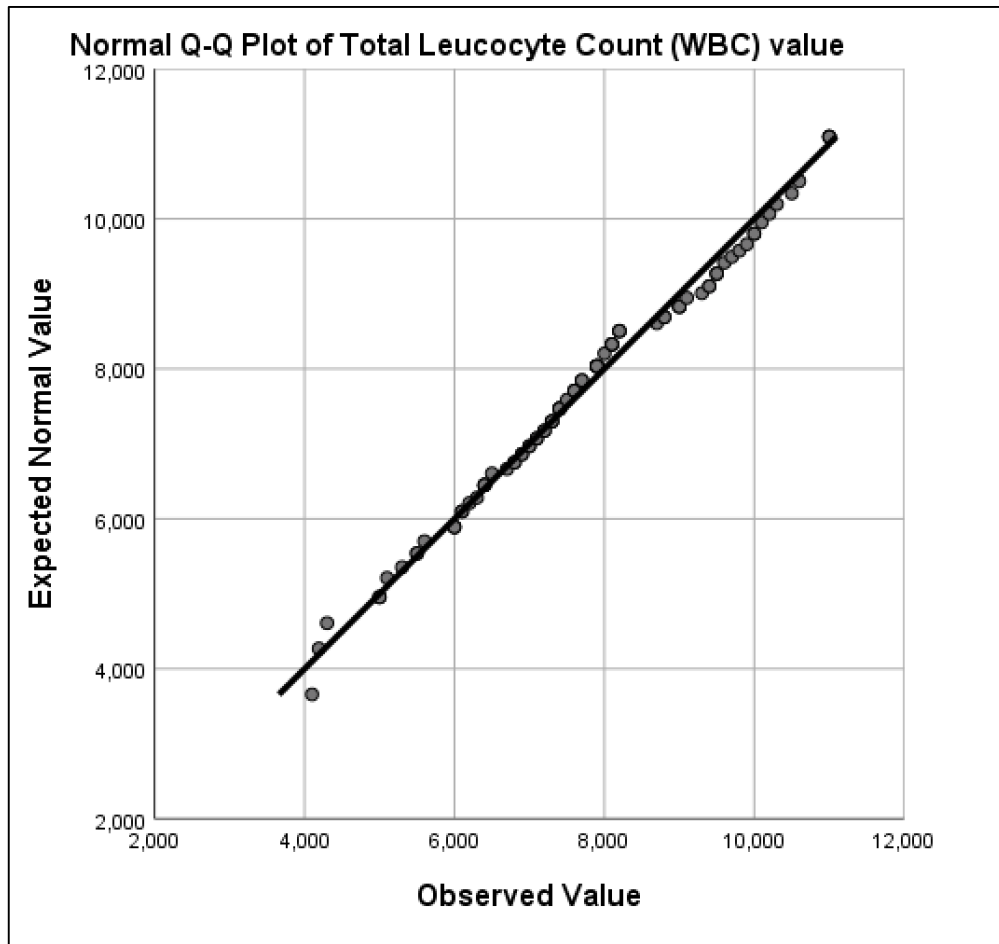


Figure 15: Q-Q plot with diagonal line

#### 4.9.3 Applying Statistical tests into the dataset

In this section, I talk about how to apply statistical tests into the dataset. Let's see how to apply the Mann-Whitney U test into the hemoglobin attribute by using the SPSS tool. Some assumptions need to hold the dataset before applying this test. We talk about them in the previous chapter. The gender and the hemoglobin attribute used for this as independent and dependent variable respectively. The independent variable has only two options as male and female. The curves belong to both options slightly the same. And, the values of the two groups are not normally distributed. The normal curves belong to these categories are shown in Figure 16 and Figure 17.



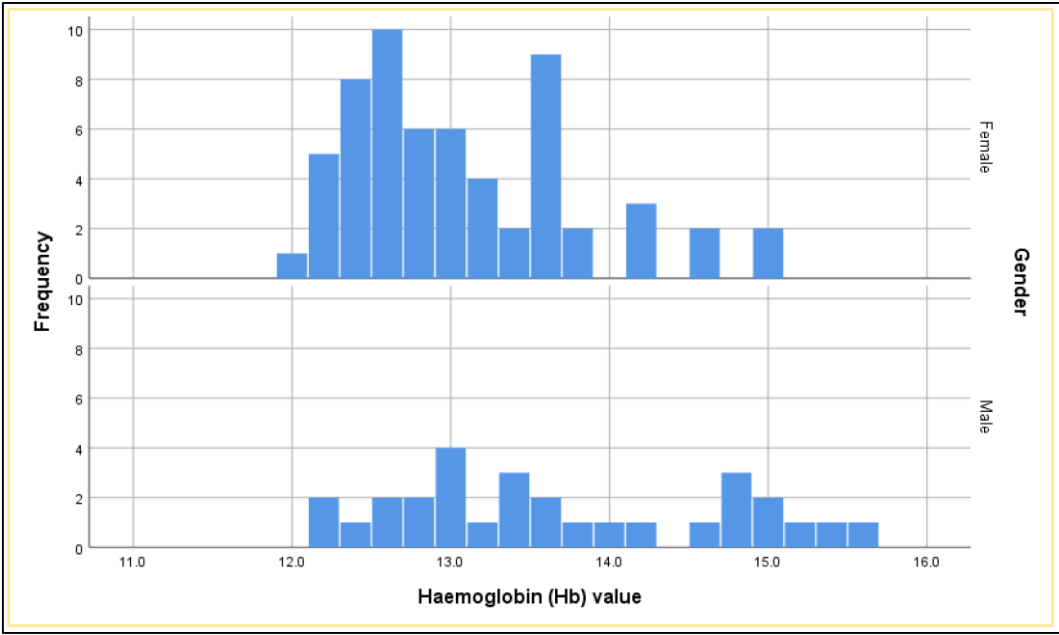


Figure 16 : Frquency graphs belong to haemoglobin

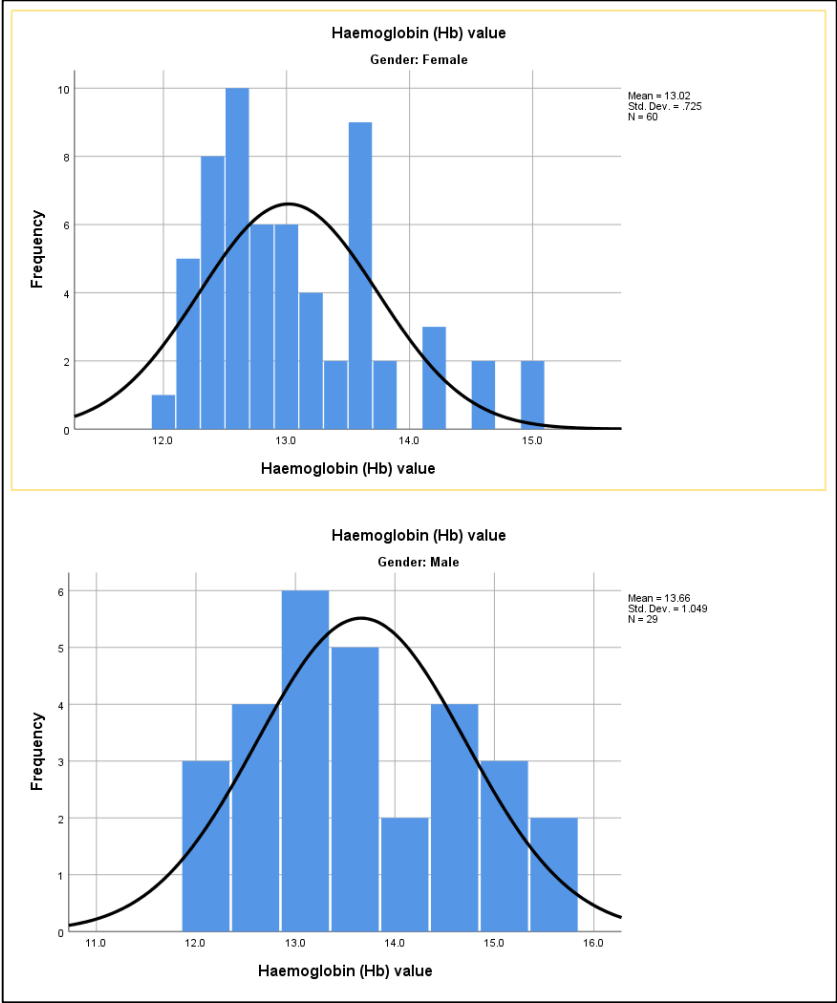


Figure 17: Frequency graphs belong to hemoglobin with the normal curve

Hence, we can apply the Mann-Whiney U test. The results received when applying the test. The result will be to talk in the next chapter.

#### **4.10 Summary**

In this chapter, I talk about how to get the dataset, how to modify the dataset according to my purpose, how to apply the classification algorithm to build a model, how to select the best model by considering the accuracy features, descriptive statistical analysis based on the healthy dataset, data visualization and understand the visualize data, how to apply statistical tests to the dataset and how to understand the output.

# Results and Evaluation

## 5.1 Introduction

We talk about what is the purpose of this research, what are the previous related works have done in researches, which kind of technologies can we use to achieve this goal, how can we achieve the final goal until this chapter begins. I have already built a model as a result of previous chapters. So, in this chapter, let's talk about how to evaluate the build model and how to achieve the final goal.

## 5.2 Test and evaluate the built model

The model built by using the Random Forest algorithm has selected the final model. In this section, we will talk about how to test this model mainly.

Weka tool can use to test the model. The test dataset created at the beginning of the research can use to accomplish this task.

Figure 14 shows the output window related to the selected model when testing it. Predictions on the test set are also attached to Appendix F.

When considering the accuracy of the testing following observations can take.

- 195 instances are correctly classified. 9 instances are incorrectly classified. 2 unhealthy instances are classified as healthy and 7 healthy instances are classified unhealthy. But correctly classified accuracy level is 95.5%.
- The kappa statistic is 0.8467. The accuracy is good hence the value near 1.

```

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.22 seconds

=== Summary ===

Correctly Classified Instances      195          95.5882 %
Incorrectly Classified Instances     9           4.4118 %
Kappa statistic                    0.8467
Total Cost                          9
Average Cost                       0.0441
K&B Relative Info Score            65.4258 %
K&B Information Score              92.5814 bits    0.4538 bits/instance
Class complexity | order 0         141.5059 bits    0.6937 bits/instance
Class complexity | scheme          40.7232 bits    0.1996 bits/instance
Complexity improvement (Sf)        100.7827 bits    0.494 bits/instance
Mean absolute error                0.1115
Root mean squared error            0.1898
Relative absolute error             36.5426 %
Root relative squared error         48.7588 %
Total Number of Instances          204

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.988   0.184   0.959     0.988   0.973     0.850   0.993    0.998    0
                0.816   0.012   0.939     0.816   0.873     0.850   0.993    0.969    1
Weighted Avg.   0.956   0.152   0.955     0.956   0.955     0.850   0.993    0.993

=== Confusion Matrix ===

 a  b  <-- classified as
164  2 |  a = 0
 7  31 |  b = 1

```

Figure 18: model testing – accuracy result

- The TP rate for classified healthy instances as healthy is 0.816 and classified unhealthy instances as unhealthy is 0.988. The values are near to the 1 too.
- FP rate to the classified healthy instance as unhealthy is 0.12 and unhealthy instance as healthy is 0.184. The values are near to the 0.
- F-measure value for healthy instances is 0.873 and for unhealthy instances is 0.973. The values are near to the 1.
- The ROC area value for both healthy and unhealthy instances is 0.993. The value is greater than 0.5.

### 5.3 Find the corresponding WBC referential range

The test set has 38 instances as healthy. but we have known already that the health code is generated without considering the WBC value. But if we consider the WBC value too, then the number of healthy instances decreases to 30. So 8 instances add to this actual healthy instances when we consider the healthy state without WBC value. When you looking at the confusion matrix you can see that 31 healthy instances are correctly classified as healthy. 31 is almost near to the 30. So we can trust the test set result.

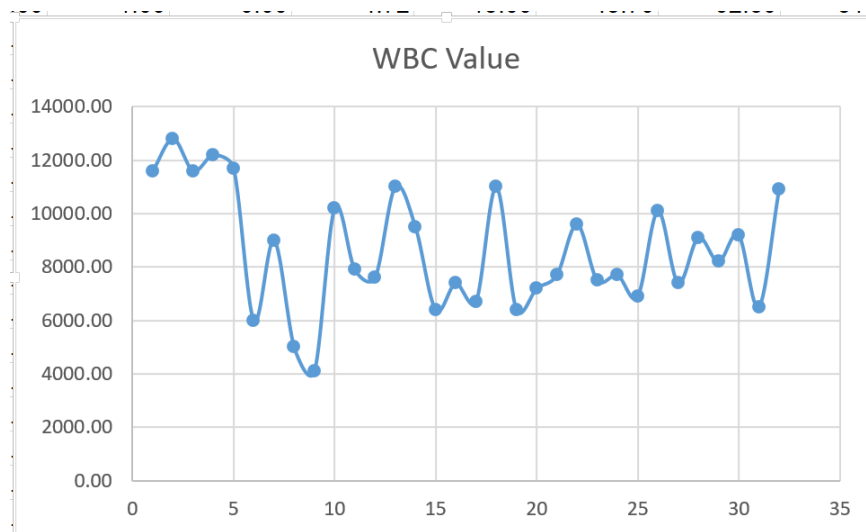


Figure 19: WBC value

Then check the predictions of the test values and identify the healthy dataset according to the test result. The data points are manually compared with an excel datasheet. The records correspond to the healthy instances which are determined in the test dataset extract from the test dataset. And take the WBC value related to them. Figure 15 shows those WBC value points.

Then take the range of those data points. It was received as  $4100\text{mm}^3 - 12800\text{mm}^3$ . The standard hematology reference range is  $4000\text{mm}^3 - 11000\text{mm}^3$ .

Finding a referential range for particular attributes is the main result of this research. And it is received in this section.

#### 5.4 Results of statistical-based data analysis

The Mann-Whitney U test is applied for the healthy hematology dataset. The following results are received.

Haemoglobin (Hb) value			
Female	N	Valid	60
		Missing	0
	Mean		13.017
	Std. Error of Mean		.0936
	Median		12.850
	Skewness		.939
	Std. Error of Skewness		.309
	Kurtosis		.300
	Std. Error of Kurtosis		.608
	Range		3.0
	Minimum		12.0
	Maximum		15.0
	Male	N	Valid
Missing			0
Mean			13.662
Std. Error of Mean			.1948
Median			13.400
Skewness			.313
Std. Error of Skewness			.434
Kurtosis			-1.193
Std. Error of Kurtosis			.845
Range			3.5
Minimum			12.1
Maximum			15.6

Figure 20 : Summary of Hemoglobin dataset

According to Figure 18, the Mean value of both Male and Female are quite similar. When considering the skewness of both graphs, it shows a positive value. So the shapes are quite similar.

<b>Descriptive Statistics</b>					
	N	Mean	Std. Deviation	Minimum	Maximum
Haemoglobin (Hb) value	89	13.227	.8916	12.0	15.6
Gender	89	.33	.471	0	1

<b>Mann-Whitney Test</b>				
<b>Ranks</b>				
	Gender	N	Mean Rank	Sum of Ranks
Haemoglobin (Hb) value	Female	60	39.73	2383.50
	Male	29	55.91	1621.50
	Total	89		

<b>Test Statistics<sup>a</sup></b>	
	Haemoglobin (Hb) value
Mann-Whitney U	553.500
Wilcoxon W	2383.500
Z	-2.775
Asymp. Sig. (2-tailed)	.006

a. Grouping Variable: Gender

Figure 21: Result of Mann-Whitney U test applying into Hemoglobin value

Here 2-tailed significant value is less than 0.05. Therefore the test performance is good. We can compare the Mean rank value of hemoglobin. It shows 39.73 for females and 55.91 for males. Therefore male has higher hemoglobin value than female.

## **5.5 Summary**

The model built in the Methodology chapter is tested in this chapter. And by using it, I found the referential range for WBC value. And, Also took meaningful results based on the statistical analysis.



### Conclusion Further work

#### 6.1 Introduction

Medical Science is a very huge area. Day by Day the area is updated. This research is also connected with Medical science. The whole research run based on some ideas such as Hematology, healthiness, unhealthiness. The word healthy is very complicated to define in medical science. It does not have a simple idea. We can define it by using various subcategories. Hematology is such a kind category. Hematology means blood. So we can tell that a person is healthy by looking at his blood reports. But the main thing you need to understand that it is not correct. There will be a person with good blood report but he may be a have eye or any other body part disability, some time his living styles (Food habits, exercises, alcohol) are not relevant to healthiness, sometime he may suffer from a disease which can not detect from blood reports, or else everything is good but he is not healthy in mentally. So healthy can not measure from the blood reports.

But when we consider if a person is suffering from a disease, then medical officers usually checked the blood reports to find/detect the disease. Hematology science is very helpful in such kind scenarios. Through this research, I hope to address the audience in this small area. Here the healthiness depends only according to the blood reports.

And also when considering the hematology attributes, some of them have an interrelationship with other attributes; depends on other attributes. But some attributes are independent. The normal value of those kinds of attributes may change according to outside reasons. For example, White blood cells have two category granulocytes and non-granulocytes. Both of these categories have five types of white blood cells in your blood as Lymphocytes, Monocyte, Eosinophil, Basophil, Neutrophil. Total white blood cell count The task of each blood cell is given below.

- **Lymphocyte:**  
It has B-lymphocyte, T-lymphocyte, and natural killer cells. Generate antibodies, fight with infection and viral cells is the main task of this type of WBC cell.
- **Monocyte:**  
This type of WBC cell attacks chronic infections if present.
- **Basophil:**  
This type of WBC cell is sensitive when occurring allergies.
- **Eosinophil:**  
This type of WBC cell is working with the immune system's responses.
- **Neutrophil:**  
This type of WBC cell helps to remove fungi and bacteria from the body.

All of these WBC cell types help to be healthy and highly affected by the hematology attributes. When you take a blood report, it shows leukocyte value (total WBC value), and also 5 types of WBC cells are listed separately too.

Inside this research, we talk about this leukocyte value. And if all WBC cell types follow normal values then there may be a high probability to take the normal value for leukocyte.

Stay in that assumption, we created a dataset that showed a healthy and unhealthy state without considering the WBC(leukocyte) value. Then built a model by using that dataset and test and evaluate it. Finally, a referential range was found.

So, this is the final chapter of the thesis and I hope to present overall achievements, limitations of the findings, achievements of the objective, and further works can do with this research.

## **6.2 Overall Achievements**

The outcome of this research is to determine a local hematology referential range for the adults in Sri Lanka. The research focused to find a local referential range for WBC value only. So a

local reference range for the WBC value is found. The standard WBC (leukocyte) referential range is  $4000\text{mm}^3 - 11000\text{mm}^3$ . But the range which is found as the result of this research is  $4100\text{mm}^3 - 12800\text{mm}^3$ . The boundaries of the result are different from the standard referential range. Several reasons can cause this result. We have already known that the standard referential range is established by considering the gaussian population. So, we do not belong to the gaussian population. Our climate, Food styles, living styles differ from Gaussian countries. Those reasons itself may cause this difference.

### **6.3 Limitations of the findings**

There are several limitations in this research and they are listed below.

- The first limitation of the research was that the data belongs to adults in Sri Lanka. Here the adults mean the people over 21 years old. So the result cannot apply the ages below 21.
- The dataset has taken from hospitals, laboratories in a particular area. So, the dataset does not cover the whole area of Sri Lanka.

### **6.4 Achievements of the objective.**

First I studied the area of blood reference ranges, blood test categories, reasons for changing blood reference ranges before stated the research as I needed surrounding knowledge about Hematology Science. It helped to identify independent attributes, dependent attributes of hematology, how the attribute's increases and decreases affect the human body, the type of blood tests, etc.

Next, I did a critical survey about technologies and areas which can be helped to accomplish my final goal. Under this, I studied data mining tools, statistical tools, data mining approach, classification methods, statistical tests, classification model's accuracy checking methods, and so on.

Finding a suitable dataset was also an objective in my research. Getting a dataset that contains human sensitive details is quite difficult as ethical approval from an acceptable organization is

a requirement to do it. Because the co-supervisor is in the medical sector, it is easy to find the dataset and it was previously taken from him for different research. When I take the dataset, I have to rebuild it according to my need. So in that phase, I have to understand the given dataset and have to find a way to rebuild it according to my purpose. I was able to go to the target as I did this thing successfully.

When considering the data set, it has 600 records that tell the healthiness and unhealthiness of a particular person without considering the WBC value. As I worked many times with this dataset manually, I knew that how many records are actually healthy, the place that records located, and what are the records newly added to the healthy group when it considers without WBC value and where those records are located and so on. These things are very helpful for me to divide the dataset into training and testing.

After I processed the dataset according to the need to build, test my model. First I select which type of classification can be applied to this model. The knowledge that I gain from critical analyzing machine learning classification algorithms is very helped me to do this task. Then I selected several classification algorithms according to the selected type of classification and build several models. After that, I analyzed them by using the accuracy features and found the best model that can be used to go to the final target.

Test and Evaluation of the model was another objective that needs to achieve. The built model is also successfully tested and using it I found a referential range according to my dataset.

## **6.5 Further Works**

Moreover, you can use the number of test datasets and find the number of referential ranges for WBC value and then take the average lower boundary value and the average upper boundary value. Similarly, you can apply the same scenario to the other appropriate attributes (if the selected attribute is not dependent on and other attributes, this method is quite bad) and find the ranges.

There is a difference between the local WBC range and the standard WBC range. Both the lower boundary and the upper boundary has been changed in the result. The lower boundary

changes may be reasonable. But the upper boundary needs to check deeper as it shows a slight big difference. However, the received local reference range is not similar to the standard referential range. So this is quite a good area to do researches to find reasons and exact value range for the local hematology range.

Another important further work is finding the reasons for these differences. Finding of How is our climate, geographical area, food, and living styles cause this difference will be a very interesting area to research.

When I work through the dataset and also when I went through the literature review, I felt the dataset is 100% not fit with the purpose. Considering this research as proof, anyone can do this research with high accuracy dataset. Considering as proof means this shows already that there may be a difference between standard hematology referential range and local referential range and we have to search accurate one for our country. So, taking high accuracy dataset is the most important thing. If someone interest in this area, he or she must find a dataset that satisfies the following points.

- It is best if you can take blood samples from the donors and prepare your dataset.
- The donors will be covered all the geographical area of Sri Lanka and selected geographical areas which show lot differences than other areas can be analyzed separately for a better outcome. Because geographical and climate changes can affect to the referential range.
- The medical history of donors is the mandatory thing that you need to give your attention and it will be helped to understand healthiness or not.
- You need to give attention to the food and living styles, any disabilities which can affect the healthiness of the donors.
- BMI value of the donors will be also a useful thing that can affect the healthiness.
- Need to check whether the donors are smoking, using alcohol, suffer from sex disease too.

- Collecting a larger dataset will help to get more accurate results.

Sometimes there may be other points that are not mentioned here, but you need to consider when you are collecting a dataset. However, making a dataset in this way will generate a cost. But it will give a high accuracy outcome.

## **6.2 Summary**

In this chapter, I talk about how I define healthiness according to my research, why I select WBC value to find a local reference range, what are the outcomes I received through this research, how I limit my solution through this research, how I achieved objectives which are mentioned chapter 1. Also, I wrote my opinions, improvements that anyone can do who like to do the researches in this field.

## References

- [1] M. Häggström, "Establishment and clinical use of reference," *WikiJournal of Medicine*, no. 26.03.2014, p. doi: 10.15347/wjm/2014.003, 26 03 2014, 1(1).
- [2] W. L Mandala, E. N. Gondwe, J. M MacLennan, M. E Molyneux and C. A MacLennan, "Age- and sex-related changes in hematological parameters in healthy Malawians," *Journal of Blood Medicine*, vol. 8, no. 28 August 2017, p. 123–130, 2017.
- [3] "Haematology Normal Adult Reference Ranges," 19 05 2017. [Online]. Available: <https://www.royalwolverhampton.nhs.uk/services/service-directory-a-z/pathology-services/departments/haematology/haematology-normal-adult-reference-ranges/>. [Accessed 09 10 2019].
- [4] I. M. Kueviakoe, A. Y. Segbena, H. Jouault, A. Vovor and M. Imbert, "Hematological Reference Values for Healthy Adults in Togo," in *International Scholarly Research Network*, 2011.
- [5] "Caucasian Countries 2019," 28 08 2019. [Online]. Available: <http://worldpopulationreview.com/countries/caucasian-countries/>. [Accessed 09 10 2019].
- [6] Abhijit Banerjee, Diganta Dey, Parbati Banerjee, Sudarshan Ray, Ratnamala Ray and Banasri Hazra, "CLSI-Derived Hematology Reference Intervals for Healthy Males in Eastern India," vol. 2, no. 2, 2013.
- [7] Timzing Miri-Dashe, Sophia Osawe, Monday Tokdung, Nenbammun Daniel, Rahila Pam Choji, Ille Mamman, Kurt Deme, Dapus Damulak and Alash'le Abimiku, "Comprehensive Reference Ranges for Hematology and Clinical Chemistry Laboratory Parameters Derived from Normal Nigerian Adults," *PLoS ONE*, vol. 9(5), no. May 15, 2014, p. e93919. doi:10.1371/journal.pone.0093919, 2014.

- [8] Elmutaz H. Taha, Mohammed Elshiekh, Mohamed Ali Alzain, Elnagi Y. Hajo, Abdelmohisen Hussein, Kamal M. Awad, Ibrahim A. Ali and Omer A. Musa, "Reference Ranges of White Blood Cells Count among Sudanese Healthy Adults," *Saudi Journal of Medicine (SJM)*, no. 30.10.2018, p. 10.21276/sjm.2018.3.10.2, 2018.
- [9] Tanzeel Huma and Usman Waheed, "THE NEED TO ESTABLISH REFERENCE RANGES," *Journal of Public Health and Biological Sciences*, vol. 2(2), pp. 188-190, 2013.
- [10] Nilgün Tekkeşin, Hüseyin Bekoz and Faruk Tükenmez, "The largest reference range study for hematological parameters from Turkey: A case control study," *Journal of Clinical and Experimental Investigations*, vol. 5 (4), no. 10.5799/ahinjs.01.2014.04.0455, pp. 548-552, 2014.
- [11] F. K. Alsheref and W. H. Gomaa, "Blood Diseases Detection using Classical Machine Learning Algorithms," (*IJACSA*) *International Journal of Advanced Computer Science and Applications*, Vols. Vol. 10, No. 7, pp. 77-81, 2019.
- [12] M. K. M. N. M. B. P. Č. M. N. & M. N. Gregor Gunčar, "An application of machine learning to haematological diagnosis," [www.nature.com/scientificreports](http://www.nature.com/scientificreports), 11.01.2018.
- [13] G. Zini, "Artificial intelligence in Hematology," *PubMed*, vol. Hematology 10(5), no. Amsterdam, Netherlands, pp. 393-400, 2005.
- [14] S. Sivapalaratnam, "Artificial Intelligence and machine learning in Haematology," *British Journal of Haematology*, vol. <https://doi.org/10.1111/bjh.15774>, no. Blackwell Publishing Inc., 2019.
- [15] G. Gunčar, M. Kukar, M. Notar, M. Brvar, P. Černelč, M. Notar and M. Notar, "An application of machine learning to haematological diagnosis," *Computer Science, Medicine, scientificreports*, vol. 8, no. DOI:10.1038/s41598-017-18564-8, 2017.
- [16] S. N. Qasem and A. Mosavi, "Novel Meta-Heuristic Model for Discrimination between Iron Deficiency Anemia and B-Thalassemia with CBC Indices Based on Dynamic Harmony Search," *AITopics*, 2020.
- [17] "Difference Between Data Mining and Data Analysis," EDUCBA, [Online]. Available: <https://www.educba.com/data-mining-vs-data-analysis/>. [Accessed 15 05 2020].



- [18] "Top 15 Best Free Data Mining Tools: The Most Comprehensive List," [Online]. Available: <https://www.softwaretestinghelp.com/data-mining-tools/>. [Accessed 15 05 2020].
- [19] B. Farnsworth, "The Top 7 Statistical Tools You Need to Make Your Data Shine," iMotions, 10 7 2019. [Online]. Available: <https://imotions.com/blog/statistical-tools/>. [Accessed 5 6 2020].
- [20] B. Godsey, "How to choose statistical software tools," Towards Data Science, 25 4 2017. [Online]. Available: <https://towardsdatascience.com/how-to-choose-statistical-software-tools-4870dd3c92a0>. [Accessed 5 6 2020].
- [21] Stephanie, "Mann Whitney U Test," 24 08 2015. [Online]. Available: <https://www.statisticshowto.com/mann-whitney-u-test/>. [Accessed 10 06 2020].
- [22] "Chi-Square Tests," [Online]. Available: [https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Chi-Square\\_Tests.pdf](https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Chi-Square_Tests.pdf). [Accessed 02 01 2020].
- [23] S. Glen, "Kruskal Wallis H Test: Definition, Examples & Assumptions From StatisticsHowTo.com," 24 2 2016. [Online]. Available: <https://www.statisticshowto.com/kruskal-wallis/>. [Accessed 18 06 2020].
- [24] "The two-dimensional Kolmogorov-Smirnov test," [Online]. Available: [https://www.researchgate.net/publication/49399948\\_The\\_two-dimensional\\_Kolmogorov-Smirnov\\_test](https://www.researchgate.net/publication/49399948_The_two-dimensional_Kolmogorov-Smirnov_test). [Accessed 20 01 2020].
- [25] N. SEMATECH, "Engineering Statistic Hand Book," [Online]. Available: <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm>. [Accessed 18 06 2020].
- [26] Keya Rani Das and Rahmatullah Imon, "A Brief Review of Tests for Normality," *American Journal of Theoretical and Applied Statistics*, Vols. vol. 5 (1), no. doi: 10.11648/j.ajtas.20160501.12, pp. pp. : 5-12, 2016.
- [27] S. Glen, "Shapiro-Wilk Test: What it is and How to Run it From StatisticsHowTo.com," 19 11 2014. [Online]. Available: <https://www.statisticshowto.com/shapiro-wilk-test/>. [Accessed 18 06 2020].
- [28] "UNDERSTANDING t-TESTS," [Online]. Available: [http://gabrielschlomer.weebly.com/uploads/2/8/8/5/28853963/understanding\\_t\\_test\\_0.pdf](http://gabrielschlomer.weebly.com/uploads/2/8/8/5/28853963/understanding_t_test_0.pdf). [Accessed 20 01 2020].

- [29] S. Glen, "T Test (Student's T-Test): Definition and Examples From StatisticsHowTo.com," [Online]. Available: <https://www.statisticshowto.com/probability-and-statistics/t-test/>. [Accessed 19 06 2020].
- [30] R. Garg, "7 TYPES OF CLASSIFICATION ALGORITHMS," [Online]. Available: <https://analyticsindiamag.com/7-types-classification-algorithms/>. [Accessed 19 06 2020].
- [31] M. Halkidi and M. Vazirgiannis, "QUALITY ASSESSMENT APPROACHES IN DATA MINING".
- [32] R. Shier, "Statistics: 2.3 The Mann-Whitney U Test," 2004. [Online]. Available: <http://www.statstutor.ac.uk/resources/uploaded/mannwhitney.pdf>. [Accessed 02 01 2020].
- [33] "Why should I use a Kruskal Wallis Test?," [Online]. Available: [https://www.researchgate.net/publication/305911256\\_Why\\_should\\_I\\_use\\_a\\_Kruskal\\_Wallis\\_Test](https://www.researchgate.net/publication/305911256_Why_should_I_use_a_Kruskal_Wallis_Test). [Accessed 20 01 2020].

# Appendix A

The model generates from Random Forest – training dataset

```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances      366          91.5 %
Incorrectly Classified Instances    34           8.5 %
Kappa statistic                    0.6884
K&B Relative Info Score            52.5209 %
K&B Information Score              142.8757 bits    0.3572 bits/instance
Class complexity | order 0         272.0358 bits    0.6801 bits/instance
Class complexity | scheme          99.0655 bits    0.2477 bits/instance
Complexity improvement (Sf)        172.9702 bits    0.4324 bits/instance
Mean absolute error                0.1271
Root mean squared error            0.2329
Relative absolute error            42.9219 %
Root relative squared error        60.6175 %
Total Number of Instances          400

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.970   0.333   0.930     0.970   0.949     0.694   0.971    0.993    0
                0.667   0.030   0.828     0.667   0.738     0.694   0.971    0.880    1
Weighted Avg.   0.915   0.279   0.911     0.915   0.911     0.694   0.971    0.973

=== Confusion Matrix ===

  a  b  <-- classified as
318 10 |  a = 0
 24 48 |  b = 1
```

## Appendix B

The model generates from Random Tree – training dataset

```
Size of the tree : 190

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances      366          91.5   %
Incorrectly Classified Instances    34           8.5   %
Kappa statistic                    0.6648
K&B Relative Info Score            59.3206 %
K&B Information Score              161.3733 bits    0.4034 bits/instance
Class complexity | order 0         272.0358 bits    0.6801 bits/instance
Class complexity | scheme          81.7646 bits     0.2044 bits/instance
Complexity improvement (Sf)        190.2711 bits    0.4757 bits/instance
Mean absolute error                0.0998
Root mean squared error            0.2233
Relative absolute error            33.6764 %
Root relative squared error        58.131 %
Total Number of Instances          400

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0.988   0.417   0.915     0.988   0.950     0.688   0.972    0.993    0
                0.583   0.012   0.913     0.583   0.712     0.688   0.972    0.874    1
Weighted Avg.   0.915   0.344   0.915     0.915   0.907     0.688   0.972    0.972

=== Confusion Matrix ===

  a  b  <-- classified as
324  4 |  a = 0
 30 42 |  b = 1
```

## Appendix C

The model generates from J48 – training dataset

```
Number of Leaves :    26
Size of the tree :    39

Time taken to build model: 0.01 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances      357           89.25 %
Incorrectly Classified Instances    43           10.75 %
Kappa statistic                    0.584
K&B Relative Info Score            38.071 %
K&B Information Score              103.5666 bits    0.2589 bits/instance
Class complexity | order 0         272.0358 bits    0.6801 bits/instance
Class complexity | scheme          132.2928 bits    0.3307 bits/instance
Complexity improvement (Sf)        139.743 bits    0.3494 bits/instance
Mean absolute error                0.1533
Root mean squared error            0.2769
Relative absolute error             51.7628 %
Root relative squared error        72.0698 %
Total Number of Instances          400

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.970   0.458   0.906     0.970   0.937     0.599   0.931    0.980    0
                0.542   0.030   0.796     0.542   0.645     0.599   0.931    0.711    1
Weighted Avg.   0.893   0.381   0.886     0.893   0.884     0.599   0.931    0.932

=== Confusion Matrix ===

  a  b  <-- classified as
318 10 |  a = 0
 33 39 |  b = 1
```

## Appendix D

The model generates from LMT – training dataset

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances      347          86.75 %
Incorrectly Classified Instances    53           13.25 %
Kappa statistic                     0.5437
K&B Relative Info Score             32.6708 %
K&B Information Score               88.8763 bits    0.2222 bits/instance
Class complexity | order 0          272.0358 bits   0.6801 bits/instance
Class complexity | scheme           142.5284 bits   0.3563 bits/instance
Complexity improvement (Sf)         129.5073 bits   0.3238 bits/instance
Mean absolute error                 0.1716
Root mean squared error             0.289
Relative absolute error             57.9182 %
Root relative squared error         75.2134 %
Total Number of Instances          400

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.924   0.389   0.915     0.924   0.920     0.544   0.925    0.983    0
                0.611   0.076   0.638     0.611   0.624     0.544   0.925    0.702    1
Weighted Avg.   0.868   0.333   0.865     0.868   0.866     0.544   0.925    0.932

=== Confusion Matrix ===

  a  b  <-- classified as
303 25 |  a = 0
 28 44 |  b = 1
```

## Appendix E

The model generates from REP Tree – training dataset

```
Correctly Classified Instances      351          87.75 %
Incorrectly Classified Instances    49           12.25 %
Kappa statistic                     0.5139
K&B Relative Info Score            32.9851 %
K&B Information Score              89.7311 bits    0.2243 bits/instance
Class complexity | order 0         272.0358 bits    0.6801 bits/instance
Class complexity | scheme          146.3226 bits    0.3658 bits/instance
Complexity improvement (Sf)        125.7131 bits    0.3143 bits/instance
Mean absolute error                 0.171
Root mean squared error             0.2924
Relative absolute error             57.7363 %
Root relative squared error         76.1148 %
Total Number of Instances          400

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.966   0.528   0.893     0.966   0.928     0.533   0.917    0.978    0
          0.472   0.034   0.756     0.472   0.581     0.533   0.917    0.664    1
Weighted Avg.   0.878   0.439   0.868     0.878   0.866     0.533   0.917    0.921

=== Confusion Matrix ===

  a  b  <-- classified as
317 11 |  a = 0
 38 34 |  b = 1
```

## Appendix F

==== Predictions on test set ====

inst#	actual	predicted	error	prediction	inst#	actual	predicted	error	prediction
1	2:1	1:0	+	0.512	42	1:0	1:0		0.983
2	2:1	2:1		0.845	43	1:0	1:0		0.999
3	2:1	2:1		0.741	44	1:0	1:0		0.977
4	2:1	2:1		0.897	45	1:0	1:0		0.997
5	2:1	1:0	+	0.557	46	1:0	1:0		0.937
6	2:1	2:1		0.897	47	1:0	1:0		0.963
7	2:1	2:1		0.934	48	1:0	1:0		0.843
8	2:1	2:1		0.87	49	1:0	1:0		0.76
9	2:1	2:1		0.732	50	1:0	1:0		0.914
10	2:1	2:1		0.842	51	1:0	1:0		1
11	2:1	2:1		0.823	52	1:0	1:0		0.827
12	2:1	2:1		0.704	53	1:0	1:0		0.753
13	2:1	2:1		0.607	54	1:0	1:0		0.978
14	2:1	2:1		0.899	55	1:0	1:0		0.866
15	2:1	2:1		0.564	56	1:0	1:0		0.99
16	2:1	2:1		0.833	57	1:0	1:0		0.99
17	2:1	2:1		0.842	58	1:0	1:0		0.911
18	2:1	1:0	+	0.503	59	1:0	1:0		0.999
19	2:1	2:1		0.686	60	1:0	1:0		0.993
20	2:1	2:1		0.77	61	1:0	1:0		0.98
21	2:1	1:0	+	0.527	62	1:0	1:0		0.992
22	2:1	2:1		0.832	63	1:0	1:0		0.99
23	2:1	2:1		0.595	64	1:0	1:0		0.98
24	2:1	2:1		0.758	65	1:0	1:0		0.907
25	2:1	2:1		0.739	66	1:0	1:0		1
26	2:1	2:1		0.777	67	1:0	1:0		0.995
27	2:1	2:1		0.728	68	1:0	1:0		0.745
28	2:1	1:0	+	0.555	69	1:0	1:0		1
29	2:1	2:1		0.762	70	1:0	1:0		0.87
30	2:1	1:0	+	0.76	71	1:0	1:0		1
31	2:1	2:1		0.706	72	1:0	1:0		0.933
32	2:1	2:1		0.707	73	1:0	1:0		0.902
33	2:1	2:1		0.68	74	1:0	1:0		0.963
34	2:1	1:0	+	0.514	75	1:0	1:0		0.896
35	2:1	2:1		0.7	76	1:0	1:0		1
36	2:1	2:1		0.837	77	1:0	1:0		0.999
37	2:1	2:1		0.833	78	1:0	1:0		0.99
38	2:1	2:1		0.658	79	1:0	1:0		0.97
39	1:0	2:1	+	0.564	80	1:0	2:1	+	0.595
40	1:0	1:0		0.96	81	1:0	1:0		0.999
41	1:0	1:0		0.98	82	1:0	1:0		0.971
					83	1:0	1:0		1
					84	1:0	1:0		1
					85	1:0	1:0		0.921
					86	1:0	1:0		0.911



inst#	actual	predicted	error prediction	inst#	actual	predicted	error prediction
87	1:0	1:0	0.981	132	1:0	1:0	0.994
88	1:0	1:0	0.943	133	1:0	1:0	0.982
89	1:0	1:0	0.985	134	1:0	1:0	0.989
90	1:0	1:0	0.896	135	1:0	1:0	0.914
91	1:0	1:0	0.89	136	1:0	1:0	0.922
92	1:0	1:0	0.808	137	1:0	1:0	0.998
93	1:0	1:0	1	138	1:0	1:0	0.973
94	1:0	1:0	0.912	139	1:0	1:0	0.854
95	1:0	1:0	0.791	140	1:0	1:0	0.995
96	1:0	1:0	1	141	1:0	1:0	1
97	1:0	1:0	0.989	142	1:0	1:0	0.868
98	1:0	1:0	1	143	1:0	1:0	0.514
99	1:0	1:0	0.844	144	1:0	1:0	0.98
100	1:0	1:0	0.503	145	1:0	1:0	0.99
101	1:0	1:0	1	146	1:0	1:0	1
102	1:0	1:0	1	147	1:0	1:0	1
103	1:0	1:0	0.852	148	1:0	1:0	0.512
104	1:0	1:0	1	149	1:0	1:0	1
105	1:0	1:0	1	150	1:0	1:0	0.98
106	1:0	1:0	1	151	1:0	1:0	0.997
107	1:0	1:0	0.813	152	1:0	1:0	0.99
108	1:0	1:0	0.968	153	1:0	1:0	0.96
109	1:0	1:0	0.96	154	1:0	1:0	1
110	1:0	1:0	0.784	155	1:0	1:0	0.87
111	1:0	1:0	1	156	1:0	1:0	0.887
112	1:0	1:0	0.78	157	1:0	1:0	0.995
113	1:0	1:0	0.961	158	1:0	1:0	1
114	1:0	1:0	1	159	1:0	1:0	1
115	1:0	1:0	0.97	160	1:0	1:0	0.92
116	1:0	1:0	0.93	161	1:0	1:0	0.999
117	1:0	1:0	1	162	1:0	1:0	0.858
118	1:0	1:0	0.976	163	1:0	1:0	0.968
119	1:0	1:0	0.958	164	1:0	1:0	0.557
120	1:0	1:0	1	165	1:0	1:0	0.738
121	1:0	1:0	0.905	166	1:0	1:0	0.68
122	1:0	1:0	0.987	167	1:0	1:0	0.99
123	1:0	1:0	1	168	1:0	1:0	0.792
124	1:0	1:0	1	169	1:0	1:0	0.98
125	1:0	1:0	1	170	1:0	1:0	0.989
126	1:0	1:0	0.981	171	1:0	1:0	0.994
127	1:0	1:0	0.998	172	1:0	1:0	0.995
128	1:0	1:0	0.99	173	1:0	1:0	1
129	1:0	1:0	0.995	174	1:0	1:0	1
130	1:0	1:0	0.997	175	1:0	1:0	0.99
131	1:0	1:0	1	176	1:0	1:0	0.527

inst#	actual	predicted	error	prediction
177	1:0	1:0	0.998	
178	1:0	1:0	0.99	
179	1:0	1:0	1	
180	1:0	1:0	0.963	
181	1:0	1:0	1	
182	1:0	1:0	0.875	
183	1:0	1:0	0.97	
184	1:0	1:0	0.957	
185	1:0	1:0	0.998	
186	1:0	1:0	0.98	
187	1:0	1:0	1	
188	1:0	1:0	0.76	
189	1:0	1:0	0.555	
190	1:0	1:0	1	
191	1:0	1:0	0.99	
192	1:0	1:0	0.99	
193	1:0	1:0	0.76	
194	1:0	1:0	0.98	
195	1:0	1:0	0.948	
196	1:0	1:0	0.99	
197	1:0	1:0	0.99	
198	1:0	1:0	0.979	
199	1:0	1:0	0.99	
200	1:0	1:0	0.897	
201	1:0	1:0	0.98	
202	1:0	1:0	1	
203	1:0	1:0	0.995	
204	1:0	1:0	1	