MALARIA PREDICTION USING BAYESIAN AND OTHER MACHINE LEARNING TECHNIQUES

A Thesis Presented to the Department of

Computer Science

African University of Science and Technology

In Partial Fulfilment of the Requirements for the Degree of

Master of Science

By

Hamisu Ismail Ahmad

(ID No. 40559)

Abuja, Nigeria

September, 2019

CERTIFICATION

This is to certify that the thesis titled “Malaria Prediction Using Bayesian and Other Machine Learning Techniques” submitted to the school of postgraduate studies, African University of Science and Technology (AUST), Abuja, Nigeria for the award of the Master's degree is a record of original research carried out by Hamisu Ismail Ahmad in the Department of Computer Science.

MALARIA PREDICTION USING BAYESIAN AND OTHER MACHINE LEARNING TECHNIQUES

By

Hamisu Ismail Ahmad

(ID No. 40559)

A THESIS APPROVED BY THE COMPUTER SCIENCE DEPARTMENT

RECOMMENDED: Supervisor, Dr. Rajesh Prasad

Head Department of Computer Science

\APPROVED: _____

Chief Academic Officer

_____

Date

ABSTRACT

Main purpose of data mining is to extract valuable information from available data. With the enormous amount of data stored in files, databases, and repositories, in the healthcare sector, it's increasingly important, if not necessary, developing powerful means for analysis and interpretation of such data for the extraction of knowledge that could help in decision-making. Classification is technique in data mining; it's defined as distinguishing, assigning object to a certain class based on its similarity to previous examples of other objects in the dataset. This thesis used four classification algorithms that include Bayesian Network, Decision tree (J48), ZeroR and OneR to build the classification models on the malaria dataset and compare the models, measure their performance with WEKA-API Library imported to JDK8, and uses the ensembles method in boosting the performance of the classification algorithms. The result identified that Decision Tree (J48) yield highest accuracy of 88.4% before and after boosting followed by Naïve Bayes, with 79.9% before and after boosting is 87.0%, then OneR with accuracy of 79.8%, 87.6% before and after boosting respectively and ZeroR have lesser accuracy of 60.9% before and after boosting, The research suggested that Decision tree is the best to be used on Malaria dataset in our hospitals.

Keywords: *data mining, pre-processing, classification, association, clustering, regression, ensembles, boosting, dataset*

ACKNOWLEDGMENTS

My gratitude goes to the Allah Almighty for giving me life, health, strength, knowledge, and courage to carry out this thesis successfully. My special gratitude goes to my parents Alh. Ismail Ahmad Aujara (Agent), Maimuna Yahaya (Ummi) and my beloved brother Musbahu for their concern and financial support on me. My weighty gratitude also goes to the entire community of African University of Science and Technology Abuja, AUST for the privilege given to me to study in the environment, my supervisor **Dr. Rajesh Prasad** will not be left for his patience, advice and time to guide me throughout the my thesis work. Finally, big thanks will be said to my friends and course mates.

TABLE OF CONTENTS

CERTIFICATION II

ABSTRACT V

ACKNOWLEDGMENTS VI

CHAPTER ONE 11

INTRODUCTION 11

1.0. About Malaria 11

1.1. Data Mining 13

1.2. Classifications 14

1.3. Problem statement 15

1.4. Research Questions 16

1.5. Research Objective & Contribution 16

1.5.1 Objectives 16

1.5.2 Contributions 16

1.6. Chapterization 17

CHAPTER TWO 18

LITERATURE REVIEW 18

2.0. Data Mining 18

2.1. Data Mining Processes 18

2.1.1 Data Mining Techniques in Health Care 19

2.1.2 Association 20

2.1.3 Clustering 20

2.1.4 Classification 21

2.2. Articles reviews 23

CHAPTER THREE 27

MATERIAL AND METHOD 27

3.0. Material and methods 27

3.1. Software's and hardware's selection 27

3.2. Data collection 28

3.3. Dataset Pre-processing 28

3.4. Opting and Building Mathematical models and Classification Algorithms 29

3.4.1 Bayesian Network Algorithm 30

3.4.2 Naïve Bayes Algorithm 32

3.4.3 Decision Tree 34

3.4.4. ZeroR 38

3.4.5 OneR 39

3.5. Evaluation Metrics 40

3.6. Boosting Classifier Method 40

CHAPTER FOUR 44

RESULTS AND DISCUSSIONS 44

4.0. Presentation of the result 44

CHAPTER FIVE 51

CONCLUSION AND FEATURE WORK 51

5.0. Conclusion 51

5.1. Limitation of Study 51

5.2. Feature Work 51

BIBLIOGRAPHY 53

List of Tables

Table 2.2 Descriptive of some Classification Algorithms 21
Table 2.3 Benefit and drawback of different classification 23
Table 3.1 Description of unsupervised dataset 28
Table 3.2 Description of supervised dataset 29
Table 3.3 Frequency table of Naïve Bayes 33
Table 3.4 Likelihood table of Naïve Bayes 33
Table 3.5.1 Frequency table for the target 36
Table 3.5.2 Frequency table for the Fever 36
Table 3.5.3 Frequency table for the Headache 37
Table 3.5.4 Frequency table for the Nausea 37
Table 3.5.5 Frequency table for the Vomiting 37
Table 3.5.6 Frequency table for of Dataset 40
Table 4.1 Percentages of correctly and incorrectly 44
Table 4.2 Performance summary of 4 classifiers 45
Table 4.3 Performance related to confusion matrix 46
Table 4.4 Performance of Naïve Bayes before and after boosting 47
Table 4.5 Performance of Decision Tree J48 before and after boosting 47
Table 4.6 Performance of ZeroR before and after boosting 48
Table 4.7 Performance of OneR before and after boosting 49
Table 4.8 Comparison and change of accuracy of all four classification models before and after boosting 50

List of Figures

Figure 1.1 Data distribution into an integrated medical database. 14
Figure 2.1 Different data mining techniques in healthcare 20
Figure 2.2 Bayesian classifier structure 22
Figure 2.3 Decision tree structure 23
Figure 3.1 Description of the process and tools 27
Figure 3.2 Different classifiers that uses frequency table 30
Figure 3.3 Model construction and evaluation in data mining 30
Figure 3.4 Structure of Naïve Bayes constructed from malaria dataset 32
Figure 3.5 Structure of dataset before constructing Decision tree 35
Figure 3.6 Entropy and Gini 35
Figure 3.6 shows the initial step of constructing Decision tree (J48) 37
Figure 3.7 progressive and complete Decision Tree (J48) 38
Figure 3.8 Target's frequency table 39
Figure 3.9 Cross validation 39
Figure 4.1. Accuracy chart of the constructed classifier model 44
Figures 4.2 Root mean square error (RMSE) for the 4 different classification 45
Figure 4.3 Time taken to build a classifier 46

CHAPTER ONE
INTRODUCTION

1.0. About Malaria

Malaria is a very common disease in Sub-Saharan African countries that is caused by a parasite. The parasite is usually transmitted to human's bodies through the bites of infected female mosquitoes which bite mainly between dusk and dawn, it's also categorised as one of the life-threatening disease that takes less than an hour to spread across the body of an individual after infection, but it is preventable and curable (WHO report on Malaria 27, March 2019), Once an infected mosquito bites a human, the parasites multiply in the host's liver before infecting and destroying red blood cells (Jill Seladi-Schulman, PhD, Mon 19 November 2018 Medical News Today MNT), Human malaria is caused by four different species of Plasmodium: P. falciparum, P. malariae, P. ovale and P. vivax, More than 90 per cent of human malaria infections in Sub-Saharan Africa are due to P.falciparum while the remainders are due to P. ovale, P. vivax, or P. malariae. Occasionally mixed infections occur (Guidelines for the Treatment of Malaria in South Africa 2018 Update).

Doctors divide malaria symptoms into two categories: Table 1.1, shows Uncomplicated and severe malaria. (Jill Seladi-Schulman, PhD, Mon 19 November 2018)

Table 1.1 Category of malaria

Uncomplicated Malaria:	***Severe Malaria.***
A doctor would give this diagnosis when symptoms are present, but no symptoms occur that suggest severe infection or dysfunction of the vital organs. *This form can become severe malaria without treatment, or if the host has poor or no immunity.* *Symptoms of uncomplicated malaria typically last 6 to 10 hours and recur every second day.* *Some strains of the parasite can have a longer cycle or cause mixed symptoms.* *As symptoms resemble those of flu, they may*	*In severe malaria, clinical or laboratory evidence shows signs of vital organ dysfunction.* *Symptoms of severe malaria include:* • *fever and chills* • *impaired consciousness* • *prostration, or adopting a prone position* • *multiple convulsions* • *deep breathing and respiratory distress* • *abnormal bleeding and signs of anemia*

remain undiagnosed or misdiagnosed in areas where malaria is less common. *In uncomplicated malaria, symptoms progress as follows, through cold, hot, and sweating stages:* • *a sensation of cold with shivering* • *fever, headaches, and vomiting* • *seizures sometimes occur in younger people with the disease* • *sweats, followed by a return to normal temperature, with tiredness* *In areas where malaria is common, many people recognize the symptoms as malaria and treat themselves without visiting a doctor.*	• *clinical jaundice and evidence of vital organ dysfunction* *Severe malaria can be fatal without treatment.*

The World malaria report 2018 estimates that there were 219 million cases of malaria in 2017. The 10 highest burden African countries saw an estimated 3.5 million more malaria cases in 2017 compared with the previous year, Malaria continues to claim the lives of more than 435 000 people each year, largely in Africa. Children under the age of 5 are especially vulnerable; the fact that every two minutes a child dies from this preventable and curable disease is unacceptable.

Most malaria cases in 2017 were in the WHO African Region (200 million or 92%), followed by the WHO South-East Asia Region with 5% of the cases and the WHO Eastern Mediterranean Region with 2%.

- Fifteen countries in Sub-Saharan Africa and India carried almost 80% of the global malaria burden. Five countries accounted for nearly half of all malaria cases worldwide: Nigeria (25%), Democratic Republic of the Congo (11%), Mozambique (5%), India (4%) and Uganda (4%).
- The 10 highest burden countries in Africa reported increases in cases of malaria in 2017 compared with 2016. Of these, Nigeria, Madagascar and the Democratic

Republic of the Congo had the highest estimated increases, all greater than half a million cases. In contrast, India reported 3 million fewer cases in the same period, a 24% decrease compared with 2016.

This 2018 report shows that after an unprecedented period of success in global malaria control, progress has stalled. Data from 2015–2017 highlight that no significant progress in reducing global malaria cases was made in this period. There were an estimated 219 million cases and 435 000 related deaths in 2017.

The *World malaria report 2018 also* draws on data from 87 countries and areas with ongoing malaria transmission; the information is supplemented by data from national household surveys and databases held by other organizations.

1.1. Data Mining

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process Osmar R. (1999).

Data mining has an infinite potential to utilize healthcare data more efficiently and effectually to predict different kind of disease. Statistical data mining tools and techniques can be roughly grouped according to their use for clustering, classification, association, and prediction.

The trend of application of data mining in healthcare today is increased because the health sector is rich with information and data mining has become a necessity. Healthcare organizations generate and collect large volumes of information to a daily basis. Use of information technology enables automation of data mining and knowledge that help bring some interesting patterns which means eliminating manual tasks and easy data extraction directly from electronic records, electronic transfer system that will secure medical records, save lives and reduce the cost of medical services as well as enabling early detection of infectious diseases on the basis of advanced data collection.

Though data mining (DM) methods and tools have been applied in different domains already for more than 40 years, their applications in healthcare are relatively young. R. D.

Wilson et al, have started to classify and collect medical publications where knowledge discovery and DM techniques were applied or researched from 1966 till 2002. Data Mining is one of the foremost motivating spaces for analysis that is mounting progressively standard in the healthcare industry. Data mining plays an efficient role in revealing the new emerging trends associated with this scenario. In the health industry, data processing provides many advantages in transactional applications like Electronic Health Record (EHR), Figure 1.1 patient satisfaction systems, lab systems, economic systems, patient identification etc. This survey highlights few applications and future issues of Data mining in medical field. It also provides a picture of a database which exists in health care organization (E. Mercy Beulah, 22 August, 2016).

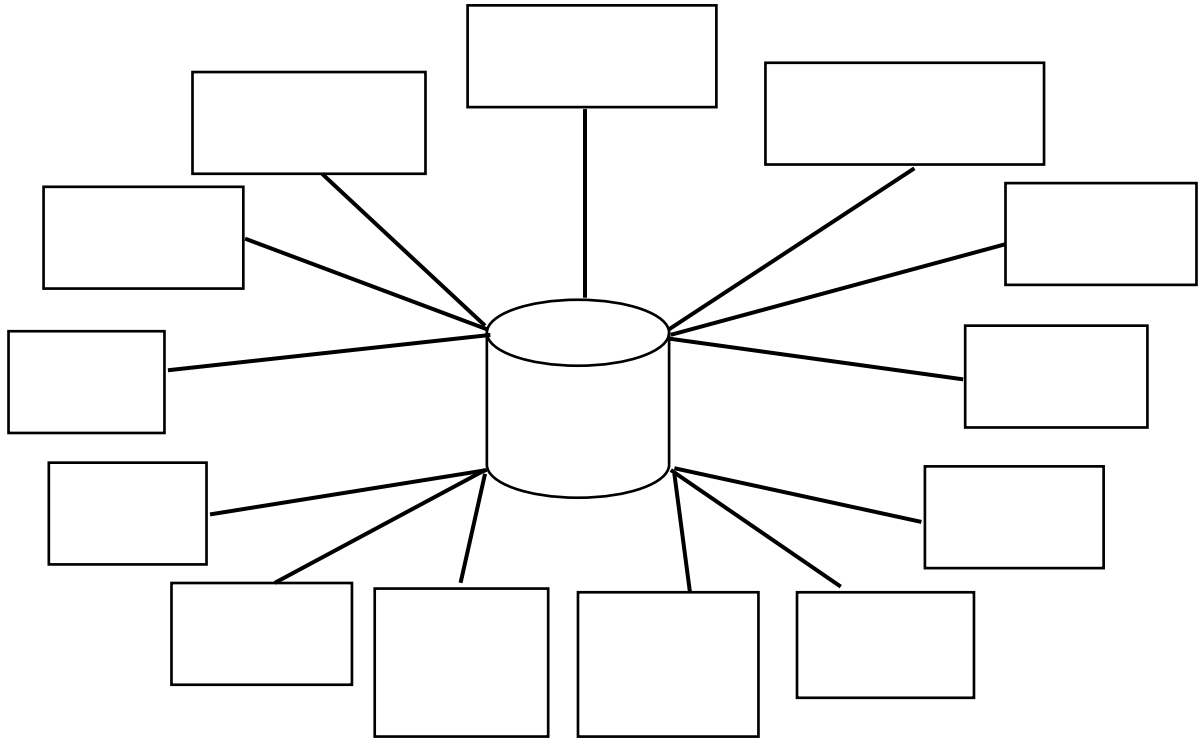


Figure 1.1 Data distribution into an integrated medical database.

1.2. Classifications

Classification is one of the most commonly used methods of data mining in Healthcare organization/sectors. Different data mining classification techniques have been used to help healthcare professionals for prediction, diagnosis, detection of various diseases such as malaria, thyroid, heart, diabetes, cancer diseases etc and also in Treatment effectiveness, Management of healthcare, Detection of fraud and abuse, Customer relationship management etc. Jammu (2017).

Classification analysis is the organization of data in given classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the Our Video Store managers could analyze the customers' behaviours vis-à-vis their credit, and label accordingly the customers who received credits with three possible labels "safe", "risky" and "very risky". Osmar R (1999).

One key area of data mining that demonstrates its application in the healthcare sector is the use of classification techniques in classifying and predicting various diseases (Rani & Govrdhan, 2010).

There are many classification techniques that includes: Simple Logistic, Instance-based k-nearest Neighbors (IBK), Naive Bayes, Stochastic Gradient Descent (SGD), Logistic model tree (LMT) and Sequential Minimal Optimization (SMO) that are used on the training dataset to categorize and predict the patients that are affected by malaria.

1.3. Problem statement

There is need of development of effective diagnostic strategies on malaria due to its continuous impact on global health sector. In malaria common regions, accurate diagnosis is hampered by technical and infrastructural challenges to many laboratories. These laboratories lack standard facilities, expertise or diagnostic supplies; thus, therapy is administered based on clinical or self-diagnosis. Now there is need for accurate and fast diagnosis strategies of malaria due to the unbroken increment in its cost of medication Francis D. (2017).

It's always difficult task evaluating the huge data generated in the healthcare sector and use the data mining techniques to discover useful knowledge and find hidden patterns for decision making. During diagnosis, management and treatment of diseases, the healthcare data must to be analysed accurately to avoid errors (Ogundele, Popoola,Oyesola, Orija, 2018).

Existing or traditional diagnosing methods of this common disease are time-consuming and tedious (Gracelyn Shi, *2017).*

Millions among the people that being affected by malaria face the severity if it's not been diagnose and treated at its earlier stage. Tremendous applications have been formed and introduced for the diagnosis and the treatment of malaria to solve the issue, but yet not fully recovered (Rahila, 2017).

1.4. Research Questions

1. What is the use of huge electronic data that is being generated on daily basis in healthcare sector?
2. How to utilize the data generated in healthcare sector to discover some hidden patterns and information i.e. Knowledge Discovery in Data (KDD)?
3. How to develop effective diagnostic strategies on several diseases due to their continuous impact on global health sector?
4. What is the diseases predictor or symptoms that will be the feature of the given dataset?
5. How to pre-process the dataset and build the mathematical model of several classification algorithms?
6. What are the performance metrics and how to evaluate the performance of different classification models?
7. How to identify the weakness of the one or more classification algorithm and use ensemble method to boost their performance?
8. Which among the classification techniques will be suitable in healthcare sector?

1.5. Research Objective & Contribution

1.5.1 Objectives

- Developing a model based on decision support system to predict the disease quite accurately because most of the conventional machine learning algorithms are showing very poor performance to classify the skewed distribution data i.e., whether a patient is affected by malaria disease or not.
- Compare several classification techniques or algorithms to major their
  - *Accuracy*
  - *Efficiency*
  - *Speed*
  - *Scalability*

1.5.2 Contributions

This research uses several classification techniques or algorithms that includes, Bayesian Network, Naive Bayes, Decision tree (J48), ZeroR and OneR on the training dataset to predict the patients that are affected by malaria base on its symptoms and

- Compare the classification algorithm in terms of high accuracy, scalability and measure their performance with prominent data mining tool WEKA-API library with Netbeans software to identify which one among those classification techniques is better to be used specifically in prediction of malaria in the hospitals.
- Boosting the ability and accuracy of the classification algorithm using ensemble method and measure the performance again.

1.6. Chapterization

The research is all about using different techniques of classification for predicting malaria using data mining approach. The entire thesis is distributed over five chapters. Each chapter highlighted different topics and subtopics.

Chapter 1 contains introduction about malaria, data mining, classification, problem statement, objective and our contribution.

Chapter 2 discusses the Data Mining, Classification concept, we also illustrate the most promising algorithms that demonstrate high performance in the task of classification namely, Naive Bayes, Decision tree, ZeroR and OneR and critically review literature.

Chapter 3 describe the methodology followed to fulfil our objectives in making a comparison between a set of classification algorithms in data mining. Secondly we pick out the suitable open source software for our work. Thirdly, we evaluate the models generated and lastly we examine each technique performance and uses ensemble method to boost the Classification Algorithm.

Chapter 4 presented the result of their accuracy performance, speed, root means square error (RMSE) and confusion matrix in forms of charts of the models before and after Boosted

Finally, Chapter 5 includes conclusion, recommendation, limitation and future work.

CHAPTER TWO
LITERATURE REVIEW

This chapter discusses the concept of data mining and Classification. We also illustrate the most promising algorithms that demonstrate high performance in the task of classification namely Simple Logistic, IBK, Naive Bayes, SGD, LMT and SMO.

2.0. Data Mining

Data Mining is an important part of machine learning and its main purpose is to extract valuable information from available data, which includes the process of sorting the large set of data to identify patterns and form the relationships to solve problems through data analysis.

Data mining techniques applied in healthcare industry play a major role in prediction and diagnosis of the diseases. Data mining concept is among the top and rapidly growing technology, generally biomedical sciences and research having being using this concept. In today's modern medicine activities huge information is stored daily in the medical database. For example, medical data can includes blood sugar, blood pressure, ECG, MRI, cholesterol levels, and so on, and also physician's analysis. Scientific decision-making for the diagnosis, treatment, and disease's prediction all depended on extracting important knowledge from the database (Sandhya, 2010).

In KDD process, data mining is the main component. Data mining involves making the choice of the data mining algorithm(s) to generate previously unknown and hypothetically beneficial information by using algorithms on the stored data in the database. This works includes making decision of which algorithms and parameters can be suitable and match a particular data mining method along with the general standards of KDD process. Data mining techniques contains Association, clustering, classification, summarization, regression, etc. (Prakash & Hanumanthappa, 2013).

2.1. Data Mining Processes

The availability of the volume of data generated in healthcare industry need to be transform into meaningful information for decision to occur. Data mining provides a great promise in analysing complexity of data to generate information. The process of data mining helps to discover knowledge which are done in five steps starting from selection stage to knowledge discovery.

- Selection

  The data is selected according to some criteria in this stage. For example, a bicycle owns by all those people, we can determine subsets of data in this way.

- Pre-processing

  This stage removes that information which is not necessary for example while doing pregnancy test it is not necessary to note the sex of a patient. It is also known as data cleansing stage.

- Transformation

  This stage transformed only those data which are useful in a particular research for example only data related to a particular demography is useful in market research.

- Data mining

  Data mining is a stage knowledge discovery process. This stage is useful for extracting the meaningful patterns from data.

- Interpretation and evaluation.

  The meaningful patterns which the system identified are interpreted into knowledge in this stage. This knowledge may be then useful for making useful decisions (Ahmad, Qamar, & Qasim Afser Rizvi, 2015)

2.1.1 Data Mining Techniques in Health Care

Supervised and unsupervised learning are the two classes of data mining techniques. Supervised learning involves a teacher that helps to learn. The learning predict an outcome based on certain criteria. Examples of such learning are classification, regression. Similarly, unsupervised learning is a techniques that does not involves a teacher. It outlines class of data without his assignment. Common example is the clustering. Table 2.1 shows the summary of the supervised and unsupervised learning (N & S, 2016).

Table 2.1 Characteristic of supervised and unsupervised learning

	Characteristic	**Techniques**
Supervised learning	• involves teacher • identify class • Mostly used	• Classification • Statistical Regression
Unsupervised learning	• No teacher involve • Classes are defined	• Clustering • Association

	• Not frequently used	Rule

Association, classification and clustering are illustrated in Figure 2.1 are among data mining techniques that are used in healthcare sector to increase their capability for building proper conclusions about patient's health from raw facts and figures (Syed & Brijesh, 2007).

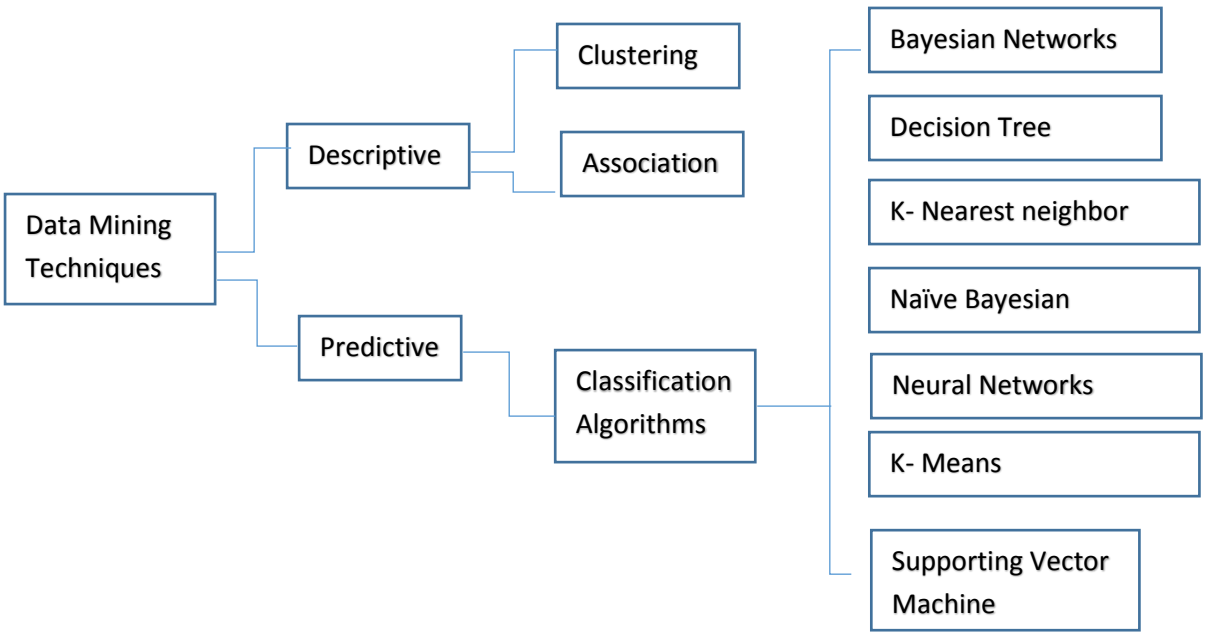


Figure 2.1 Different data mining techniques in healthcare

2.1.2 Association

Association is among the Data mining techniques that have great impact in the healthcare sector to identify the relationships between diseases, state of human health and the symptoms of disease. Ji et al., used association in order to learn uncommon casual relationships in Electronic health databases (J.Yanqing & H.Ying & J.Tran & P.Dews & A.Mansour & R.Michael Massanari, 2011).

2.1.3 Clustering

Clustering is also part of Data mining techniques it differ from classification; it has no predefined classes. A huge database is partitioned into number of subgroups named clusters. Data is divided base on its similarities. Clustering algorithms identifies collections of the data such that objects within same cluster are more identical to each other, than other groups (Sharmila & Vethamanickam, 2015)

2.1.4 Classification

Classification is one of the most commonly used methods of data mining in healthcare organization/sectors. Different data mining classification techniques have been used to help healthcare professionals for prediction, diagnosis, detection of various diseases such as thyroid, heart, diabetes, cancer diseases, and also in Treatment effectiveness, Management of healthcare, Detection of fraud and abuse, Customer relationship management etc. The most common classification data mining techniques used in healthcare illustrated in Table 2.2 are: neural network, decision tree, nearest neighbour algorithms, support vector machine, Bayesian Methods (Sidiq & Khan, 2017)

Table 2.2 Descriptive of some Classification Algorithms

Techniques name	**Abbreviation**	**Developed by**	**Year**	**Category**
Simple Logistic classifier	Simple logistic	David cox	1985	Logistic Regression Model
Instance- based K-nearest Neighbour	IBK	Aha et al.	1991	K-nearest Neighbours Classifier
Naïve Bayesian Classifier	Naïve Bayes	John and Langly	1995	Bayesian Classification
Stochastic gradient descent	SGD	Robbins and Monro	1951	Stochastic approximation method
Logistic Model Tree	LMT	Lan Wehr et al.	2005	Decision Tree classification
Sequential Minimal Optimization	SMO	John Platt	1998	Support Vector Classification

❖ *Naive Bayes*

For probabilistic learning method Bayesian classification is used. With the help of classification algorithm we can easily obtained it. Bayes theorem of statistics plays huge role in it, but in medical domain, attributes such as patient symptoms and their health conditions are correlated with each other but Naive Bayes Classifier assumes that all attributes are independent with each other as shown in Figure 2.2. This is the key drawback with Naive Bayes Classifier. If attributes are independent with each other than Naïve Bayesian classifier has shown high accuracy. In healthcare field they play very important roles. Hence, researchers across the world used hem there are main two advantages of BBN. First One is it helps to makes computation process very easy. Second one is that for huge datasets it has better speed and accuracy. Bayesian Belief Network is widely used by many

researchers in healthcare domain. Liu et al. develop a decision support system using BBN for analysing risks that are associated with health. Curiac *et al.*, used BBN in making significant decision regarding patient health suffering from psychiatric disease to analyse the psychiatric patient data and performed experiment on real data obtain from Lugoj Municipal Hospital. The structural model is represented as a directed graph where the nodes represent attributes and arcs represent attribute dependency (Sidiq & Khan, 2017)

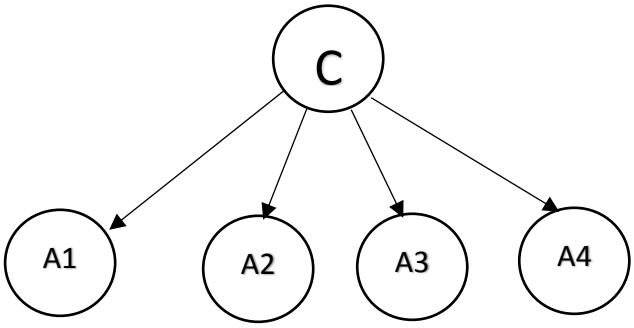


Figure 2.2 Bayesian classifier structure

❖ *Decision Tree*

Decision Tree is usually used by various researchers in healthcare field. Decision tree is among the classification technique that solves large and complex problems by providing rules in an understandable form. It is a knowledge representation structure as depicted in Figure 2.3 it consists of nodes and branches organized in a tree shaped such that, every internal non-leaf node is labelled with values of the attributes and the branches coming out from an internal node are labelled with values of the attributes in that node. Every node is labelled with a class (a value of the goal attribute). Tree based models which include classification and regression trees, are the common implementation of induction modelling. Decision tree models are best suited for data mining as they are easy to interpret, inexpensive to construct, easy to integrate with database system and they have comparable or better accuracy in many applications. By Using Decision Tree, decision makers can choose best traversal and alternative from root to leaf indicates unique class separation based on maximum information gain (Sidiq & Khan, 2017).

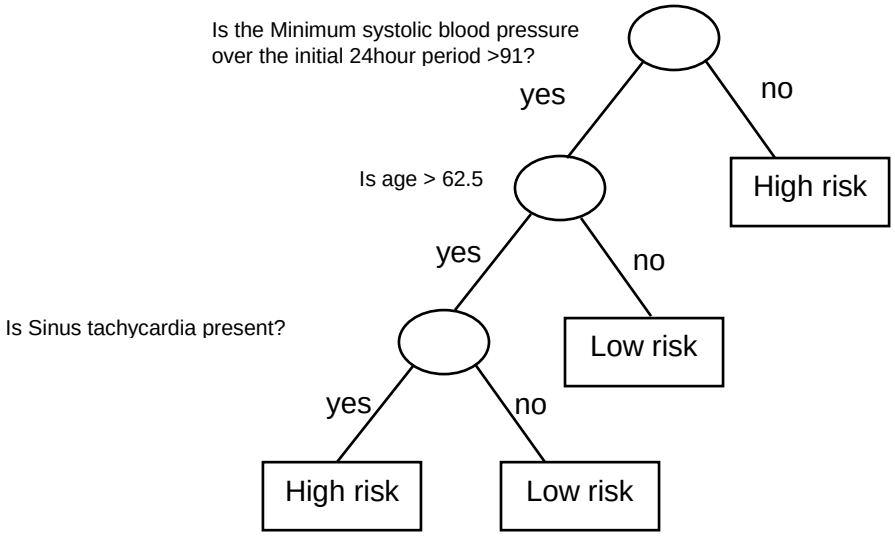


Figure 2.3 Decision tree structure

The classification techniques have benefits and drawbacks with respect to scenario or type of the dataset as shown in Table 2.3

Table 2.3 Benefit and drawback of different classification

Techniques	**Benefits**	**Drawback**
Decision Tree	• Building decision tree does not requires prior knowledge • Reduces anomaly and assign specific values to problem. • The diversity of data can easily be processed • Easy to understand • Numeric and categorical data are only processed	• Only requires one attributes. • It generates categorical output • Can be unstable because the data are dependent on the dataset features. • May suffer from overfitting • Classifies by rectangular partitioning
Bayesian Network	• Easy to implement • Can handle continuous and discrete data. • Not sensitive to irrelevant features • Lesser training data are required	• Dependency of the variable might result in inconsistency of the result • Computational infeasible • unautomated

2.2. Articles reviews

A 3-stage expert system (FS-PSO-SVM) based on a hybrid support vector machines method for diagnosing thyroid disease was developed by (HuiLing, 1963). The first stage (FS) tends

to construct various feature subsets with diverse discriminative capability. While in stage 2, the obtained feature subsets are used for the training of the designed SVM classifier, for training an optimal predictor model whose parameters are optimized using particle swarm optimization (PSO). The proposed system achieved the better classification accuracy stated so far by 10-fold cross-validation method, with the mean accuracy of 97.49% and with the maximum accuracy of 98.59%.

Ali Keles (2008) et al**.,** Proposed an Expert system for diagnosing thyroid disease (ESTDD).they found fuzzy rules by using neuro fuzzy method, which will be in ESTDD system. The accuracy of ESTDD is 95.33% while diagnosing thyroid diseases.

Depending on the size and attributes of the dataset classification techniques behave differently. The technique with the highest rate of accuracy and the lowest rate of error over a given dataset is selected as the better classification technique for that particular dataset. After applying diversified classification techniques the results obtained shows that SVM has the most favourable results for PIMA Indian Diabetes dataset followed by StatLog Heart Disease dataset with 96.74% and 99.25% rate of accuracy respectively, and C4.5 decision tree for BUPA Liver-disorders dataset has an accuracy rate of 79.71% whereas for Wisconsin Breast Cancer dataset Bayes Net, SVM, kNN and RBF-NN all have shown almost the same results with high rate of accuracy of 97.28% (Gupta, Kumar & Sharma, 2011).

The comparative analysis of SVM and K- Nearest Neighbor (KNN) in accuracy of predictions of Hypothyroid is done by (K. Saravana & Kumar, 2014**)** he has used SVM and KNN methods for collected dataset to predict hypothyroid and he clearly observed the prediction accuracy rate is 94.4336% in SVM and KNN has accuracy rate of 96.3430%. As the difference / variance is 1.9094. Therefore, he concluded that KNN has better performance than the SVM while working or predicting thyroid disease.

The experiment over the three datasets done by (**Rajeswara, Pellakuri & Ramya, 2015)** and the result shows that the FT algorithm is the better classifier between the opposite algorithms which are RandomForest, LMT and SimpleCart. But the results is restricted to the weka tool solely. What the experiments shows is that the accuracy between associate algorithms depends on the attribute number that dataset has. The results also can vary once if different tools like tanegra, rapid mining etc. are used on similar datasets, this experiment can also be expanded by using additional classification algorithms an additional datasets from various domains.

A good research in this field conducted by (**Hota & Dewangan, 2016)** used many machine learning techniques and are explored along with rank based feature selection technique to classify heart data. Experimental results are obtained using WEKA which shows that CART is performing better than other two DTs even after applying FST as 84.82% accuracy with only four features.

(Tejeswinee, Shomona, & Athilakshmi, 2017) carried out an investigation. The objective of their investigation was to examine the performance of the classification algorithms such as Support Vector Machine (SVM), Random Forest, Decision Tree, Naïve Bayes, Adaboost and K-NN, on the generated dataset. The accuracy of classification by these algorithms was measured using two units – Accuracy and Matthew's Correlation Coefficient (MCC) – and the results were tabulated. It was evident from the study that prior to feature selection, Random forest and K-NN classifiers predicted the diagnostic classes with high accuracy (~82%) when weighed against the other classification techniques. SVM gave the best accuracy (~94%) with CFS subset evaluation. In Gain Ratio method, Random Forest showed impressive results (~85%). It was followed by SVM and Decision tree classifiers.

The interesting systematic efforts are made in building the predicting system. During the building, three different classification techniques are studied and assessed on various measures. Experiments were done on Pima Indians Diabetes Database. The results determine the adequacy of the built system with an achieved rate of accuracy of 76.30 % using the Naive Bayes classification techniques. In upcoming moments, the built system with the used data mining classification techniques can be used precisely to predict or diagnose different diseases. Their work can be expanded also and improved for the automation analysis of the diabetes including some other data mining algorithms (**Sisodia & Sisodia, 2018)**.

Naïve Bayes classifier has shown its capability of better classification over categorical attributes, despites they used few dataset Naïve Bayes was able to approximate probabilities and correctly classified the given instance. In the future work, they will build the system over the big dataset with the size approximately 2,000 patients. The system was implemented in Java programming and support GUI facility also (**Aminu, Prasad & Mathias, 2018)**.

Safae Sossi Alaoui, Yousef Farhaoui and **Brahim Aksasse (2018)** carried out a research for choosing the suitable classifier for health problems datasets concerning the following diseases; breast cancer, diabetes and hypothyroidism, among the six powerful classification algorithms in Data Mining specifically; Simple Logistic, IBK, Naive Bayes, SGD, LMT and

SMO. Experimental results generated by the open source software Weka 3.8, demonstrate the high accuracy of LMT and Simple Logistic; with respectively an average of 83,84% and 82,88%, consequently they have proved their capacity to predict the class label efficiently for all datasets.

(Rady & Anwar, 2019) carried out similar study whereby they applied four data mining algorithms on a clinical/laboratory dataset consisting of 361 chronic kidney disease patients. The results of the addressed algorithms have been compared to define the most accurate algorithm results in classifying the severity stage of CKD. This study recommends that the Probabilistic Neural Networks algorithm is the best algorithm that can be used by physicians in order to eliminate diagnostic and treatment errors. Finally, they observed that the Probabilistic Neural Networks algorithm gives the highest overall classifcation accuracy percentage of 96.7%, compared to other algorithms in classifying the stages of CKD patients. On the other hand, the Multilayer Perceptron requires a minimum execution time (3 s) whereas the Probabilistic Neural Network requires 12 s to finalize the analysis.

CHAPTER THREE

MATERIAL AND METHOD

This section describes the methodology followed to fulfil our objectives in making a comparison between a set of classification algorithms in data mining. Secondly we picked out the suitable open source software for our work. Thirdly, we evaluated the models generated and lastly we examined each technique performance and boost all four classification algorithm.

3.0. Material and methods

In this work we have used the malaria's data collected from FMC Yola, as a training and testing dataset then we used Java programming language with Weka-api library to pre-process the dataset and build four (4) different classification models and then compared the result, we boosted all algorithm using ensembles method, we also tested the prediction of all the classification models with unknown class dataset, the process and tool used are illustrate in Figure. 3.1.

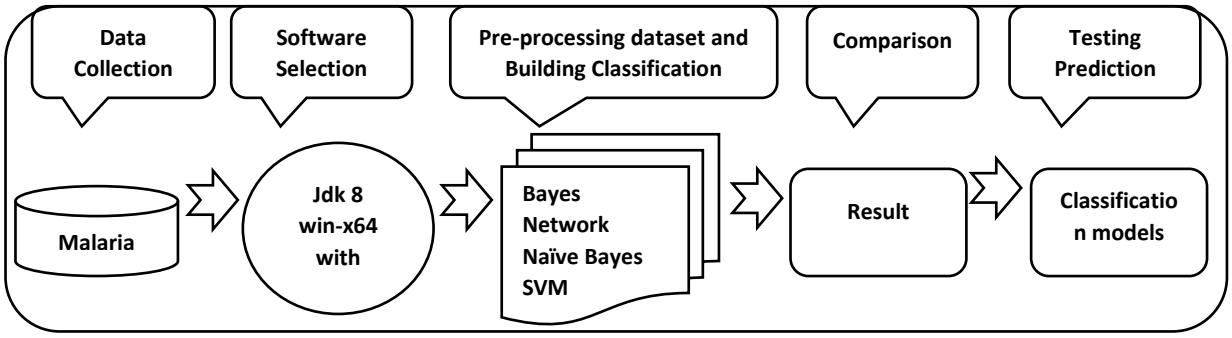


Figure 3.1 Description of the process and tools

3.1. Software's and hardware's selection

The improvement of knowledge discoveries in data in this years is driven by a many of available open source software in the market (Mikut and Reischl, 2011); In this paper, we had chosen the powerful object Oriented programming Language called Java, we used Java Development Kid (JDK 8) with latest Weka-api 3.8 library from (Waikato Environment for Knowledge Analysis), we imported weka-api library into the JDK in NetBeans editor which is a collection of machine learning classes, written in the Java programming language developed at the University of Waikato in New Zealand. In addition, we had used several classes of filtering to pre-process the collected dataset, and make it suitable for building the classification models. In fact, we had executed the classifications programs in a personal computer with the following features:

Processor Intel(R) core(TM) i5-2410M

CPU @ 2.30GHz

Installed Memory (RAM) 6.00GB

Windows 7 Ultimate, 64-bit Operating System

3.2. Data collection

Generally speaking, data collection is a very crucial task that influences every kind of research. Effectively, as we mentioned above my research employed a credible dataset from Federal Medical Center, Yola. The details of dataset are illustrated in Table 3.1.

Table 3.1 Description of unsupervised dataset

Name of the Dataset	Number of instance	Number of attributes	Types of attribute	Type of classification	Class variable	File Format
Malaria	699	6	Numeric and Nominal	Binary	Class	Comma Separated Value (.csv)

3.3. Dataset Pre-processing

Actually the Federal Medical Center, Yola were not keeping disease's diagnostic data electronically, so we had transformed the data to into the electronic medium with the help of MS-Excel 2013, the dataset consisted of six different attributes namely (FileNO, Fevar, headache, nausea, vomiting and class), and there is no dependency between the attributes, the first attribute is an Identity number of the patients visited the unit not more than six months and followed by four attributes that indicates the symptoms of malaria and last attribute is the class attribute that makes decision whether the patient has malaria or not based on the values of four different symptoms attributes, all the first five (5) attributes are of type Numeric attributes, we used 1 to represent the presence of a symptoms and 0 to indicate its absence, and null to indicate a particular field did not exist, the class attribute is of type nominal attribute that consist of two distinct values (YES and NO) categorically, it's called class attribute because it's based on its values that the classifier categorize each of the record of the dataset, we used YES to indicated and conclude the patients is malaria positive and NO to conclude malaria Negative.

So up to this step the dataset was in the Comma, Separated Value file (CSV Unicode UTF-8), the file was converted into the Attribute relation file format through the converter class which is the standard Java Weka-api library file format, we then used the filter class to automatically remove the irrelevant feature of the dataset, which means conversion of dataset from unsupervised to the supervised dataset. The new look of the supervised dataset is illustrated in Table 3.2.

Table 3.2 Description of supervised dataset

Name of the Dataset	Number of instance	Number of attributes	Types of attribute	Type of classification	Class variable	File Format
Malaria	699	5	Numeric and Nominal	Binary	Class	Attribute Relational File Format (.arff)

So from this step the dataset is suitable to be used in the building of classification models and the reason why we have chosen the classification techniques is that the class attribute of the dataset is of type nominal and its predictive dataset, had it been it was numeric class type and descriptive dataset we could have used regression techniques for comfort ability and suitability.

3.4. Opting and Building Mathematical models and Classification Algorithms

In this research, we choose the classification algorithm that uses a frequency table only such as Bayesian Network, Naïve Bayes, Decision tree, zeroR and OneR, as a result of no dependency among the attributes of the dataset Naïve Bayes will be used instead of Bayesian Network, as we all knew Naïve Bayes is subtype of the Bayesian Network and they both uses conditional probability and frequency table as shown in Figure 3.2.

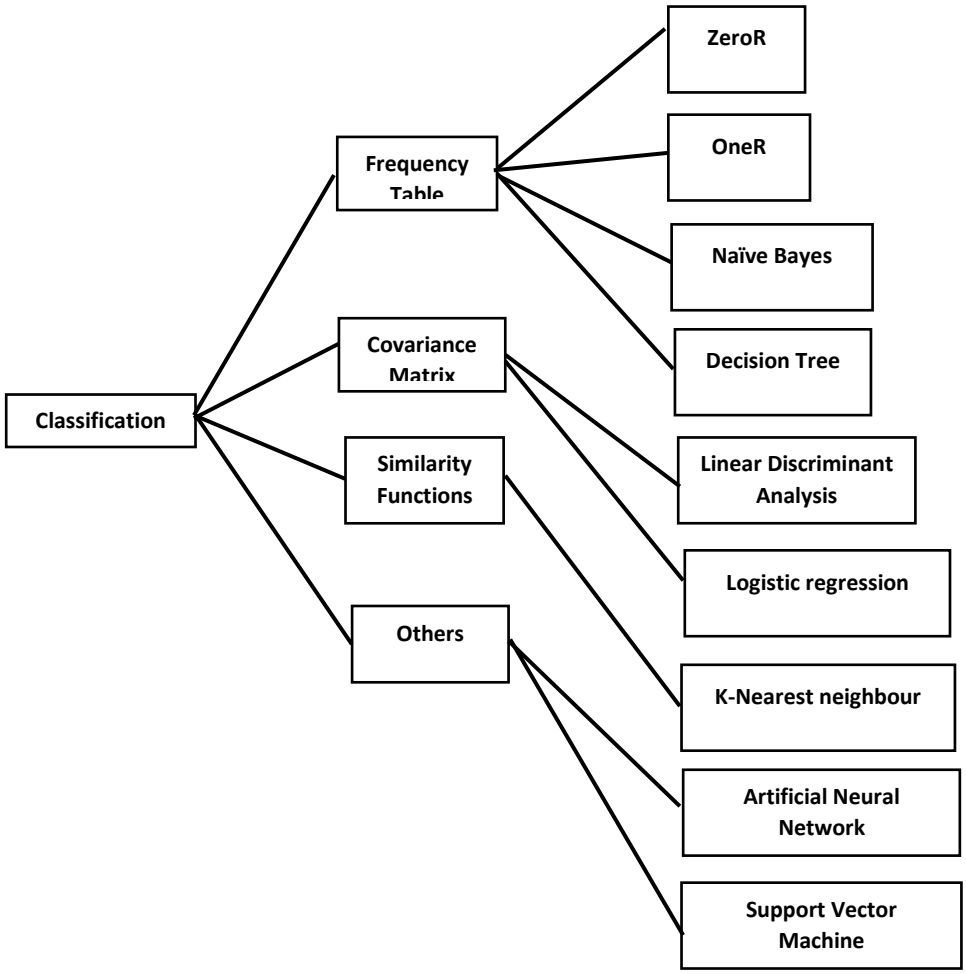


Figure 3.2 Different classifiers that uses frequency table

Before building the mathematical and classification model, Figure 3.3 will show the process of building the classification model and evaluation process

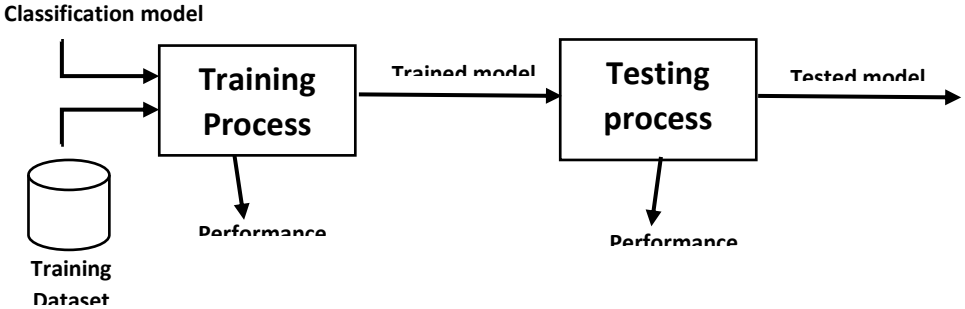


Figure 3.3 Model construction and evaluation in data mining

This figure above illustrated the building model and the evaluation process, now let see how the mathematical model of Bayesian network and other three different classifiers are built?

3.4.1 Bayesian Network Algorithm

A Bayesian network represents the causal probabilistic relationship among a set of random variables, their conditional dependences, and it provides a compact representation of a joint probability distribution, Murphy (1998). It consists of two major parts: a directed acyclic graph and a set of conditional probability distributions. The directed acyclic graph is a set of random variables represented by nodes. For health measurement, a node may be a health domain, and the states of the node would be the possible responses to that domain. If there exists a causal probabilistic dependence between two random variables in the graph, the corresponding two nodes are connected by a directed edge, Murphy (1998), while the directed edge from a node *A* to a node *B* indicates that the random variable *A* causes the random variable *B*. Since the directed edges represent a static causal probabilistic dependence, cycles are not allowed in the graph. A conditional probability distribution is defined for each node in the graph. In other words, the conditional probability distribution of a node (random variable) is defined for every possible outcome of the preceding causal node(s).

- Bayes Theorem.

For decades conditional probabilities of events of interest have been computed from known probabilities using Bayes' theorem.

Theorem 1.1 shows the formulae of conditional probability (Bayes) Given two events E and F such that P(E) 6= 0 and P(F) 6= 0, we have

$$\mathbf{P(E|F) = \frac{P(F|E)P(E)}{P(F)}} \qquad \mathbf{(1.1)}$$

Furthermore, theorem 1.2 have given mutually exclusive and exhaustive events E1, E2, . . . En such that P(Ei) 6=0 for all i, we have for $1 \leq i \leq n$,

$$\mathbf{(Ei \mid F) = \frac{P(F|Ei)P(Ei)}{P(F|E1)P(E1) + P(F|E2)P(E2) + \cdots P(F|En)P(En)}} \qquad \mathbf{(1.2)}$$

From the dataset five (5) different attribute exist including class attribute, suppose that Fever attribute, denoted by *F*, occurs P [*F*], Headache is denoted by *H*, occurs P [*H*], Nausea attribute, denoted by *N*, occurs P [*N*] and vomiting is donated by V, occurs P [*V*], It is

reasonable to assume the above attributes are independent. Furthermore, in this setting, the probability of patient having malaria P [*C* = P] and P[C = N]

3.4.2 Naïve Bayes Algorithm

Naive Bayes is a classification Algorithm with a concept which defines all attributes are independent and also unrelated to each other. It describes that status of a specific attribute in a class does not have effect on the status of another attribute. Since it is based on conditional probability it is measured as a powerful classifier employed for classification purpose. It's working well for the dataset that has unbalancing problems and lost values. Naive Bayes is a machine learning classifier which uses the Bayes Theorem of Conditional Probability. Using Bayes theorem posterior probability

***P*(*C* | *X*)** is calculated from **P(C), P(X)** and ***P*(*X* | *C*)**

Therefore, **P(C | X) = (P(X | C) P(C)) / P(X)**

Where,

***P*(*C* | *X*)** = target class's posterior probability .

***P*(*X* | *C*)** = predictor class's probability.

P(C) = class C's probability being true (before seeing any data).

P(X) = predictor's prior probability.

Naïve Bayes assumes that the effect of the value of a predictor (X) on a given class (C) is independent of the values of other predictors, as shown in figure 3.4, this assumption is called conditional independence.

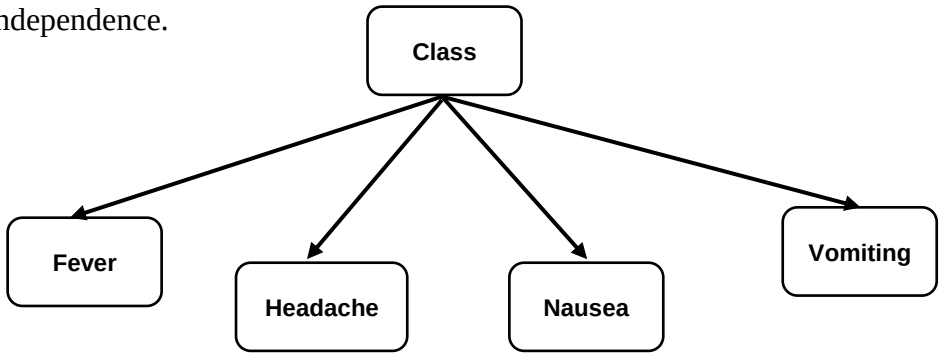


Figure 3.4 Structure of Naïve Bayes constructed from malaria dataset

The formulae will be used to develop Naive Bayes model below:

- From the malaria dataset the posterior probability can be calculated by first constructing a frequency table for each attribute against the target as illustrated in Table 3.3.
- Then transform the frequency table to likelihood table in Table 3.4, and finally using the naïve Bayes equation to calculate the posterior probability of the class
- The class with the highest posterior probability is the outcomes of the prediction

Table 3.3 Frequency table of Naïve Bayes

P(p)=426/699

P(n)= 273/699

The above is the class C's probability (before seeing any data).

Frequency Table		Class	
		P	N
Fever	1	403 /426	118/273
	0	22/426	155/ 273

Frequency Table		Class	
		P	N
Headache	1	278 /426	93/273
	0	146/426	180/ 273

Frequency Table		Class	
		P	N
Nausea	1	205 /426	85/273
	0	218/426	118/ 273

Frequency Table		Class	
		P	N
Vomiting	1	187 /426	97/273
	0	237/426	176/ 273

Table 3.4 Likelihood table of Naïve Bayes

Likelihood Table		Class		
		P	N	
Fever	1	403 /426	118/273	521/699
	0	22/426	155/ 273	177/699
		425/699	273/699	

Likelihood Table		Class		
		P	N	
Headache	1	278 /426	93/273	371/699
	0	146/426	180/ 273	326/699
		424/699	273/699	

Likelihood Table		Class		
		P	N	
Nausea	1	205 /426	85/273	290/699
	0	218/426	118/ 273	336/699
		423/699	203/699	

Likelihood Table		Class		
		P	N	
Vomiting	1	187 /426	97/273	284/699
	0	237/426	176/ 273	413/699
		424/699	273/699	

et assume that has the patient is recognised with the following symptoms:

Fever = 1

Headache = 1

Nausea = 1

Vomiting =1

Likelihood of Yes and No can be calculated the table above

Likelihood of Yes =P(Fever | 1) * P(Headache | Yes) * P(Nausea | 1) * P(Vomiting | 1) * P(Yes)=

403/426 * 278/426 * 205/426 * 187/426 * 426/699 =0.079676

Likelihood of No =P(Fever | 0) * P(Headache | 0) * P(Nausea | 0) * P(Vomiting | 0) * P(No)=

187/273 * 93/273 * 85/273 * 97/273 * 273/699 = 0.010082

Now likelihood of Yes and No will be normalised to get the probability of Yes and No

P(Yes) = 0.079676/(0.079676+0.010082) = 0.887675

P(No) = 0.010082/(0.079676+0.010082) = 0.112324

Based on the symptoms the probability is calculated and the result of prediction shows that the P(Yes) > P(No), means patient has malaria.

3.4.3 Decision Tree

Decision tree build classification or regression model in the form of tree structure, It breaks the dataset down into smaller subset while at the same time an association decision tree is incrementally developed

- The final result a tree decision node and leaf nodes
- A decision node has two or more branches
- Leaf node represent the classification or decision
- The topmost decision of the tree represent or correspond to the best predictor called root note
- Decision tree can handle both categorical and numerical data

Baefore constructing decision tree Figure 3.5 structure of the dataset.

Predictors Target

fever	headache	nausea	vomiting	class
1	0	0	1	N
1	1	1	1	P
1	1	0	1	P
1	0	0	1	P
1	1	0	0	P
1	1	0	0	P
1	1	0	1	P
1	1	0	0	P
1	0	0	0	N
1	1	1	0	P
1	1	1	0	P
1	1	0	0	P
0	0	1	1	N
1	1	0	0	P
1	0	0	1	N
1	1	1	0	P
0	1	0	0	N

Figure 3.5 Structure of dataset before constructing Decision tree

The core algorithm for building trees is called ID3 by J.R Quinlan which employs a top down, greedy search through the space of possible branches with no backtracking,

ID3 use entropy and information gain to construct a decision tree.

A decision tree is built top-down from the root node and involves partitioning the data into subset that contains instances with the similar values (homogeneous)

ID3 algorithm uses entropy to calculate the homogeneity of a sample

If the sample is completely homogeneous as shown in Figure 3.6, the entropy is zero and if the sample is an equally divided it has entropy of one.

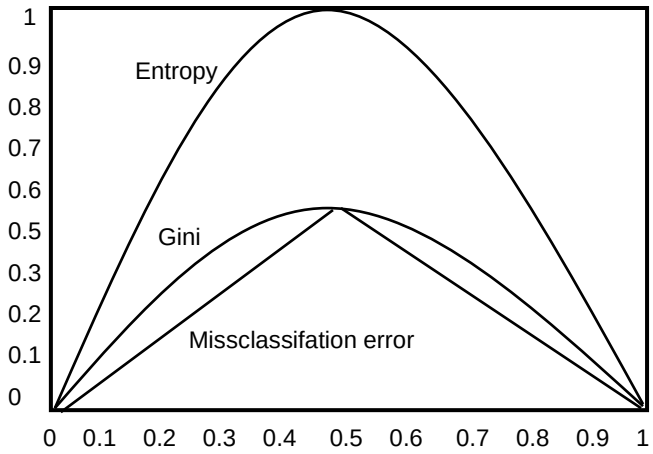


Figure 3.6 Entropy and Gini

To build a decision tree there is need to calculate two types of entropy using frequency Table 3.5.1 as follows:

Entropy using the frequency table of the one attribute

A) (Entropy of the target):

$$E(S) = \sum_{i=1}^{c} -pi\ log2\ pi$$

Table 3.5.1 Frequency table for the target

Class	
Yes	No
426	273

Entropy (Class) = Entropy (273, 426)

=Entropy (0.293, 0.609)

= - (0.293 $\log_2$ 0.293) – (0.609 $\log_2$ 0.609)

E (Class)= **0.28737**

Entropy using the frequency table of two attributes:

$$E(T,X) = \sum_{ceX} P(c)E(c)$$

We choose to split using fever attribute in Table 3.5.2 at initial

Table 3.5.2 Frequency table for the Fever

Frequency Table		Class		
		Yes	No	
Fever	1	403 /426	118/273	521
	0	22/426	155/273	177
				699

E(Class, Fever)= P(1)*E(403,118)+P(0)*E(22,155)

=(521/699)*0.27+(177/699)*0.12

E(Class, Fever)=**0.23**

We selected Fever attribute to be the root note of the decision tree from Figure 3.6

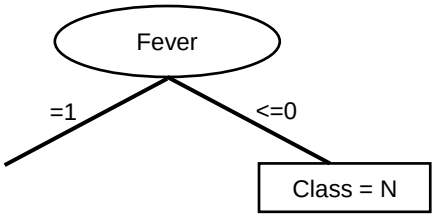


Figure 3.6 shows the initial step of constructing Decision tree (J48)

Table 3.5.3 Frequency table for the Headache

Frequency Table		Class		
		Yes	No	
Headache	1	278 /426	93/273	366
	0	146/426	180/ 273	326
				699

From the frequency Table 3.5.3, E (Class, Headache) = P (1)*E (278,93)+P(0)*E(146,180)

= (366/699)*0.27+ (326/699)*0.12

E (class, Headache) = **0.12**

Table 3.5.4 Frequency table for the Nausea

Frequency Table		Class		
		Yes	No	
Nausea	1	205 /426	85/273	290
	0	218/426	118/ 273	336
				699

From the frequency Table 3.5.4, (Class, Nausea)= P (1) *E (205,85)+P(0)*E(218,118)

= (290/699)*0.25 + (336/699)*0.12

E (Class, Nausea) = **0.16**

Table 3.5.5 Frequency table for the Vomiting

Frequency Table		Class		
		Yes	No	
Vomiting	1	187 /426	97/273	284
	0	237/426	176/ 273	413
				699

From the frequency Table 3.5.5, E (Class, Vomiting) = P(1)*E(187,97)+P(0)*E(237,176)

= (284/699)*0.28 + (413/699)*0.3

E (Class, Vomiting) = **0.29**

The weighted average of the Gini Index for the decedent nodes is

For Headache 0.61*0.12 + 0.4*0.12 =**0.12**

For Nausea 0.6*0.16 + 0.3*0.16 = **0.15**

For Vomiting 0.6*0.29 + 0.4*0.29 = **0.29**

Since the subset for attribute Headache has smaller Gini Index among the attribute, is prepared over Nausea and vomiting attributes. The base on our calculation we draw the decision tree in Figure 3.7

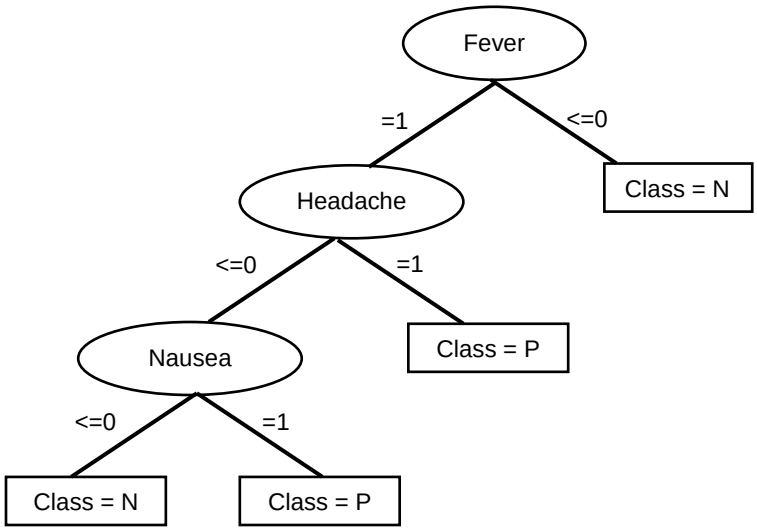


Figure 3.7 progressive and complete Decision Tree (J48)

3.4.4. ZeroR

ZeroR is the simplest classification method which relies on the target and ignores all the predictors, zeroR classifier is simple predicts the majority category (class)

It's very useful for determining a baseline performance as a benchmark for other classification methods.

It's also using a frequency table for the target and select its most frequent value as illustrated in Figure 3.8.

Figure 3.8 Target's frequency table

Class = Yes is the ZeroR model for the above dataset with accuracy of **0.6**

3.4.5 OneR

One Rule, short for "One Rule", is a simple, yet accurate classification algorithm. It generate one rule for each predictor in the data, then select the rule with the smallest total error as its "One rule"

To create a rule for a predictor, there is need to construct a frequency table for each predictor against the target

The algorithm is below:

For each predictor,

For each value of the predictor, make a rule as follows:

Count how often each value of target (Class appears)

Find the most frequent class

Make the rule assign that class to this value of the predictor

Calculate the total error of the each predictor

Choose the predictor with smallest total error

Table 3.5.6 Frequency table for of Dataset

Frequency Table		Class	
		Yes	No
Fever	1	403	118
	0	22	155

Frequency Table		Class	
		Yes	No
Headache	1	278	93
	0	146	180

Frequency Table		Class	
		Yes	No
Nausea	1	205	85
	0	218	118

Frequency Table		Class	
		Yes	No
Vomiting	1	187	97
	0	237	176

IF Fever = 1 THEN Class =Yes

IF Fever = 0 THEN Class =No

Predictors Contribution

Simply, the total error calculated from the frequency table 3.5.6 is the measure of each predictor contribution, therefore a low total error means a higher contribution to the predictability of the model.

And the attribute with higher contribution is Fever therefore is selected based on the lowest total error given among the rest attributes

3.5. Evaluation Metrics

In practice, there are two types of dataset; the training dataset; which used by the classifier to build up the model by learning from the data, and the testing dataset (known as validation dataset); which tends to guess the performance of the predictive model. In this case, we had selected for K-fold cross-validation (shown in Figure 3.9) which counts on partitioning the training dataset into k subsets with equal sizes. In each iterations, one part is always reserved to validation dataset and the rest k-1 splits are reserved as training data.

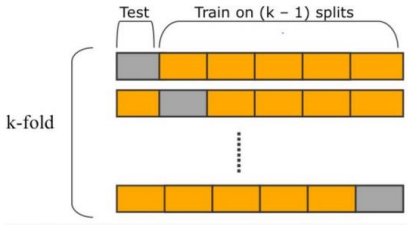


Figure 3.9 Cross validation

In addition, we outline an overview of the different metrics used in evaluating the chosen classifiers. In general, measuring accuracy, speed, scalability, interpretability and robustness of algorithms can be done by comparison between Classifiers. Sossi (2017). The work consists of taking into account the **accuracy** of classification algorithms which is defined as the capacity of any classifier to predict the class label efficiently, then the **speed** which means the time required to produce and build the classification model and the **scalability** which describes the ability to build a model efficiently when the classifier is applied to a huge dataset (STEFANOWSKI, 2008).

Furthermore, **Confusion matrix.** Provost and Kohavi, (1998) is method which presents the predicted and actual classification. Thus, multiple standards are presented based on confusion matrix such as **Precision**, **Recall** and **F-Measure**. Formulas are outlined in Figure 3.9.1 with 2*2 confusion matrix.

- **True Positive Rate (TPR) or Sensitivity (Recall):** A true positive is when the outcome of a prediction is said 'P' and the classifier have actually predicted the value to be same 'P'. It is a measure of comprehensiveness or magnitude.

  True Positive Rate (TPR) or sensitivity = ∑ true positive / ∑ conditional positive

- **True Negative (TNR) or Specificity:** is also called *specificity*, it indicates the number of tuples classified as false while they were actually false. It is also the ability of the classifier to classify those that were false correctly
- **False Positive (FPR):** is when a record is classified to be false while is actually supposed to be predicted as true.
- **False Negative (FNR):** It signifies the number of tuples classified as false while they were actually true.

- **Precision:** Precision is the portion of retrieved cases that are relevant. It is the measure of correctness or excellence.
  Precision = TP/ (TP + FP).
- **Accuracy**; is the ability of a model to appropriately predict the class label of previously unseen data or new data. It is a measure of how well the classifier makes a prediction on average. A good classification algorithm will try to minimize the number of times it makes the wrong prediction.
  Accuracy = (TP+TN) / (TP+FP+FN+TN).

3.6. Boosting Classifier Method

Boosting is a general method which attempts to boost the accuracy of any given learning algorithm, it was originally designed for classification problems but it can profitably be extended to regression as well. Boosting [Freund Schapire1996, Schapire1990] encompasses a family of methods.

The focus of boosting methods is to produce a series of "weak" learners in order to produce a powerful combination.

A weak learner is a learner that has accuracy only slightly better than chance.

Idea: given a weak learner, run it multiple times on (reweighted) training data, then let the learned classifiers vote

- On each iteration t:
  - weight each training example by how incorrectly it was classified
  - Learn a hypothesis
  - A strength for this hypothesis
- Final classifier:
  - A linear combination of the votes of the different classifiers weighted by their strength

The training set used for each member of the series is chosen based on the performance of the earlier classifier(s) in the series.

Examples that are incorrectly predicted by previous classifiers in the series are chosen more often than examples that were correctly predicted.

Thus Boosting attempts to produce new classifiers that are better able to predict examples for which the current ensemble's performance is poor.

Unlike Bagging, the resampling of the training set is dependent on the performance of the earlier classifiers.

In boosting by sampling for classification, this two AdaBoost are used: AdaBoost.M1 for two-class problem and AdaBoost.M2 for multiple-class problem as well (Freund & Schapire, 1996) and boosting using regression is done by AdaBoostR (Drucker, 1997)

AdaBoost Algorithm

1. Initialize the observation weights

$w_i = 1/N$, i = 1, 2, . . . , N.

2. For m = 1 to M repeat steps (a)–(d):

(a) Fit a classifier $G_m(x)$ to the training data

using weights w_i.

(b) Compute $\sum_{i=1}^{N} wi\ I(yi \neq Gm(xi)) / \sum_{i=1}^{N} wi$

(c) Compute $\alpha_m = \log((1 - err_m) / err_m)$.

(d) Update weights for i = 1, ..., N:

$w_i \leftarrow w_i \cdot \exp[\alpha_m \cdot I(y_i != Gm(x_i))]$

and renormalize to w_i to sum to 1.

Output

$$G(x) = sign[\sum_{m=1}^{m} \alpha m\ Gm(x)]$$

CHAPTER FOUR

RESULTS AND DISCUSSIONS

4.0. Presentation of the result

Different experimental results are found to compare the assigned classification algorithms mainly Naïve Bayes, Decision tree, ZeroR and OneR, on the malaria dataset introduced previously and using the process of 10-folds cross validation in JDK8 with imported library of Weka-API . Firstly, the accuracy percentages are illustrated in Figures 4.1

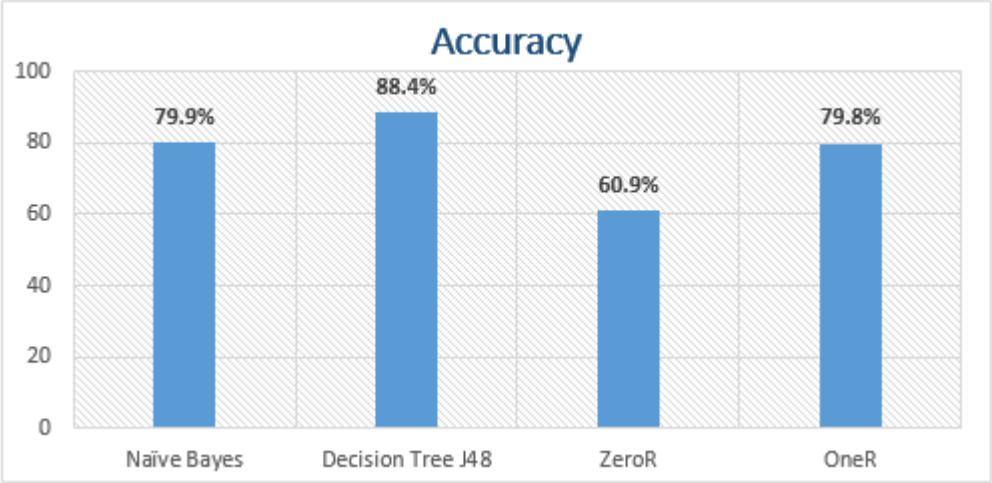


Figure 4.1. Accuracy chart of the constructed classifier model

Figure 4.1 declares the accuracy obtained in training Malaria dataset by the 4 different classification algorithms. It indicates that Decision tree (J48) has the greater accuracy with 88.4% followed by Naïve Bayes with 79.9%, then OneR and ZeroR with the accuracy of 79.8% and 60.9% respectively.

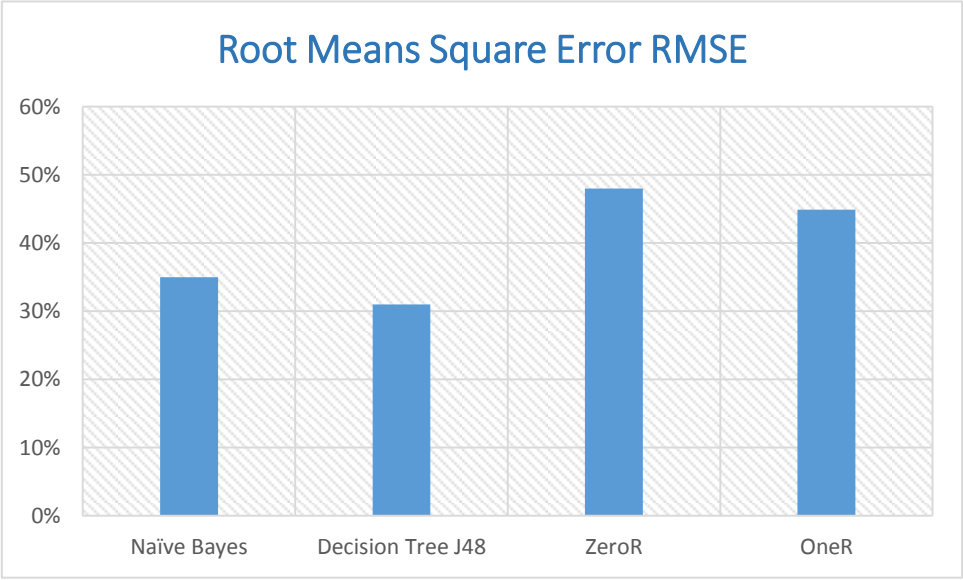
Table 4.1 Percentages of correctly and incorrectly

SN	Classification Algorithm	Correctly Classified Instance	Incorrectly Classified Instance
1.	Naïve Bayes	79.9%	20.1%
2.	Decision Tree (J48)	88.4%	11.6%
3.	ZeroR	60.9%	39.1%
4.	OneR	79.8%	20.2%

Table 4.1 shows the percentages of correctly and incorrectly classified instances for each Algorithm building using the malaria datasets. Indeed, the results indicates that the Decision tree (J48) is having the highest correctly classified instance followed by Naïve Bayes then OneR and ZeroR.

Table 4.2 Performance summary of 4 classifiers

SN	Classification Algorithm	Kappa Statistic	Mean Absolute Error	Root Mean Square Error	Relative Absolute Error	Root Relative Sq. Error	Speed
1.	Naïve Bayes	55%	27%	35%	57.1%	73.4%	0.2s
2.	Decision Tree (J48)	76%	19%	31%	41%	63.9%	0.3s
3.	ZeroR	0%	47%	48%	100%	100%	0.1s
4.	OneR	54.8%	20%	44.9%	42.3%	92%	0s



Figures 4.2 Root mean square error (RMSE) for the 4 different classification algorithm using malaria datasets

The Root mean square error (RMSE) as shown in Figure 4.2, is used as a factor to verify experimental results, RMSE indicates the standard deviation of the differences between predicted figures and real or observed figures.

Figure 4.2 illustrates that Decision tree have the lowest RMSE with percentage value of **31%**, then followed by Naïve Bayes with 35% also ZeroR and OneR has RMSE of 48% and 44.9% respectively.

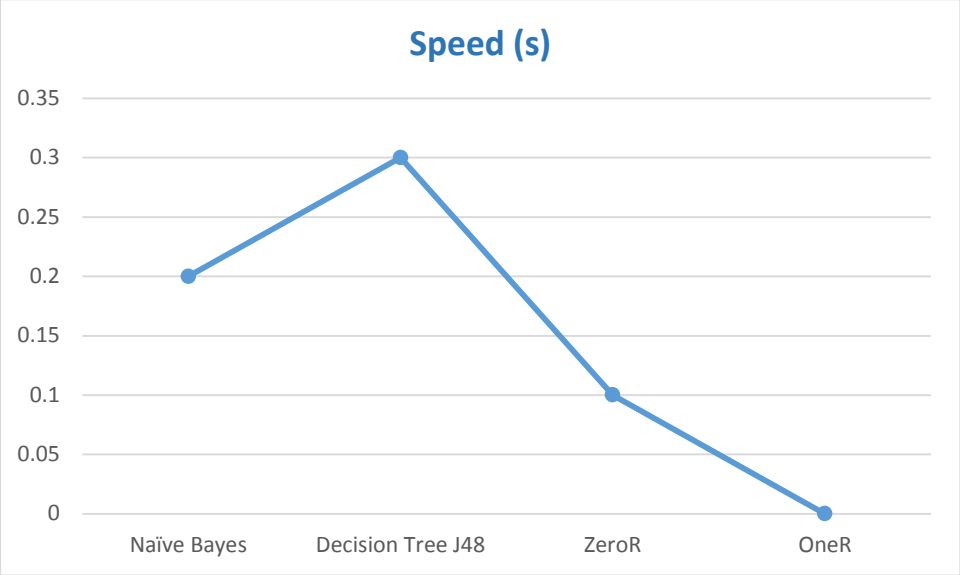


Figure 4.3 Time taken to build a classifier

Figure 4.3 clarifies that oneR is faster in speed then the rest, followed by ZeroR then Naïve Bayes and the lowest speed is Decision tree classifier.

Table 4.3 Performance related to confusion matrix

SN	Classification Algorithm	Precision	Recall	F-Measure
1.	Naïve Bayes	0.77	0.94	0.85
2.	Decision Tree (J48)	0.93	0.87	0.90
3.	ZeroR	0.6	1.0	0.75
4.	OneR	0.77	0.94	0.85

After we performed the ensemble method, boosting Naïve Bayes Classifier using Adaboost m1 we found out its accuracy increases from 79.9% to 87.0% and correctly classified instance increases to 87% and incorrectly classified instance reduces to 13.0%, the rest of performance metrics are shown in Table 4.4

Table 4.4 Performance of Naïve Bayes before and after boosting

SN	Performance Metrics	Naïve Bayes (before Bosting)	Naïve Bayes (after Bosting)
1.	**Correctly Classified Instance**	79.9%	87.0%
2.	**Incorrectly Classified Instance**	20.1%	13.0%
3.	**Kappa Statistic**	55%	72.9%
4.	**Mean Absolute Error**	27%	18.1%
5.	**Root Mean Square Error**	35%	30.1%
6.	**Relative Absolute Error**	57.1%	38.2%
7.	**Root Relative Sq. Error**	73.4%	61.7%
8.	**Speed (s)**	0.2s	0.1s
9.	**Precision**	0.77	0.86
10.	**Recall**	0.94	0.93
11.	**F-Measure**	0.85	0.90

Table 4.5 Performance of Decision Tree J48 before and after boosting

SN	Performance Metrics	Decision Tree (before Boosting)	Decision Tree (after Boosting)
1.	Correctly Classified Instance	88.4%	88.4.%
2.	Incorrectly Classified Instance	11.6%	11.6%
3.	Kappa Statistic	76%	76%

4.	Mean Absolute Error	19%	19%
5.	Root Mean Square Error	31%	31%
6.	Relative Absolute Error	41%	41%
7.	Root Relative Sq. Error	63.9%	63.9%
8.	Speed (s)	0.3s	0.3s
9.	Precision	0.93	0.93
10.	Recall	0.87	0.87
11.	F-Measure	0.90	0.90

Table 4.6 Performance of ZeroR before and after boosting

SN	**Performance Metrics**	**Zero R (before Boosting)**	**Zero R (after Boosting)**
1.	Correctly Classified Instance	60.9%	60.9%
2.	Incorrectly Classified Instance	39.1%	39.1%
3.	Kappa Statistic	0%	0%
4.	Mean Absolute Error	47%	47%
5.	Root Mean Square Error	48%	48%
6.	Relative Absolute Error	100%	100%
7.	Root Relative Sq. Error	1000%	100%

8.	Speed (s)	0.1s	0.1s
9.	Precision	0.66	0.66
10.	Recall	1.0	1.0
11.	F-Measure	0.75	0.75

Table 4.7 Performance of OneR before and after boosting

SN	**Performance Metrics**	**One R (before Bosting)**	**One R (after Bosting)**
1.	Correctly Classified Instance	79.8%	87.6%
2.	Incorrectly Classified Instance	20.2%	12.40%
3.	Kappa Statistic	54.8%	73.7%
4.	Mean Absolute Error	20%	20.3%
5.	Root Mean Square Error	44.9%	30.1%
6.	Relative Absolute Error	42.3%	42.7%
7.	Root Relative Sq. Error	92%	61.7%
8.	Speed (s)	0.s	0.1s
9.	Precision	0.77	0.86
10.	Recall	0.94	0.94
11.	F-Measure	0.85	0.90

Table 4.8 Comparison and change of accuracy of all four classification models before and after boosting

S/N	Classifier	Accuracy (before boosting)	Accuracy (After boosting)	% Change
1	Naïve Bayes	79.9%	87.0%	7.1%
2	Decision Tree	**88.4%**	**88.4.%**	0%
3	Zero R	60.9%	60.9%	0%
4	One R	79.8%	87.6%	7.8%

CHAPTER FIVE

CONCLUSION AND FEATURE WORK

5.0. Conclusion

This research focused on building Naïve Bayes and three different classification model namely; Decision tree, ZeroR and OneR, choosing the most suitable model among them for health problems concerning malaria disease prediction. Experimental results generated using JDK8 packages with the open source Weka-API 3.8 Library, have shown that the Decision tree has higher accuracy average of 88.4% (Figure 4.1) consequently it has shown its capacity to predict the class label efficiently and accurately for malaria dataset. Also, the algorithm has the lowermost average of root mean square error (RMSE); 31% (Figure 4.2) in comparison with other classification algorithm. However, the four models are all boosted using AdaBoost m1 to see whether we can achieve more optimised accuracy, Naïve Bayes and OneR are boosted successfully (Table 4.4 and 4.7) and their accuracies have increased but still they are not greater than the Decision Tree, ZeroR and Decision tree cannot be Boosted using AdaBoost as shown in Table 4.5 their result before and after boosting are the same, then in Table 4.3 OneR and ZeroR are faster than Naïve Bayes and Decision tree because in reality; decision tree methods have tendency to be slow whenever the number of class are too huge. Therefore, ZeroR model is the worst accurate because it uses as performance base line for other classifiers, in short, based on our research we identify that Decision tree (J48) is the appropriate classification technique because of accuracy level shown in (Figure 4.1) followed by Naïve Bayes which they can be the great solution for Malaria prediction and decision making in Healthcare System. Performance of the classifiers was also checked majorly using confusion matrix. Performance of the classifiers was also checked majorly using confusion matrix.

5.1. Limitation of Study

The research study is limited to be used only on malaria cases within health care environment and is limited to only Classification models that use a frequency table, and uses only one of the method for Increasing performance of classification algorithm which is Boosting, the study also highlighted the significance and power of Decision tree (J48) over Naïve Bayes, ZeroR and OneR classifiers in supervised learning.

5.2. Feature Work

This study can further be experimented by using other ensembles methods such as bagging, stacking and voting to increase in performance of the classification algorithm on

similar disease and another classification technique as well, one or more attributes or predictor can be added for better performance. The experiment can be done using different programming language. The performance can also be checked using other performance evaluation metrics made for checking the accuracy and evaluating a given classification model such as *Gain and Lift Chart, Gini Coefficient, Kolmogorov Smirnov Chart, Concordant (Discordant Ratio).*

BIBLIOGRAPHY

Jill Seladi-Schulman, P. (2018). *Swine flu: Causes, symptoms, and treatment*. [online] Medical News Today. Available at: https://www.medicalnewstoday.com/articles/147720.php [Accessed 20 Nov. 2018].

K. Sharmila & Vethamanickam, (2015) Survey on Data Mining Algorithm and Its Application in Healthcare Sector Using Hadoop Platform. *International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 5, Issue 1, January 2015)*

Chen, H., Yang, B., Wang, G. et al. A Three-Stage Expert System Based on Support Vector Machines for Thyroid Disease Diagnosis. *Journal of Medical Systems* 36, 1953–1963 (2012). https://doi.org/10.1007/s10916-011-9655-8

K. Saravana & Kumar, (2014) Support vector machine and K- nearest neighbor based analysis for the prediction of hypothyroid. January 2014 *International Journal of Pharma and Bio Sciences 5(4):B447-B453*

Osmar, (1999). Principles of Knowledge Discovery in Data. University of Alberta

Rajeswara, Pellakuri & Ramya (2015). Performance Analysis of Classification Algorithms Using Healthcare Dataset. *D. Rajeswara Rao et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1103-1106*

Ahmad, P., Qamar, S., & Qasim Afser Rizvi, S. (2015). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications*, *120*(15), 38–50. https://doi.org/10.5120/21307-4126

Gupta, S., Kumar, D., & Sharma, A. (2011). *P Erformance a Nalysis of V Arious D Ata M Ining C Lassification T Echniques on*. *3*(4), 155–169.

Hota, H. S., & Dewangan, S. (2016). *Classification of Health Care Data Using Machine Learning Technique*. *5*(9), 17–20.

Joshi, S., & Nair, M. K. (2018). Survey of Classification Based Prediction Techniques in Healthcare. *Indian Journal of Science and Technology*, *11*(15), 1–19. https://doi.org/10.17485/ijst/2018/v11i15/121111

N, Y., & S, M. (2016). A Review on Text Mining in Data Mining. *International Journal on*

Soft Computing, *7*(2/3), 01–08. https://doi.org/10.5121/ijsc.2016.7301

Rady, E. H. A., & Anwar, A. S. (2019). Prediction of kidney disease stages using data mining algorithms. *Informatics in Medicine Unlocked*, *15*(December 2018), 100178. https://doi.org/10.1016/j.imu.2019.100178

Rajeswara rao, D., Pellakuri, V., & Ramya Harika, T. (2015). Performance Analysis of Classification Algorithms Using Healthcare Dataset. *(IJCSIT) International Journal of Computer Science and Information Technologies*, *6*(2), 1103–1106. Retrieved from http://ijcsit.com/docs/Volume 6/vol6issue02/ijcsit2015060239.pdf

Rani, B. K., & Govrdhan, A. (2010). *Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks*. *02*(02), 250–255.

Sidiq, U., & Khan, R. A. (2017). Data Mining for diagnosis in Healthcare Sector-a review. *International Journal of Advances in Scientific Research and Engineering*, *3*(8), 1–9. https://doi.org/10.7324/ijasre.2017.32486

Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, *132*(Iccids), 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

Tejeswinee, K., Shomona, G. J., & Athilakshmi, R. (2017). Feature Selection Techniques for Prediction of Neuro-Degenerative Disorders: A Case-Study with Alzheimer's and Parkinson's Disease. *Procedia Computer Science*, *115*, 188–194. https://doi.org/10.1016/j.procs.2017.09.125

Aminu Aliyu (2018). A Framework for Predicting Malaria using Naïve Bayes Classifier.

Al-Hassan, N. A., & Roberts, G. T. (2002). Patterns of presentation of malaria in a tertiary care institute in Saudi Arabia. *Saudi Medical Journal*, *23*(5), 562–567.

Al-radaideh, Q. A. & Nagi, E.A. (2012). Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance, *(IJACSA) International Journal of Advanced Computer Science and Applications*, *3*(2), 144–151.

Ameta, M. A., & Jain, M. K. (2017). Data Mining Techniques for the Prediction of Kidney Diseases and Treatment : A Review, *International Journal Of Engineering And Computer Science 6*(2), 20376–20378. https://doi.org/10.18535/ijecs/v6i2.37

Archana, S., & Elangovan, K. (2014). Survey of Classification Techniques in Data Mining,

International Journal of Computer Science and Mobile Applications 2(2), 65–71.

Arévalo-Herrera, M., Lopez-Perez, M., Medina, L., Moreno, A., Gutierrez, J. B., & Herrera, S. (2015). Clinical profile of Plasmodium falciparum and Plasmodium vivax infections in low and unstable malaria transmission settings of Colombia. *Malaria Journal*, *14*(1), 154. https://doi.org/10.1186/s12936-015-0678-3

Baby, M. N., & Priyanka, L. T. (2012). Customer Classification And Prediction Based On Data Mining Technique, *International Journal of Emerging Technology and Advanced Engineering 2*(12), 314-318.

Bartoloni, A., & Zammarchi, L. (2012). Clinical aspects of uncomplicated and severe malaria. *Mediterranean Journal of Hematology and Infectious Diseases*, *4*(1). https://doi.org/10.4084/MJHID.2012.026

Bhardwaj, B. K. & Pal, S. (2011). Data Mining : A prediction for performance improvement using classification, *(IJCSIS) International Journal of Computer Science and Information Security*, *9*(4).

Bohra, H., Arora, A., Gaikwad, P., Bhand, R., & Patil, M. R. (2017). Health Prediction and Medical Diagnosis using Naive Bayes, *International Journal of Advanced Research in Computer and Communication Engineering* ISO *6*(4), 32–35. https://doi.org/10.17148/IJARCCE.2017.6407

Borgwardt, H. K. K. M., Kröger, P., Pryakhin, A., & Zimek, S. A. (2007). Future trends in data mining. *Data Mining and Knowledge Discovery*, *15*(1), 87–97. https://doi.org/10.1007/s10618-007-0067-9

Calderaro, A., Piccolo, G., Gorrini, C., Rossi, S., Montecchini, S., Loretana, M., … Arcangeletti, M. C. (2013). Accurate identification of the six human Plasmodium spp . causing imported malaria , including Plasmodium ovale wallikeri and Plasmodium knowlesi, *Malaria Journal*, *1*2(1), 1–6.

Cdc, C. F. D. C. and P. (2013). Treatment of Malaria (Guidelines For Clinicians). *Treatment of Malaria (Guidelines for Clinicians)*, (July), 1–8. https://doi.org/10.1016/S0140-6736(05)66420-3

Chai, T., Draxler, R. R., & Prediction, C. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,

Geoscientific Model Development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Cherian, V., & Bindu, M. S. (2017). Heart Disease Prediction Using Naïve Bayes Algorithm and Laplace Smoothing Technique, *International Journal of Computer Science Trends and Technology (IJCST) 5*(2), 68–73.

Chotivanich, K., Silamut, K., & Day, N. P. (2007). Laboratory diagnosis of malaria infection-A short review of methods. *New Zealand Journal of Medical Laboratory Science*, *61*(1), 4.

Dangare, C. S., & Cse, M. E. (2012). Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, *International Journal of Computer Applications 47*(10), 44–48.

Durairaj, M., & Ranjani, V. (2013). Data Mining Applications In Healthcare Sector: A Study. *International Journal Of Scientific & Technology Research* Volume 2, Issue 10, October 2013 Issn 2277-8616.

Femi Aminu, E., Ogbonnia, E. O., & Shehu, I. S. (2016). A Predictive Symptoms-based System using Support Vector Machines to enhanced Classification Accuracy of Malaria and Typhoid Coinfection. *International Journal of Mathematical Sciences and Computing*, *2*(4), 54–66. https://doi.org/10.5815/ijmsc.2016.04.06

Ghumbre, S., Patil, C., & Ghatol, A. (2011). Heart Disease Diagnosis using Support Vector Machine. *International Conference on Computer Science and Information Technology (ICCSIT), pg 84-88*

Gu, X., Chen, H., & Yang, B. (2015). Heterogeneous data mining for planning active surveillance of malaria. In *Proceedings of the ASE BigData & SocialInformatics 2015* (p. 34). ACM.

Hand, D. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems, *Machine Learning,* vol. 45, 171–186.

Howes, R. E., Jr, R. C. R., Battle, K. E., Longbottom, J., Mappin, B., Ordanovich, D., … Hay, S. I. (2015). Plasmodium vivax Transmission in Africa, *PLoS Neglected Tropical Diseases*, 9(11), 1–27. https://doi.org/10.1371/journal.pntd.0004222

Idro, R., Marsh, K., John, C. C., & Newton, C. R. (2010). Cerebral Malaria; Mechanisms Of

Brain Injury And Strategies For Improved Neuro-Cognitive Outcome. *Pediatric Research,* 68(4): 267–274.. 2010 October ; doi:10.1203/PDR.0b013e3181eee738.

Jothi, N., Rashid, N. A., & Husain, W. (2015). Data Mining in Healthcare - A Review. *Procedia Computer Science*, *72*, 306–313. https://doi.org/10.1016/j.procs.2015.12.145

Krejcie, R. V, & Morgan, D. (1970). ACTIVITIES, *Educational And Psychological Measurement*, 607–610.

Laishram, D. D., Sutton, P. L., Nanda, N., Sharma, V. L., Sobti, R. C., & Carlton, J. M. (2012). The complexities of malaria disease manifestations with a focus on asymptomatic malaria, *Malaria Journal*, vol. 11(1) pg. 29.

Langarizadeh, M. & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction : a Systematic Review, *Acta Informatica Medica*, *24*(5), 364–369. https://doi.org/10.5455/aim.2016.24.364-369

Lucini, F. R., Fogliatto, F. S., da Silveira, G. J. C., Neyeloff, J., Anzanello, M. J., de S. Kuchenbecker, R., & Schaan, B. D. (2017). Text mining approach to predict hospital admissions using early medical records from the emergency department. *International Journal of Medical Informatics*, *100*, 1–8. https://doi.org/10.1016/j.ijmedinf.2017.01.001

Manjusha, K K, Sankaranarayanan K, & Seena P. (2015). Data Mining in Dermatological Diagnosis : A Method for Severity Prediction, *International Journal of Computer Applications 117*(11), 11–14.

Masethe, H. D., & Masethe, M. A. (2014). Prediction of Heart Disease using Classification Algorithms, *Proceedings of the World Congress on Engineering and Computer Science 2014*, *II*, 22–24.

Medhekar, D. S., Bote, M. P., & Deshmukh, S. D. (2013). Heart Disease Prediction System using Naive Bayes, *International Journal Of Enhanced Research In Science Technology & Engineering*, *2*(3), 290-294.

Mohapatra, B. N., Jangid, S. K., & Mohanty, R. (2014). GCRBS score: a new scoring system for predicting outcome in severe falciparum malaria. *The Journal of the Association of Physicians of India*, *62*(1), 14–17.

Muhe, L., Oljira, B., Degefu, H., Enquesellassie, F., & Weber, M. W. (1999). Clinical

algorithm for malaria during low and high transmission seasons. *Archives of Disease in Childhood*, *81*(3), 216–220. https://doi.org/10.1136/adc.81.3.216

Mutanda, A. L., Cheruiyot, P., Hodges, J. S., Ayodo, G., Odero, W., & John, C. C. (2014). Sensitivity of fever for diagnosis of clinical malaria in a Kenyan area of unstable, low malaria transmission. *Malaria Journal*, *13*(1), 163. https://doi.org/10.1186/1475-2875-13-163

Ndyomugyenyi, R., Magnussen, P., & Clarke, S. (2007). Diagnosis and treatment of malaria in peripheral health facilities in Uganda: findings from an area of low transmission in south-western Uganda, Malaria journal, Vol. 6(1) *pg. 39.* https://doi.org/10.1186/1475-2875-6-39

Neeta, R., & Abraham, J. (2012). Different Clinical Features of Malaria. *Asian Journel of Biomedical & Pharmaceutical Sciences*, *2*(12), 28–3.

Nikam, S. S. (2015).A Comparative S tudy of Classification Techniques in Data Mining Algorithms. *Oriental journal of computer science and technology,* Vol. 8(1) pg. 13-19.

Oguntimilehin, A. (2013). A Machine Learning Approach to Clinical Diagnosis of Typhoid Fever. *International Journal of Computer and Information Technology 2*(4), 671–676.

Oguntimilehin, A, Adetunmbi, A. O & Abiola, O.B. (2015). A Review of Predictive Models on Diagnosis and Treatment of Malaria Fever. *International Journal of Computer Science and Mobile Computing 4*(5), 1087–1093.

Patil, R. R. (2014). Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing, *International Journal of Advanced Research in Computer and Communication Engineering* Vol. *3*(5).

Patil, R., Chopade, P., Mishra, A., Sane, B., & Sargar, Y. (2016). Disease Prediction System using Data Mining Hybrid Approach. *Communications on Applied Electronics (CAE) Foundation of Computer Science FCS, New York, USA* Volume 4 – No.9, April 2016 – www.caeaccess.org.

Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. *International Journal Of Computer Science And Applications* Vol. *6*(2), 256-261.

Pirnstill, C. W., & Coté, G. L. (2015). Malaria Diagnosis Using a Mobile Phone Polarized

Microscope. *Scientific Reports*, *5*, 13368. https://doi.org/10.1038/srep13368

Raj, T. F. M., & Prasanna, S. (2013). Implementation of ML Using Naïve Bayes Algorithm for Identifying Disease-Treatment Relation in Bio-Science Text. *Research Journal of Applied Sciences, Engineering and Technology, 5*(2), 421–426.

Rani, B. K., & Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering*, *2*(2), 250–255.

Razzak, M. I. (2015). Automatic detection and classification of malarial parasite. *International Journal of Biometrics and Bioinformatics (IJBB)*, *9*(1), 1–12.

S.Indhumathi, .G.Vijaybaskar. (2015). Web Based Health Care Detection Using Naive Bayes Algorithm, *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 4*(9), 3532-3536.

Saa, A. A. (2016). Educational Data Mining & Students ' Performance Prediction, *International Journal of Advanced Computer Science and Applications*, *7*(5), 212–220.

Sah, R. D., & Sheetalani, J. (2017). Review of Medical Disease Symptoms Prediction Using Data Mining Technique, *IOSR Journal of Computer Engineering (IOSR-JCE). 19*(3), 59–70. https://doi.org/10.9790/0661-1903015970

Sala al-Din Abdullah, A. (2016). Using Data Mining Techniques to identify the causes of deaths in al-gedaref hospital. *European Journal of Computer Science and Information Technology*, *4*(2), 1–8.

Sharma, V., Kumar, A., Panat, L., & Karajkhede, G. (2015). Malaria Outbreak Prediction Model Using Machine Learning, *International Journal of Advanced Research in Computer Engineering & Technology, 4*(12), 4415–4419.

Shinde, R., Arjun, S., Patil, P., & Waghmare, P. J. (2015). An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm. *International Journal of Computer Science and Information Technologies, 6*(1), 637–639.

Soni, J. (2011). Predictive Data Mining for Medical Diagnosis : An Overview of Heart Disease Prediction. *International Journal of Computer Applications, 17*(8), 43–48.

Stauffer, W., & Fischer, P. R. (2003). Diagnosis and Treatment of Malaria in Children.

Clinical Infectious Diseases, Vol. 37(10), 1340-1348.

Taneja, A. (2013). Heart Disease Prediction System Using Data Mining Techniques. *Oriental Journal Of Computer Science & Technology,* Vol. 6(4), 457-466.

Tarekegn, G. B. & Sreenivasarao, V. (2016). Application of Data Mining Techniques to Predict Students Placement in to Departments, *International Journal of Research Studies in Computer Science and Engineering 3*(2), 10–14.

Tribhuvan, A. P., Tribhuvan, P. P., & Gade, J. G. (2015). Advances in Computational Research Applying Naive Bayesian Classifier For Predicting Performance Of A Student Using Weka. *Advances in Computational Research, 7*(1), 239–242.

Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. *International Journal of Innovative Science, Engineering & Technology*, *2*(9), 441–444.

Vijayarani, S., & Deepa, S. (2014). Naïve Bayes Classification for Predicting Diseases in Haemoglobin Protein Sequences, *International Journal of Computational Intelligence and Informatics,3*(4), 278–283.

Vijayarani, S., & Dhayanand, S. (2015). Liver Disease Prediction using SVM and Naïve Bayes Algorithms. *International Journal of Science, Engineering and Technology Research (IJSETR), 4*(4), 816–820.

Vitorino, R., De, A., Mendonça, D., & Goreti, M. (2011). Severe Plasmodium falciparum malaria, *Rev Bras Ter Intensiva*, *23*(3), 358–369.

Wasan, H. K. and S. K. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. *Journal of Computer Science*, *2*(2), 194–200.

Zewdu, T. (1998). Prediction of HIV Status in Addis Ababa using Data Mining Technology. *HiLCoE Journal of Computer Science and Technology, Vol.2(2) pg. 51-71*