T.C.

ISTANBUL ALTINBAS UNIVERSITY

GRADUATE SCHOOL OF SCIENCES ENGINEERING

DIAGNOSES OF CORONARY HEART DISEASE (CHD) USING DATA MINING TECHNIQUES BASED ON CLASSIFICATION

Mustafa Adil Fayez

Master of Information Technology

Thesis Supervisor

Asst. Prof. Dr. Oğuz Ata

Istanbul (2018)

DIAGNOSES OF CORONARY HEART DISEASE (CHD) USING DATA MINING TECHNIQUES BASED ON CLASSIFICATION

by

Mustafa Adil Fayez Fayez

Information Technology

Submitted to the Graduate School of Science and Engineering

in partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAŞ UNIVERSITY

2018

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Oğuz Ata

Asst. Prof. Dr. Oğuz ATA

Supervisor

Examining Committee Members (first name belongs to the chairperson of the jury and the second name belongs to supervisor)

Assoc. Prof. Dr.Metin ZONTUL	Software Engineering, Istanbul Aydin University	_______
Asst. Prof. Dr.Oğuz ATA	Software Engineering, Istanbul Altinbas University	Oğuz Ata
Assoc. Prof. Dr.Oğuz BAYAT	Electrical Engineering, Istanbul Altinbas University	_______

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Oğuz Ata

Asst. Prof. Dr.Oğuz ATA

Head of Department

Assoc. Prof. Dr.Oğuz BAYAT

Director

Approval Date of Graduate School of

Science and Engineering: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Mustafa Adil Fayez Fayez

Signature

ACKNOWLEDGEMENTS

My thanks and appreciation to my supervisor Asst. Prof. Dr. Oğuz Ata to his interest and gave his attention, support, patience and guidance to me throughout the period of this study as well as my thanks to my family, who was supported for my arrival to this stage and also my friends who have been helping me throughout these years.

Table of Contents

LIST OF ABBREVIATIONS

Cases

ANN: Artificial Neural Network	4
CART: Classification And Regression Trees	4
CHD: Coronary Heart Disease	4
CSV: Comma Separated Value	4
CVD: Cardiovascular Disease	4
DT: Decission Tree	4
ECG: Electrocardiography	4
ECHO: Echocardiography	4
KNN: K-Nearest Neighbors	4
NB: Naive Bayes	4
python: Programing Language	4
RF: Random Forest	4
SVM: Support Vector Machine	4
UCI: California University Irvine	4
WEKA:Waikato Environment Of Knowledg Learning	4

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DIAGNOSES OF CORONARY HEART DISEASE (CHD) USING DATA MINING TECHNIQUES BASED ON CLASSIFICATION

FAYEZ, Mustafa Adil Fayez,
M.S., Information Technology, Altınbaş University,
Supervisor: Asst. Prof. Dr. Oğuz Ata

Date: May 2018
Pages: 67

Coronary heart disease (CHD) has attracted the most attention around the world because it leads to death. These days, data mining in many fields, including commercial fields and medical fields, where medical fields are the most productive of large data on a continuous basis, and which must find different ways to extract information, may be important in predicting the spread of this disease. We have designed a system to help the diagnosis of CHD with better reduction of costs and time required for the process by using a programing language with data mining classification techniques. These algorithms produced good results and high accuracy. We applied our study to various CHD datasets. We obtained the best accuracy at 99% through the use of the Random Forest (RF) algorithm with Hungarian two classes. With Cleveland, we obtained 94% accuracy using the same algorithm while the better accuracy with the same dataset in the previous study was 58% when using the SVM algorithm. Moreover, with the Hungarian five class dataset, we obtained 99% as the best accuracy using random Forest Classifier algorithm rather than the accuracy that was achieved with this dataset in previous work, which was close to 67% using the SVM algorithm. In addition, we obtained 88% as a better accuracy using the AdaBoost classifier with the Hungarian data set and 87% accuracy using the Logistic Regression classifier with the heart.csv dataset. With the Switzerland dataset, we had 95% as the best accuracy using Random Forest and 91% best accuracy with the Long-Beach dataset using the same classifier. Finally, with the Switzerland dataset, we achieved a 78% better accuracy using the AdaBoost and Logistic Regression classifier. With Long-Beach, we had 80% using the AdaBoost classifier and 76%

using the Logistic Regression classifier. Also with the heart.csv dataset, we achieved 87% best accuracy using the Logistic Regression classifier and 86% accuracy when using the AdaBoost classifier. We used a train test split and preprocessing for the CHD dataset in this study and processed the missing values that were found with attributes with a less complicated system. This process differs significantly from previous study is proposed results and accuracy for this purpose with the same CHD dataset.

Keywords: CHD, Classification techniques, Python, Data mining.

ÖZET

SINIFLANDIRMA TEMELLİ VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK KORONER KALP HASTALIĞI (KKH) TANISI

Mustafa Adil Fayez, FAYEZ
M.S., Bilgi Teknolojisi, Altınbaş Üniversitesi
Danışman: Yrd. Dr. Oğuz Ata

Tarih: Mayıs / 2018
Sayfalar: 67

Koroner kalp hastalığı (KKH) ölümle sonuçlanan bir hastalık olduğu için bütün dünyada çok fazla ilgiyi çekmiştir. Bu günlerde, veri madenciliği özellikle ticari ve medikal alanda olmak üzere birçok alanda kullanılmaktadır. Özellikle medikal alan, veri üretiminin sürekli olması ve farklı öznitelik çıkarımı yöntemlerinin bulunmasından dolayı hastalığın yayılmasına dair çözümler önermektedir. Veri madenciliği sınıflandırma teknikleri ve bir programlama dili kullanarak süreç için gereken maliyet ve zamanın daha iyi azaltılması için KKH teşhisine yardımcı olacak bir sistem tasarladık. Bu algoritmalar iyi sonuçlar ve yüksek doğruluk elde etmiştir. Çalışmamızı çeşitli KKH veri kümelerine uyguladık. Hungarian iki sınıflı verisetinde Rastgele Orman(Random Forest - RF) algoritması kullanılarak en iyi doğruluğu % 99 oranında elde ettik. Cleveland veri seti ile, aynı algoritmayı kullanarak % 94 oranında doğruluk elde ettik, kıyasladığımız bir başka çalışmadaki sonuçta aynı veri kümesinde elde ettikleri doğruluk oranı SVM algoritması ile % 58 idi. Ayrıca, Hungarian beş sınıflı veri kümesi ile kıyasladığımız önceki çalışmada SVM algoritması kullanılarak % 67 doğruluk oranı elde edilmişken biz Rastgele Orman(RF) algoritması ile %99 doğruluk oranı elde ettik. Buna ek olarak, AdaBoast algoritması ile Hungarian veri setinde %88 ve heart.csv veri setinde Logistic Regression algoritması ile %87 doğruluk oranı elde ettik. Ayrıca Switzerland veri seti ile Rastgele Orman(RF) algoritması kullanarak %95 ve Long-Beach veri seti ile aynı algoritmadan %91 doğruluk oranı elde ettik. Son olarak, Switzerland veri seti ile AdaBoost ve Logistic Regression algoritmaları ile %78, Long-Beach veri setinde AdaBoost algoritması ile %80, Logistic Regression algoritması ile %76, heart.csv veri setinde Logistic Regression ile %87 ve AdaBoost algoritması ile %86 doğruluk oranı elde ettik.

Bu çalışmada KKH için farklı veri setleri için ortak önişlem ve eğitim-test veri bölmesi kullandık. Bu işlem önceki çalışmadan önemli ölçüde farklıdır ve aynı KKH veri setleri ile elde edilen sonuçlardan daha başarılı sonuçlar almamıza katkıda bulunmuştur.

Anahtar Kelimeler: KKH, Sınıflandırma teknikleri, Python, Veri madenciliği.

CHAPTER 1

1.1 Introduction

Coronary heart disease (CHD), also known as coronary artery disease, is a state that refers to how platelets accumulate inside the arteries close to major valves. These arteries provide the heart with sufficient oxygen and blood. The platelets consist cholesterol, calcium with other materials. Arteriosclerosis occurs when platelets accumulate in the arteries of the heart [1-5]. The main causes of this disease include high blood pressure, fat and cholesterol, smoking and high blood sugar. Around the world, especially in the U.S, CHD is a major cause of death. Every year, more than a billion Americans die due to CHD specific circumstances or practices identified as danger factors, including sleep apnea (a common complaint such that breathing stops during sleep), stress, and alcohol. Excessive alcohol consumption may destroy heart strength and increase other danger factors for cardiovascular diseases and preeclampsia[6].

To analyze CHD, doctor will based on the medicinal with household is health past, the danger factors, heart illness, bodily examinations with the outcomes of trials and processes. There are no individual trials that can identify coronary heart disease when doctors consider the possibility of CHD. From the tests that should be performed, EKG is an effortless, uncomplicated check that finds and registers electrical activity of heart. The trial displays the speed and rhythms of the heartbeat (fixed or unbalanced). EKG also uses the power and timing of electrical signals as they permit over each portion of the heart. Stress testing and echocardiography use sound waves to show moving images of the heart. Such a check can provide information on the size and form of the heart and describe in how efficiently the heart cavities and valves function. ECHO can also display the zones of blood movement to the heart, and by taking X-ray photographs of the chest and the organs inside the chest, the levels of cholesterol, sugar and proteins in the blood and the (ECT) can be determined. The ECT is a check that looks for patches of calcium (i.e., calcification) on the walls of the coronary arteries[7].

The diagnosis process can clearly recognize disease from its marks and symptoms. A few prognosis trials are available to support physicians who depend on observations and examinations of medical histories. In the twentieth century, there were many technological developments in the field of medicine which led to the advancement of a wide range of diagnostic checks and novel methods of diseases diagnoses. These developments significantly improved the capability of physicians to conduct comprehensive diagnoses[8].

In health care, information and data remain to upsurge and intension by providing large amounts of complex information. There are many applications in the medical field to help predict and diagnose different kinds of diseases. Data mining techniques are one of these procedures which continue advancing better methods to provide the best medical solutions. The shift from written health records to electronically stored records has played a great role in advocating for the use of clinical data of patients to improve health care. The adoption of electronic health records allows health professionals to disseminate knowledge in all health care sectors, which in turn helps to reduce medical errors, provide comprehensive documentation, and improve patient care and satisfaction.

Data mining or extraction is also expected to help to reduce costs. If the US healthcare industry continues to use massive data to increase efficiency and quality, the potential value may reach more than $300 billion per year according to the 2011 McKinsey World Report [9]. The future of health care may depend on the use of data extraction to reduce health care costs, identify treatment plans, implement best practices with measure efficacy, detect fraudulent insurance with medical claims and ultimately improve patient care. From our experience so far, many hospitals have not understood what health care data mining means. In its simplest form, data mining can be defined as the use of methods to extract and determine patterns from enormous quantities of data. These outcomes include the use of databanks with information along with computer analysis with previous studies and a collection of discussions. DM has gained momentum in the healthcare industry because it provides benefits for all stakeholders, including caregivers, patients, health care organizations, researchers, and insurance companies. Care providers can use data mining to recognize effective remedies and best practices [9]. Patients with high-risk diseases can have access to better and affordable health care services while using appropriate treatment interventions and protocols. Healthcare organizations can use data mining to improve patient relief by providing more patient care and by reducing costs with increases in operational efficiency while maintaining high-quality care.

DM can be used not only to identify exact outcomes but also to identify and anticipate any events nearby in order to increase proactive prevention of events at the outset [10, 11]. In our research, we concentrate on the classification of diseases with a data mining classification algorithm because the diagnosis of a disease takes a long time and has high costs in medical procedures. Therefore, we are attempting a different approach to classification in order to lessen time and reduce prices to help in the diagnosis of CHD in the simplest and most active manner.

1.2 Definitions

1.2.1 Coronary heart disease: is a tapering of minor blood vessels that supply the heart with blood and oxygen. Coronary heart disease is also called coronary artery disease [12].

1.2.2 Blood pressure: High blood pressure (HBP) is when greater-than-normal pressure of blood against blood vessel walls [13].

1.2.3 Cholesterol: is a waxy material found naturally in the blood, mostly formed in the liver but also originating in foods such as red meat, butter and eggs [14].

1.2.4 Data mining: is the process of isolation of enormous datasets to distinguish patterns and to find relationships by extracting beneficial information to solve problems through data analysis [15].

1.3 Statement of problem

For many years, heart disease has been considered the most common reason for death worldwide. Health experts use their knowledge and skills to complete diagnoses of CHD for patients and generally, most medical data garnered from patients are kept in files or folders. These vast quantities of unorganized health registers make no sense to users. Data mining is a classification technique we can use to resolve this problem by recalling preceding cases and recycling information and knowledge. In such cases, these techniques turn data into beneficial information which can help to support the system for CHD diagnosis. This system is intended to provide the best contribution to the physician and health experts in the diagnosis of patient data. Therefore, the system can assist physicians and health experts in identifying and analyzing patients’ health status. This research aims primarily at developing a system of taking a clinical dataset to determine the condition of patients, which will save time and money and add value to the quality decisions of physicians, which in turn prefer patient to healing and can be applied where there is a shortage of professionals. This study attempts to help determine which attributes are more important for the diagnosis of CHD and which classification algorithm is most appropriate to establish a diagnostic system for CHD.

1.4 Aim and hypothesis

The main objective of our study is to know how to deal with coronary heart disease by using the best method to reduce the amount of time that doctors lose to make diagnosis process and at lowest costs by applying data mining classification techniques. The system comprises multiple data mining techniques in its classification processes to find the best way to deal with any loss of data in medical datasets.

The hypothesis of this study is one such that the inaccuracy of data obtained from medical sources may be a major cause of errors during the diagnosis of coronary heart disease. Misinformation provided by the patient to the physician and the lack and inaccuracy of information are some of the causes of the aggravation of the disease.

Coronary heart disease cannot be reduced or eliminated completely through a specific method without reducing the causes of the disease. Moreover, many commentators attribute 40% of the latest decline in coronary heart disease may be due to improved treatment rather than a decrease in the occurrence of disease.

CHAPTER 2

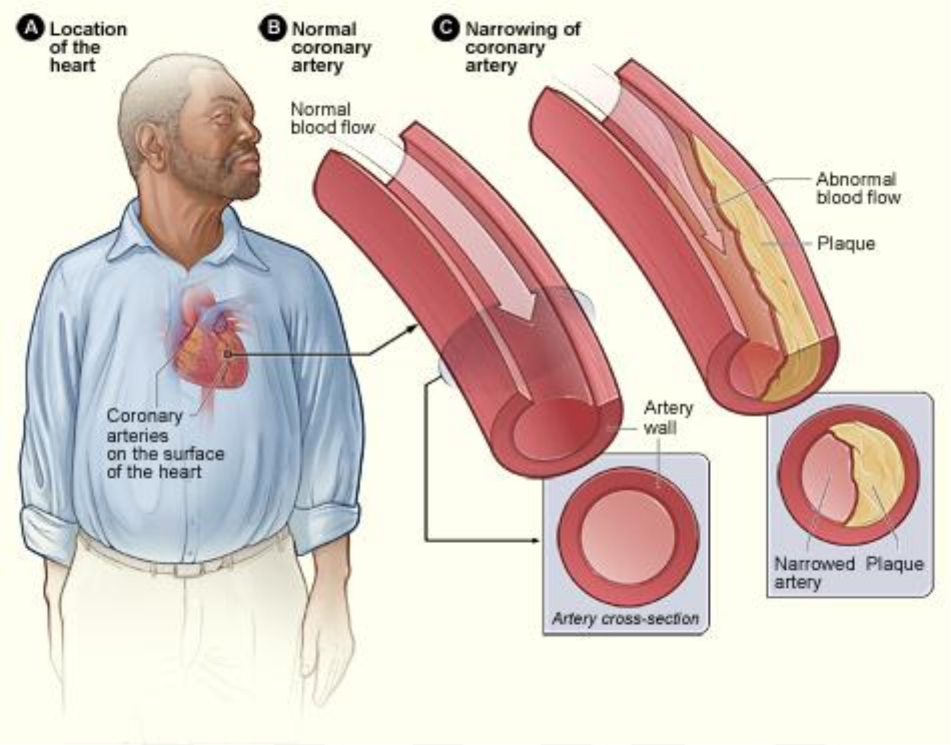
2.1 Introduction

Through the study of a number of scientific, technical and medical sources, coronary heart diseases (CHDs) have the highest rate, among non-communicable diseases, that leads to death in most countries around the world. Therefore, it is necessary to produce a diagnosis model or system for CHD to reduce management costs. Thus, prevention is important. Nowadays, many advanced studies have proposed methods and systems that predict and diagnose CHD using data mining algorithms with artificial intelligence and automated learning techniques. CHD based data extraction models use different DM techniques, such ANN, decision trees, Bayesian theory and genetic algorithms.

2.2 Coronary Heart Disease

Coronary Heart Disease occurs when the major blood vessels that feed the heart with blood and oxygen becomes damaged or diseased [16]. There are three major forms of heart disease, namely Cardiovascular Disease (CVD), Coronary Heart Disease (CHD) and ischemic heart diseases. CHD and CVD occur in the majority of cases and cause death in multiple countries, especially in the United States and China for both males and females at different ages. Deposits of plaque in the heart arteries are usually the reason for CHDs. The figure below shows the form of CHD and its effect on the arteries of the heart.

Figure 2. 1: Coronary heart disease[17].



When sediments accumulate, they narrow the coronary arteries and decrease blood delivered to the heart [18]. This decrease may cause chest pain, shortness of breath and other signs and symptoms of CHD. Full blockages can cause a heart attack and the main causes of this disease are smoking, high blood pressure, high cholesterol and a sedentary life style.

2.3 Diagnosis problem

An important factor in preventing CHD is an efficient health check, which includes a study of the patient’s personal and family medical history as well as measurements of blood pressure, cholesterol and physical fitness [19]. This may be time consuming and costly. Data mining consumed long studied in healthcare field by the application of DM tools and equipment to improve development of data analysis in large and complex medicinal datasets. An approved or applied DM algorithm in the medication sides is of great importance in diagnosis, prediction and deeply understood of health care dataset. The goals of these submissions contain action center analyses to improve and avoid any mistakes in hospitals with the early diagnosis of disease and reductions of hospital deaths. With data mining algorithms, we can decrease time and minimize the cost, thereby making the diagnosis problem easier through the design of a model to find the best solution at high accuracy to diagnose CHD disease.

2.4 Background of CHD with data mining

The diagnosis of, and research into, CHD is of particular interest and concern to the community because of the great human loss among people, especially in European countries. The major challenge of health care organizations is the provision of correct patient diagnosis and the effective administration of treatments. Poor clinical decisions lead to serious and unacceptable results. A clinical decision-making process in the health care setting needs to be supported with more advanced technology, including a computer based information system [20]. DM science includes a group of physical and classification techniques which can be utilized to discover hidden information and knowledge from a medical dataset. This information provides healthcare professionals with an extra source of knowledge to make intelligent clinical decisions, which in turn enhances patient safety, increases patients' quality of life and improves the outcomes of patient diagnoses. There are many systems created by different researchers to support experts and doctors to detect heart disease. Specific techniques are more common, such as NB, DT and KNN. There are other data mining based classification algorithms such as kernel density, neural networks, bagging algorithms, sequential minimal optimization, direct kernel self-organizing maps and SVM, which are applied to diagnosing CHD and used in data mining tools. Data mining classification algorithms are used to help in the diagnosis processes.

2.5 Related work

With Dr. A.Govrdhan, K.Srinivas and B.Kavihta Rani[21] their research about design a system to help predict heart attacks used one dependency augmented naïve bayes classifier (ODANB) and Bayesian network (BN) on medical dataset obtained from UCI Machine Learning Repository for dataset. They got 84% as best accuracy with (BN) technique applied in WEKA tool.

Also Ching-Hsue Cheng, Yen-Wen Chen and Duen-Yian Yeh[22] they had created a prediction model for CHD disease. Different kind of data mining algorithm approved in this study like DT, bayesian classifier and back propagation neural network.

Also with A. Sheik Abdullah and R.R.Rajalaxmi [23] their research or model to predict CHD disease. In this study Random forest (RF) classification algorithm which used on Cleveland CHD dataset taken from UCI. They had gained best accuracy 63.33% used RF technique.

B.L. Deekshatulu, M. A. Jabbar and P. Chandra [24]. They have worked on classification of heart disease used KNN and genetic algorithm. They have proposed new algorithm which combine KNN machine learning technique with genetic algorithm for classification process. In their proposed they have got a total accuracy as 92% with multiple heart disease dataset. So that as we said there are many works and models approached to diagnoses of (CHD) with different accuracy.

T. Santhanam and E.P. Ephzibah [25] they have used PCA and feed forward neural network (NT) algorithms to classify CHD heart disease. They had used attribute selection with classification techniques on their dataset which got from UCI repository to get the better model for that purpose. Also they achieved best accuracy with principal components analysis (PCA) techniques.

Pramod Kumar Yadav, Shamsher Bahadur Patel and Dr. D. P. Shukla [26] created a model to predict and diagnosis of Heart Disease used Naïve Bayes (NB) and decision tree algorithm by utilized WEKA tool. Also they had used Genetic algorithm to determine which attribute contribute more towards diagnosis of heart ailments. They have got best accuracy with decision tree (D Tree) technique for heart disease prediction system.

Zahra Mahmoodabadi and Mohammad Saniee Abadeh [27] have designed a model to diagnosis of CHD they used Cleveland and Hungarian clinical datasets. Their proposed system created by using MATLAB with ICA and PSO algorithms. Also they got 94.92% as a better accuracy in their research used ICA algorithm.

B.Venkatalakshmi, M.V Shivsankar [28] have created a predictive data mining model to diagnosis of heart disease. They used medical dataset got from UCI with apply Decision Tree (DT) and Naïve Bayes (DM) methods used WEKA data mining tool. Also they have got a best accuracy with NB classifier as 85% in their model.

Feature analysis for CHD designed by Randa El-Bialy , Mostafa A. Salamay, Omar H. Karam and M.Essam Khalifa [29] They suggested the results analyzed by the WEKA tool which used in their model. Their CHD dataset taken from UCI and they have achieved accuracy of this collected dataset like (77.5%, 78%) which is greater than the average of classification accuracy for all their datasets.

B. Bahrami and M. Hosseini Shirvani [30]. They proposed a system by used several kind of classification algorithm like KNN and J48 on clinical dataset which involved 209 records with no missing values. In their model they had achieved 83.7% as better accuracy with J48 classifier applied in WEKA tool.

Wiharto, MCom, Hari Cusnanto and Herianto [31]. Their proposed were about interpretation of medical data to create a system to interpretation clinical examination results for identification of CHD based on C4.5 and decision tree techniques. They got sensitivity of 74.7%, with specificity of 93.7%, a PPV of 74.2%, and NPV of 93.7%, with accuracy AUC of 84.2%.

With M.Swathi Lakshmi and Dr. D.Haritha [32] they had proposed a model to diagnosis of heart disease. They applied SVM and naïve bayes (NB) data mining (DM) algorithms for forecast of heart disease with Cleveland dataset taken from UCI. In their system they have achieved 84.87% as a better accuracy used NB classification algorithm.

Mohammad Tajfard, Maryam Tayefia, Ali Reza Amirabadizadeh, H. Esmaeily, A. Taghipour, G. A. Ferns, Majid Ghayour-Mobarhan worked a researched about CRP is efficiently associated with CHD disease[33]. They had used decision tree algorithm on dataset which contain 2346 individuals with 1159 healthy participants. In their model that created to identify the Risk Factors of CHD disease with sensitivity specificity with good accuracy of 96%.

In Sneha Susan Varghese and Laya Devadas [34] research they proposed a model for CHD prediction system. Their system consist of two artificial intelligence methods rule based expert system and deep learning method. They used CLIPS which is a public Software material for building proficient systems.

An decision support model proposed with Thendral Puyalnithi and Madhuviswanatham Vankadara [35] in their research they got a dataset from UCI repository applied with different type of Data Mining (DM) algorithms used orange tool. They have obtained a high accuracy used ID3 Naïve Bayes (NB) algorithms about 93% as a better accuracy.

2.6 Data mining algorithms

2.6.1 Naïve Bayes

Bayes supervised theory was named by Reverend Thomas Bayes, who lived between1702 and 1760. It is best for use in classification processes or tasks and it is simple to use and apply when there are attributes and variables which are independent of each other [36]. In the medical field, there are many data mining models based on this technique that help in the prediction process and diagnose problems, especially heart disease in clinical datasets.

The Naïve Bayes is basically a group of algorithms sharing a similar set of properties or ideologies. It is the most prevalent or popular method which uses a group of algorithms in machine learning. This technique permits us to test predictions or classifications by offering a group of features using probability theory.

The advantage of this technique is that it is fast and informal to build, it is trainable with small datasets, it is not sensitive to irrelevant attributes and there is an availability of online applications with simple emotion modeling.

The disadvantages include the fact that it assumes every feature to be independent and computation intensive. In many academic sources, it can be seen that the Naïve Bayes classifier performance almost matches other classifiers in most cases. The Naïve Bayes classifier technique runs better in most clinical data classification problems.

2.6.2 SVM algorithm

The SVM supervised technique industrialized by V. N. Vapnik and Alexy Y. Chervonenkis since 1963 with it is original form. It is used for in nonlinear classification process through apply the Kernel Trick to Maximum-Margin Hyper Planes which established by Bernhard E. Boser and Isabelle M. Guyon. And because it is very powerful in binary classification, it is used in many application fields. However, it is mostly used for classification problems with binary data type[37]. We can say about SVM that it is light and acceptable in use for healthcare dataset[38]. The function of the resolution executed by SVM can be written as:

$$f(x) = \text{sign } (\sum_{i=1}^{N} y_i a_i k(\bar{x}, \bar{x}_i) + b) \qquad (2.2)$$

There are two kernel functions are listed below:

Polynomial function k $(\bar{x}_i, \bar{x}_j) = (\bar{x}_i \bullet \bar{x}_j + 1)^{d}$ (2.3)

Radial basis function (RBF) k $(\bar{x}_i, \bar{x}_j) = \exp\,(\text{ã} \,|\bar{x}_i - \bar{x}_j |^{2})$ (2.4)

For the advantages of SVM that it is pliability in select of brink form robustness with minor number of dataset also we can do any think with right kernel. It is work well when the structure is unknown. In addition, SVM technique considers the best choice when the quantity of dimensions is upper than the available samples.

The disadvantages in SVM that it is terrible and slow during test phase also the serious problem with this algorithm that it is high complexity and need to large memory for quadratic program in addition SVM may computationally expensive.

2.6.3 Decision tree algorithm

A decision tree supervised is a data mining technique advanced by J. Ross at Sydney University. The head method of the decision tree is named iterative dichotomies ID3. This technique can be better used with linear decision boundary problem types and applied when high accuracy and time are important [39]. It uses a tree as a graph to provide decisions and it is simple depiction of instances classifier and this algorithm widely used for supervised (ML). The decision tree technique creates a type of flowchart involving a node or leaf with a set of decisions to be completed depending on this node. Each node of this algorithm is marked with a probability distribution over the class. The classification process or rule is represented by the path from the root to the leaf node. There are three basic algorithms which use DT techniques: ID3, C4.5 and CART.

The advantage of a decision tree (DT) is that the analysis process of this tool permits a team to clearly lay out and reflect all available options. Moreover, it is comparatively informal to visualize the costs quickly and accurately when compared with other classification DM methods. In addition, the C4.5 Decision Tree algorithm has the ability to provide high precision and the CART algorithm is used when we want to make quick decisions. The decision tree technique also has the ability to deal with numerical and factual data.

The disadvantages of DT occur in cases when there are several choices to reflect and each selection has several likely results. Moreover, this technique occasionally requires complex preparation as well as additional time and effort to determine the different possible outcomes for each decision. It is not the best when there are several uncorrelated variables in the dataset.

2.6.4 KNN algorithm

KNN is a statistical data mining classification technique that depends on the closest training example in an attribute space. Moreover, it is a lazy algorithm because it is used only to re-educate the Search Space with instances whose class is known; however, when the class is unknown, it is presented for evaluation. The KNN technique is very fast, but the testing stage is costly in terms of time and memory [40]. This algorithm involves two phases of training and classification in the training step. The training example is a vector and in the classification step, K is a constant. KNN compares the input attribute vector with the library of reference vectors and a query point with the nearest class as a label of the library feature vector.

The advantage of KKN is that it is a non-parametric method. The path or tactic is easy to learn and it is very powerful with noisy data [41].

For the disadvantage is that there is no possibility of recognizing a class value that does not exist in the training data. The selected neighbors are assigned the same weight.

2.6.5 Random forest algorithm

The RF technique is a supervised data mining technique for classification purposes. It was developed by Tin Km Ho in 1995 and it can be used for regression problems [42]. We often use the random forest classifier technique when we need high accuracy in a short time. This algorithm involves the DT group. The difference between Random Forest (RF) and Decision Tree (DT) is that the Random Forest technique randomly selects observations and features to build several decision trees followed by averaging the results. Another difference is that deep decision trees might suffer from over fitting. Random Forest prevents over fitting most of the time by creating random subsets of the attributes and building smaller trees using these subsets followed by combining the sub trees.

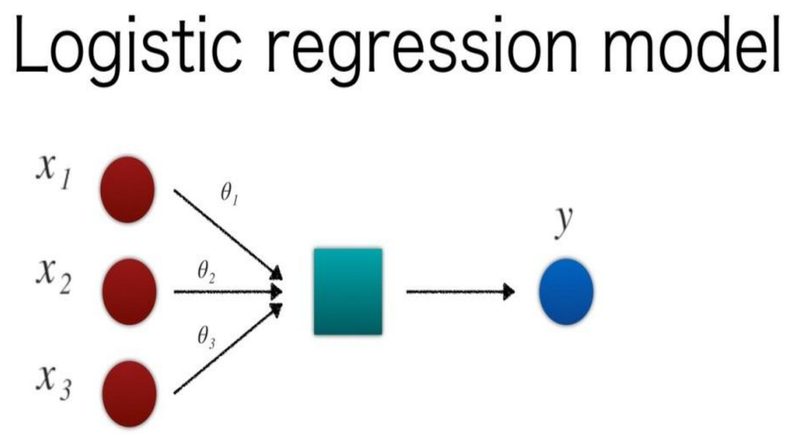
The advantages of random forest are that it is very good at maintaining accuracy, especially when there are missing values. Moreover, it is efficient at saving data preparation time while providing the benefit of implicit feature selection. It is generally very effective with large datasets and it has a higher classification accuracy. It can handle thousands of variables without a deletion variable [43].

Also there is a disadvantage for this algorithm that it has a difficult theoretical analysis and it makes no guesses outside the range of response standards in the training data.

2.6.6 Logistic regression

Logistic regression is a type of data mining and a predictive and classification model which can be used for the outcome variable is a class variable include of two categories for example yes or no, disease or no disease. It is endeavor to results are expected based on group of independent variables, but if the academics include wrong independent variables then the model will be little or unpredictable [44]. We can also think of this technique as a special case of linear regression when the result variable is categorical since we use the probability log as a dependent variable. Logistic regression was developed by David Cox in 1958. It can be can be utilized to guess the probability of a binary response on the basis of one or more predictions. A risk factor increases the probability of producing a certain result by a certain percentage. The figure below expresses the logistic regression model.

Figure 2. 2: Logistic regression model [45].

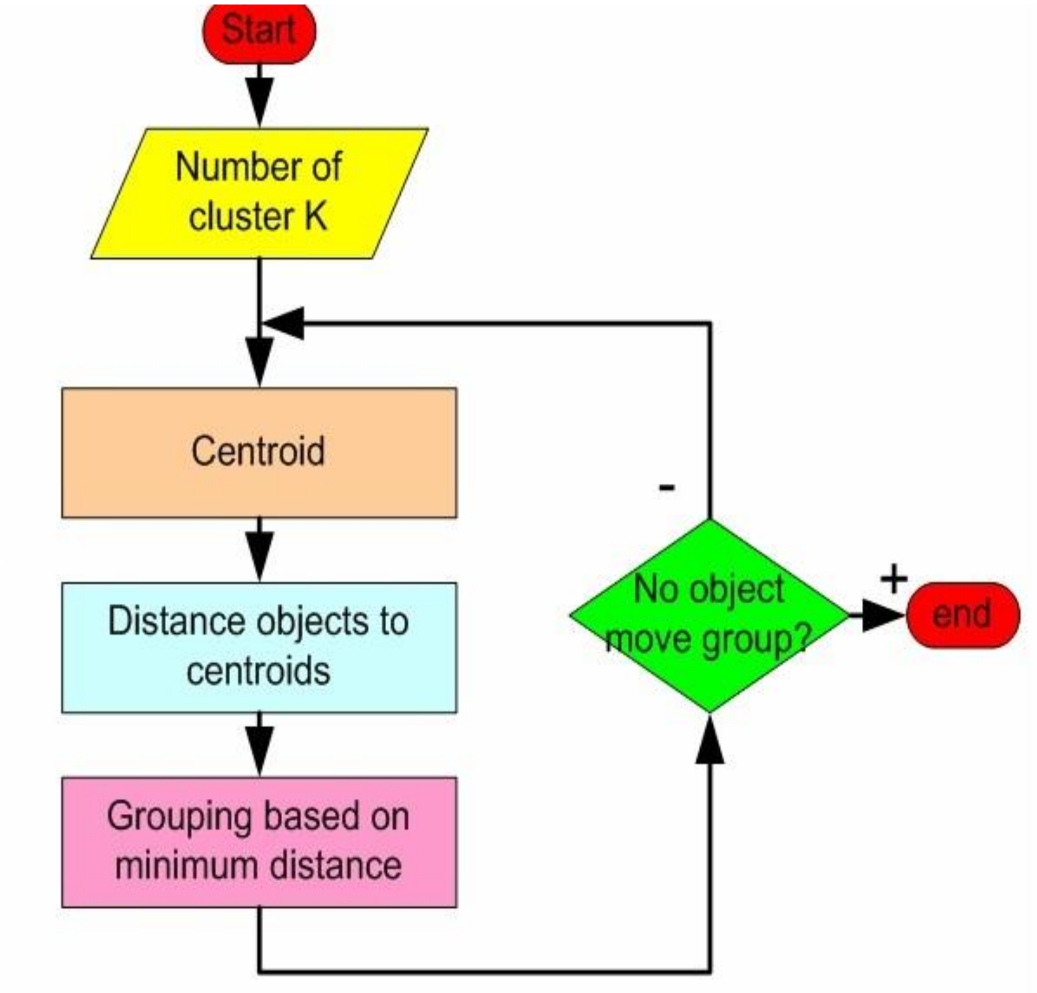


From the advantages of this technique that error terms don't have to be normally distributed and no linear relationship between the IV and DV that must be assumed also it can handle nonlinear effect and power terms[46]. The disadvantages of this technique that it is need to a large sample size to achieve stable results with limited outcome variable.

2.6.7 K-Means algorithm

K-clustering means is a type of unsupervised machine learning algorithm used for clustering purposes or unregistered data. The purpose of this algorithm is to find collections in the dataset with the number of collections being represented in the K variable. This algorithm works repeatedly to assign each data point to a K group according to the features provided. The data points are grouped based on similarities in features [46]. This clustering algorithm can be used to find groups that are not explicitly categorized in the dataset and which can be used to confirm business assumptions about existing group types or to identify unknown groups in complex datasets. The figure below expresses and describes the K-means cluster technique.

Figure 2. 3: K-means cluster algorithm [47].

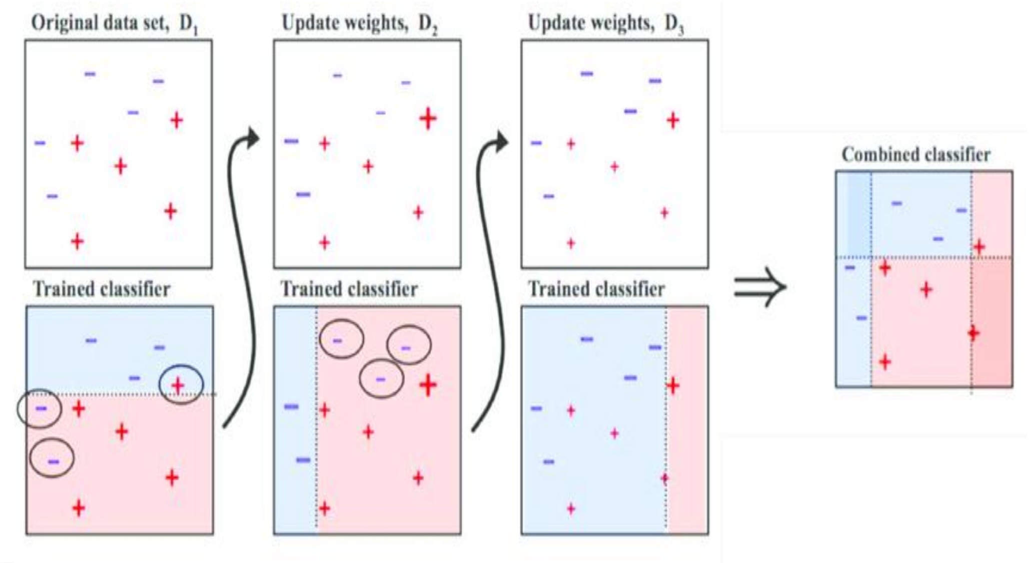

Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

From the advantages of k-means technique that it is faster than hierarchical clustering with the large amount of data and it is easy to implement also the instance can change cluster or move to another cluster. The disadvantages of this algorithm that it is difficult to predict the cluster is number and it has a strong impact on the final results also it is sensitive to scale.

2.6.8 Ada Boost algorithm

Ada Boost is an adaptive boosting classifier for parallel enhancement. It is basically an automatic machine learning algorithm used as a classifier. We can use this algorithm when there is a large amount of data that needs to be divided into different categories; therefore we need a good classification algorithm to do so [48]. Typically use of Ada Boost sees it used in conjunction with other machine learning algorithms to improve their performance. The figure below explains the steps of the Ada Boost classifier.

Figure 2. 4: Ada Boost classifier steps [49].

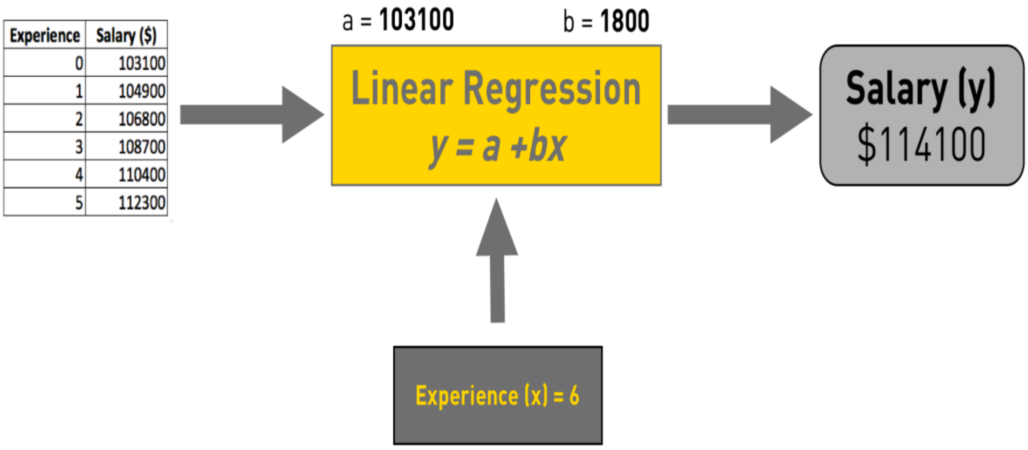


The word boosting in the promotion of other algorithms is a general way to improve the accuracy of any particular learning algorithm. From the advantages of this algorithm that it is simple or easy to perform and it is simple classifier for feature selection purpose also it have good generalization. The disadvantages of this technique that it is have suboptimal solution and it is sensitive to noisy data.

2.6.9 Linear regression algorithm

The regression task is used to predict values that have the ability to provide answers to questions such as the sum of points in a game. Regression techniques collapse under the supervision of learning. Regression analysis is used to evaluate the connection between two or more variables. It is a data mining regression technique which models or describes the relationship between two variables or features by installing a linear equation for observed data. One feature can be considered to be an explanatory variable, such as x, and the other feature may be considered to be a dependent or outcome variable (y) [50]. The figure below expresses and describes the general steps of the linear regression machine learning algorithm.

Figure 2.5: General steps of linear regression [51].



Experience	Salary ($)
0	103100
1	104900
2	106800
3	108700
4	110400
5	112300

From the advantages of linear regression technique that it is one of the most statistical tools and it doesn’t fit the data exactly also it considered as simple or easy to understand and perform[52]. From the disadvantages of this algorithm that it is limited to use for linear task to find the relationship between two independent variables and it is look only at the mean of that variable also it is sensitive to outliers.

CHAPTER 3

3.1 Methodology

The capability of machine learning (ML) to analyze enormous amounts of data will deliver doctors and patients the necessary opinions in real time. This will permit healthcare professionals to diagnose, and offer cures for, diseases. In our study, we used a data mining classification method to obtain or create a system that may be used to classify and diagnose CHD by applying multiple techniques with different types of medical datasets using an appropriate programing language to help people in the healthcare field.

3.2 Heart disease dataset and it is description

In our system, we used five different kinds of medical dataset acquired from the California University Irvine (UCI) repository. The datasets that we used were the Cleveland, Hungarian, Long-Beach and Switzerland datasets [53]. We also used the processed two class Hungarian dataset obtained from the same web site and the heart.csv data set from the Kaggle web site[54]. The description of our dataset showed in the tables below.

Table 3. 1: The description of datasets.

Name of Dataset	No. Features	No. instances	No. classes
Cleveland	13	303	5
Hungarian	13	294	5
Long-beach	13	200	5
Switzerland	13	123	5
Two classes hungarian	13	294	2
Heart.csv two class	13	304	2

Cleveland, Hungarian, long-beach and Switzerland datasets classes are used to infer the presence (values 1, 2, 3, and 4) or absence (value 0) CHD disease. The features are (1) age, (2) sex, (3) chest pain type, (4) resting blood pressure, (5) cholesterol, (6) fasting blood sugar, (7) resting electrocardiographic results, (8) maximum heart rate, (9) exercise induced angina, (10) depression induced by exercise relative to segment, (11) slope of peak exercise, (12) number of major vessels, and (13) thal.

3.3 Data preprocessing

Medical or clinical datasets that often have a missing value problem and this can affect the efficiency and accuracy of our system. We find missing values as question marks in our dataset and we solve this problem according to academic solutions [55]. The procedures to handle missing data include:

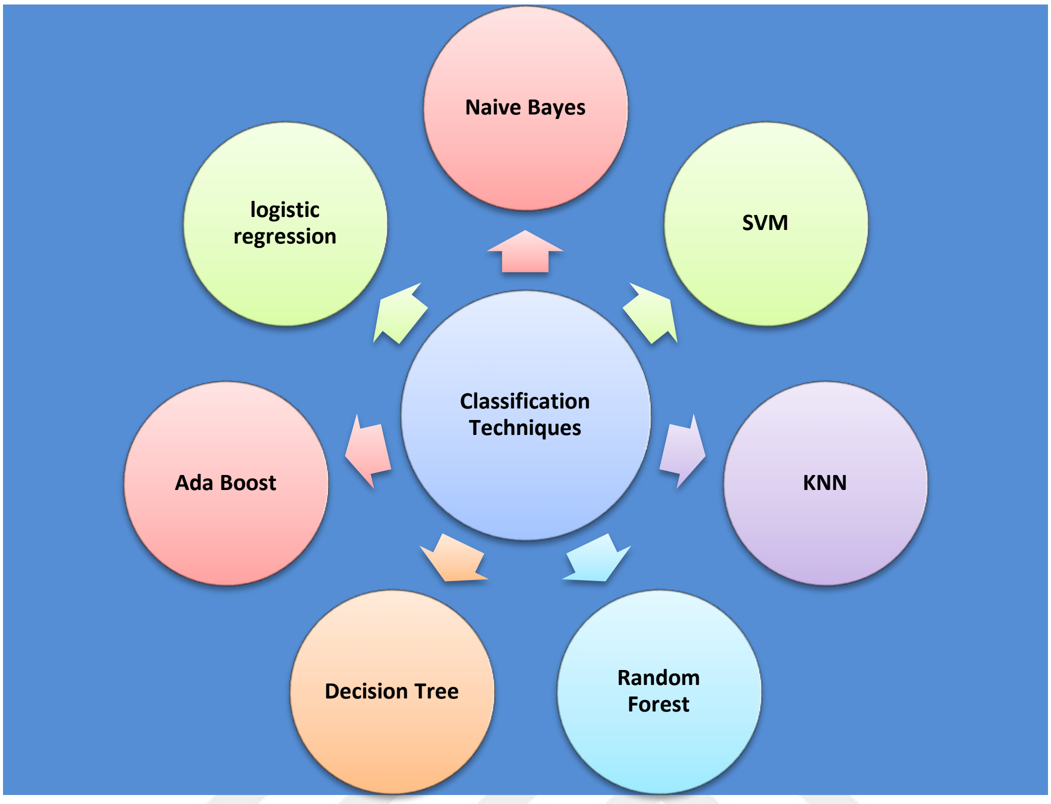
- Replace missing values with numerical data like question mark with 0 or the mean of the attribute.
- Delete the attribute which contain high number of missing values.
- Replace the numeric missing values with -9 or 999.

Missing feature values may cause problems such as errors of data measure and errors of data understanding. We process dataset attributes to identify and remove unneeded attributes with high missing values so as to be able to affect the accuracy of the data mining classification algorithm [56]. We processed our Hungarian five class and two class datasets as well as the Long-Beach and Switzerland datasets. Three attributes with large numbers of missing values were removed from each of the Hungarian datasets and two attributes were removed from the Long-Beach and Switzerland datasets in order to increase and enhance the accuracy [57]. Moreover, we processed the other missing values by replacing those values with 0 (zero) according to data preprocessing procedures.

3.4 Classification techniques

Data mining classification methods have become simple and popular in machine learning, most notably in healthcare. The large amount of production of clinical datasets has become a problem, so it has become necessary to find new solutions, such as classification or prediction systems using classification algorithms [58]. The figure below presents a number of classification algorithms that are available.

Figure 3. 1: The different kinds of classification techniques.

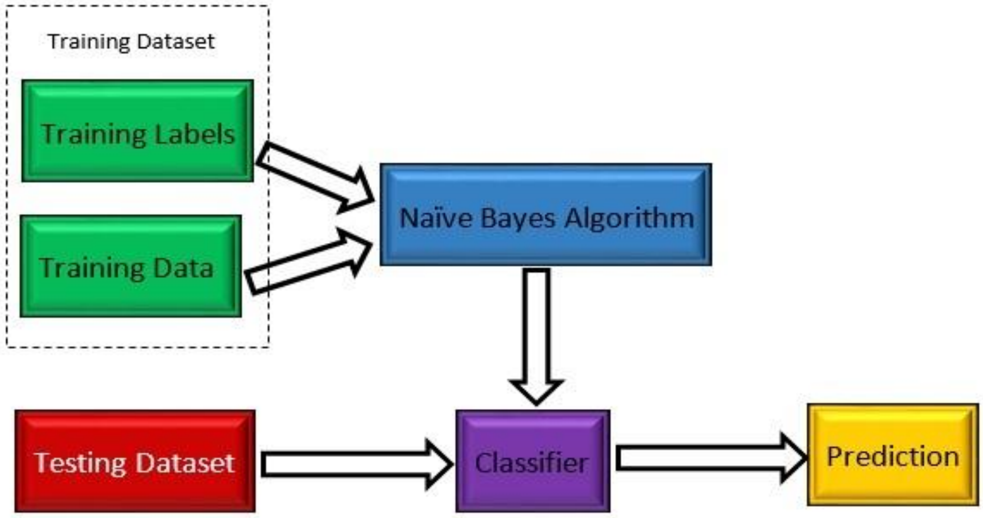


The two types of learning include supervised and unsupervised with supervised. The system attempts to learn from any previous examples given. Learning under supervision is where we have both input values (X) and output values (Y) and we can use this technique to implement the assignment function from the input to the output [59]. With unsupervised learning, the techniques are left to themselves to discover interesting structures in the data, including input data (X) only without any corresponding output variables. The advantages of classification include the fact that it is simple and the structure exists independent of any job. The disadvantages include judgments being subjective and the benchmark being utilized for comparison may have inherent biases that may affect assigned groups of employees and a number of functions may appear to fit more than one job.

3.4.1 Naïve Bayes Classifier

The NB classifier is a machine learning classification method based on strong hypotheses on the independence of common variables in the application of Bayes theory [60]. The NB classifier assumes there is independence between the conditional expectation variables on the response and the numerical distribution of the mean and standard deviation digital indicators from the training dataset. The figure below shows how the NB classifier works.

Figure 3. 2: Workings of the Naive Bayes classifier [61].

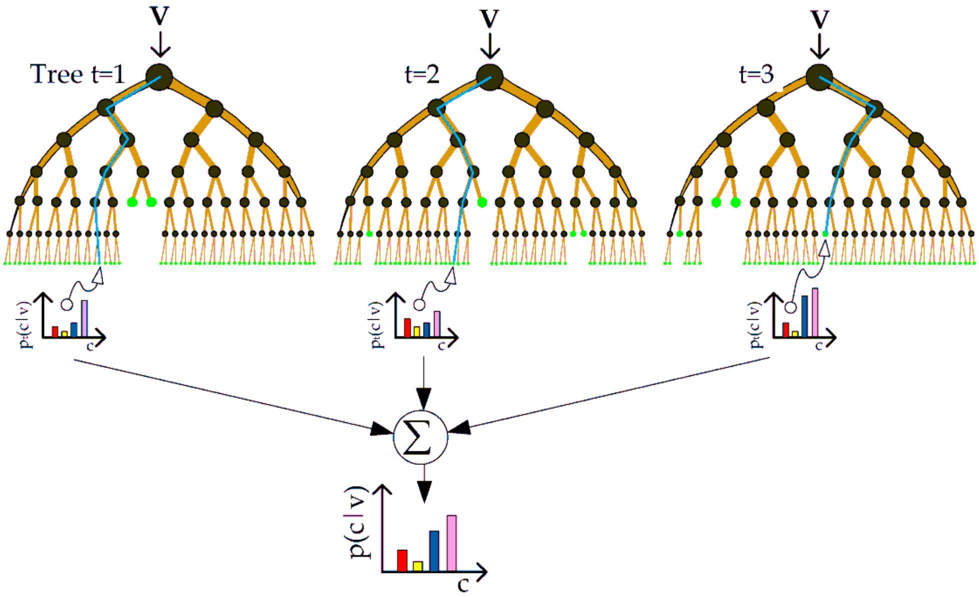


Naïve Bayes models are mostly used as an ersatz to the decision tree classification technique to solve classification problems. When we construct an NB classifier, each row in the training data which contains at least one NA will be exceeded totally. If the test data have missing values, then these predictors are ignored in the probability calculation during the prediction. There are three types of Naive Bayes classifier: Gaussian, multinomial and Bernoulli with Gaussian. If the attribute values are continuous, it is assumed that the values are related to each category distributed according to the Gaussian normal distribution [62]. The Bernoulli classifier is used with the data distributed according to multivariate Bernoulli distributions and it is better with binary data. Multinomial is better to use with data that is multinomially distributed. It is considered one of the standard classic algorithms applied in text classification problems. We have applied these collections of NB classifier techniques with our system on the different kinds of clinical datasets.

3.4.2 Random Forest Classifier

The Random Forest data mining algorithm is defined as a supervised classification technique. From its seemingly self-explanatory name, this classifier works to create a forest and make it random. This technique is often called random because in this algorithm, each individual decision tree is expert in different subsections of the training data, and each node for each decision tree is broken down using a randomly selected attribute from the dataset. By introducing this element of randomization, the technique is able to initialize models that are not related to each other if a training dataset is entered with an outcome and features in DT, which means you a set of rules is formatted [63]. These rules can be used to make predictions. The Random Forest works in different stages, so it randomly selects *f* features from a total of *m* features where $k < m$, and among the *f* features, it will compute node *d* using a better split point, after which it works to divide the node into a child node by splitting. Finally, it attempts to form a forest by iterating the stages *n* times to form *n* trees. The figure below shows the working shape of the random forest classifier.

Figure 3. 3: Working form of the random forest [64].



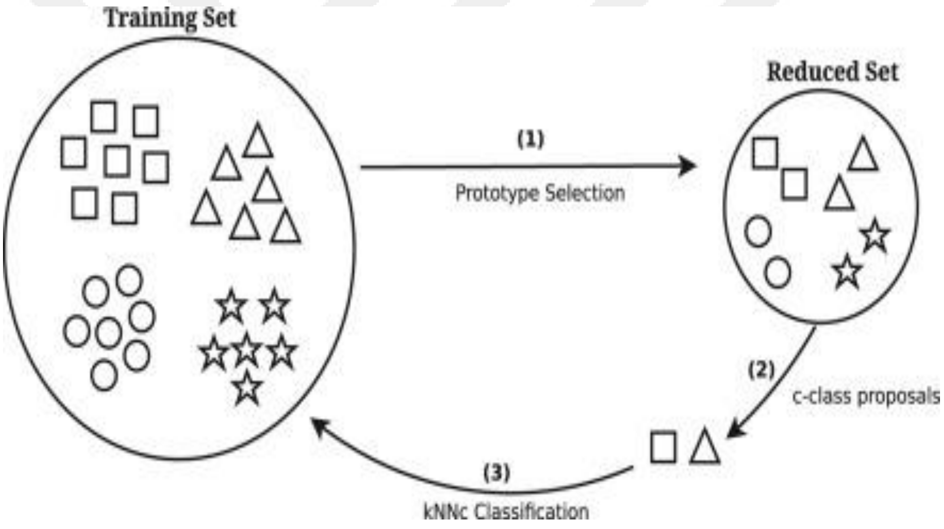
If we wanted to predict whether one’s son likes animated movies, one would need to collect the previous animated movies that one liked and take some of the movies’ features as the input. Then, through a decision tree algorithm, rules can be created. One can then enter the features of this movie to ascertain whether it will appeal to the son. The calculation process of these nodes and the formation of the rules are used to gain information [65]. The main difference between the Random Forest and the Decision

Tree is that the process of the random forest is to find the root node. The splitting attribute node will run haphazardly, which is not possible with the Decision Tree technique. We selected the Random Forest technique because of its ability to be used for both classification and regression. Additionally, this classifier will not over fit the system with the possibility of handling any missing values.

3.4.3 KNN Classifier

The KNN classifier is a nonparametric learning algorithm and one of the supervised data mining techniques with which we can classify data points into a particular category with the help of the training group. Therefore, it captures information in all training cases and classifies new cases according to any similarities [40]. The figure below shows the KNN classifier.

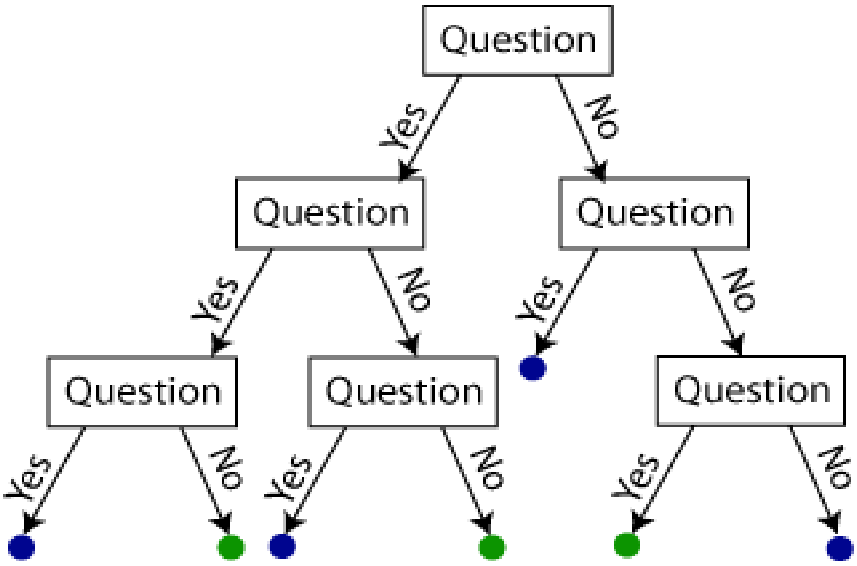
Figure 3. 4: KNN classifier [66].



3.4.4 Decision Tree Classifier

DT adopts classification or regression models in the form of a tree structure. It divides datasets into smaller subsections while the linked decision tree actually advances. The final outcome is a tree holding the decision node with a holding leaf node [67]. The decision node takes two or more branches and signifies a classification or decision. The uppermost verdict node in the tree that agrees with the best predictor is named the root node. Decision trees can hold definite and numeric data. The figure below shows the general form of the Decision Tree classification technique.

Figure3. 5: General form of the Decision Tree classifier [68].

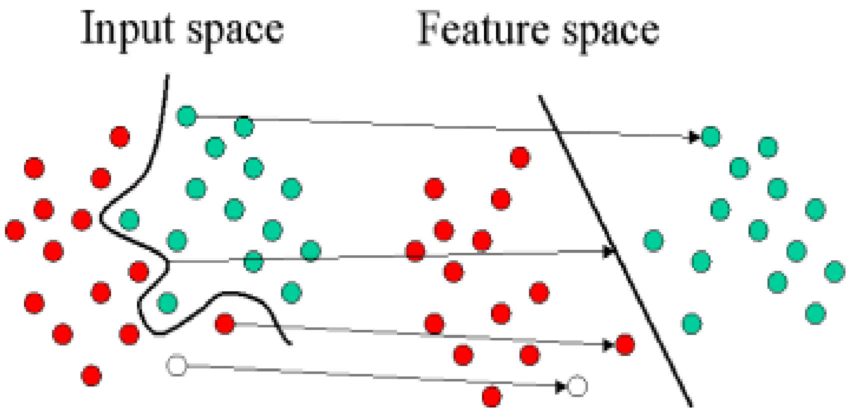


From the advantages of decision tree machine learning technique that it is require relatively few effort from users to set up data and the nonlinear relationships between parameters do not affect tree performance also it is considered as the best feature of using trees for analytics. The dis advantages of this algorithm that it is will not give the right answer but it will give many possible answers and it is tried several combinations of variables to get the better split.

3.4.5 SVM Classifier

The SVM rating is based on the concept of decision levels that define decision boundaries. The decision level separates a group of objects with different class memberships. SVM finds vectors ("vector support") that specify commas that give the largest separated categories. SVM supports both binary and multiclass targets. For linear SVM, the largest and smallest hyperplane margins which divide the training dataset should be found [69]. If the training data is detachable, two hyperplanes are selected in a manner that separates the data. There are no points between them and the distance between them is known as a margin. The figure below shows the general SVM classification technique.

Figure 3. 6: Form of the SVM classifier [70].

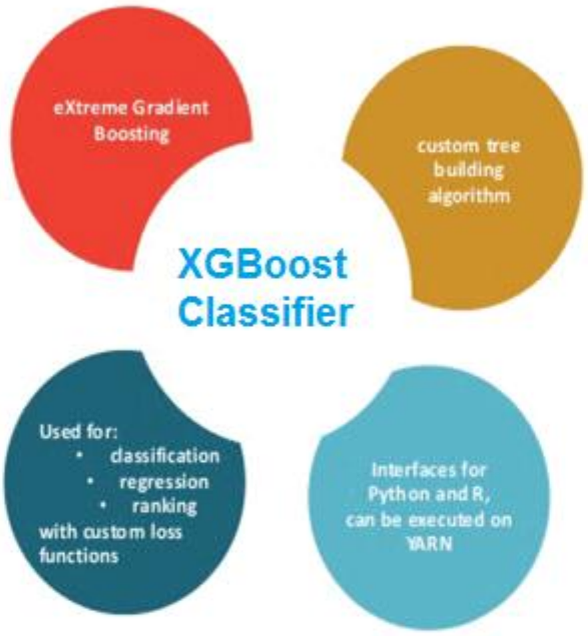


The extension of the SVM approach to nonlinear classification depends on the conversion of input variables and the ability to adapt SVM procedures effectively to converted input spaces. The idea of converting an input space using basic functions is an obvious way to extend linear techniques to a nonlinear setup [71]. The kernel is a function which converts input data to a higher-dimensional space where the problem is resolved. There are many kernel shapes, such as RBF and polynomial.

3.4.6 XGBoost classifier

Boosting or enhancement techniques improve the accuracy of a predictive function by applying the function repeatedly in a string and merging the output of each function with the weighting so that the overall error of the prediction is reduced. In many cases, the predictive accuracy of such a series exceeds the accuracy of the basic function used alone [72]. XGBoost is a machine learning supervised algorithm consisting of a library for the development of high-speed, high-performance boosting tree models. Its speed is due to the parallel computing behind the scenes. It has been very popular in recent years because of its scalability and effectiveness. The figure below shows the XGBoost algorithm.

Figure 3. 7: XGBoost algorithm [73].


XGBoost Classifier
eXtreme Gradient Boosting
custom tree building algorithm
Used for:
• classification
• regression
• ranking
with custom loss functions
Interfaces for Python and R, can be executed on YARN

From the advantages of this technique that it is easy to use and implementation because it contain a library which can perform with R and python also it computational efficiency with ability to achieve a high accuracy in addition it is feasibility to tune parameter and modify the objective[74]. From the disadvantages of this technique that the high costs or the price this is a cloud service and limited storage in the free version.

CHAPTER 4

4.1 Introduction

In our study, we used seven data mining classification techniques and applied them in appropriate programing languages, such as Python. These seven techniques include the Random Forest, SVM, Naïve Bayes, the logistic regression classifier, the K-nearest neighbor classifier, the Decision Tree and the AdaBoost classifier. Different results and levels of accuracy were achieved with five different types of processed (CHD) medical datasets.

In addition we have contribute this study used a different five kinds of clean and preprocess coronary heart disease datasets in different manner from previous study like ([75],[76],[77],[78],[79],[31]) implemented with several classification techniques to improve and achieve the highest accuracy which is the most important thing in the medical field.

4.2 Results and discussion

Multiple data mining classification algorithms with five different (CHD) datasets were used in our study. We have achieved varying results with accuracy applied using the Python programing language. Also, we have used train test split to divide our dataset into train and test with (5-15) Random State and we used 30% of our dataset as a test size and 70% as training size. When we used this method the accuracy had changed from 80% to 95% in our system.

4.2.1 Naïve Bayes classifier

Our results with the most popular data mining algorithm, the Naïve-Bayes, were applied in the Python programing language with the CHD datasets. With the Cleveland dataset, we achieved 72% as accuracy and with the Hungarian two classes dataset, we achieved 85% as a better accuracy. Moreover, with the Hungarian five class, we achieved 98% best accuracy. The Long-Beach dataset produced 55% accuracy and the Switzerland dataset gave us 41% accuracy in addition to having achieved 82% as a best accuracy using the heart.csv dataset.

To show our results, we used a classification report to describe the precision which is the proportion of properly expected positive remarks to the complete number of predicted positive remarks. Recall is the proportion of fittingly foretold positive remarks to the total number of remarks in the real class. The F1 Score is the slanted average of Precision and Recall and the Support is the number of samples of the true response that lies in that class. The class field refers to the presence of CHD disease in the patient. It is integer valued from 0 (no presence) to 4. The CHD five classes datasets have concentrated on simply attempting to distinguish presence (values 1, 2, 3, 4) from absence (value 0). The tables below show the results with different datasets by classification report with class number.

Table 4. 1: Classification report of cleveland dataset.

Class NO.		**Prec**	**Rec**	**f1. Scor**	**Supp**
CHD Absence (class 0)		0.96	0.91	0.94	58
distinguish CHD presence with four classes	**(Class1)**	0.43	0.60	0.50	15
	(Class2)	0.39	0.50	0.44	14
	(Class3)	0.60	0.27	0.37	11
	(Class4)	0.0	0.0	0.0	2
Avg / total		0.74	0.72	0.72	100

Table 4. 2: Classification report of hungarian two class dataset.

Class NO.	**Prec**	**Rec**	**f1. scor**	**Supp**
CHD Absence (class 0)	0.88	0.89	0.89	65
CHD presence (Class1)	0.78	0.76	0.77	33
Avg / total	0.85	0.85	0.85	98

Table 4. 3: Classification report of hungarian five class.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		1.00	1.00	1.00	63
distinguish CHD presence with four classes	**(Class1)**	1.00	1.00	1.00	12
	(Class2)	1.00	1.00	1.00	5
	(Class3)	0.86	1.00	0.92	12
	(Class4)	1.00	0.67	0.80	6
Avg / total		0.98	0.98	0.98	98

Table 4. 4: Classification report of switzerland dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.0	0.0	0.0	4
distinguish CHD presence with four classes	**(Class1)**	0.41	1.00	0.59	17
	(Class2)	0.0	0.0	0.0	9
	(Class3)	0.0	0.0	0.0	10
	(Class4)	0.0	0.0	0.0	1
Avg / total		0.17	0.41	0.24	41

Table 4. 5: Classification report of long-beach dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.92	0.52	0.67	23
distinguish CHD presence with four classes	**(Class1)**	0.20	0.24	0.22	17
	(Class2)	0.30	0.43	0.35	14
	(Class3)	0.25	0.27	0.26	11
	(Class4)	0.0	0.0	0.0	1
Avg / total		0.48	0.38	0.41	66

Table 4. 6: Classification report of heart.csv dataset.

Class NO.	Prec	Rec	f1. scor	Supp
CHD Absence (class 0)	0.90	0.78	0.83	58
CHD presence (Class1)	0.74	0.88	0.80	42
Avg / total	0.83	0.82	0.82	100

4.2.2 Random forest classifier

With the Random Forest classifier, we achieve the best results and accuracy in minimum time. Using the Cleveland dataset, we achieved 94% as the best accuracy. The Hungarian five classes gave us 99% as the best accuracy in our study and the Hungarian two classes achieved 99% as a better accuracy. The Switzerland dataset gave 95% as a better accuracy and the Long-Beach dataset achieved 91% as the best accuracy. The heart.csv dataset achieved 84% as a better accuracy with results show in the tables below.

Table 4. 7: Classification report of cleveland dataset.

Class NO.		**Prec**	**Rec**	**f1. scor**	**Supp**
CHD Absence (class 0)		1.00	1.00	1.00	67
distinguish CHD presence with four classes	**(Class1)**	0.71	1.00	0.83	10
	(Class2)	0.89	0.73	0.80	11
	(Class3)	1.00	0.70	0.82	10
	(Class4)	0.67	1.00	0.80	2
Avg / total		0.95	0.94	0.94	100

Table 4. 8: Classification report of hungarian five class.

Class NO.		**Prec**	**Rec**	**f1. scor**	**Supp**
CHD Absence (class 0)		1.00	1.00	1.00	68
distinguish CHD presence with four classes	**(Class1)**	1.00	1.00	1.00	12
	(Class2)	1.00	1.00	1.00	8
	(Class3)	0.91	1.00	0.95	10
	(Class4)	1.00	0.50	0.67	2
Avg / total		0.99	0.99	0.99	100

Table 4. 9: Classification report of hungarian two class datasets.

Class NO.	Prec	Rec	f1. Scor	Supp
CHD Absence (class 0)	0.98	1.00	0.99	57
CHD presence (Class1)	1.00	0.98	0.99	41
Avg / total	0.99	0.99	0.99	98

Table 4. 10: Classification report of Switzerland datasets.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		1.00	1.00	1.00	2
distinguish CHD presence with four classes	**(Class1)**	1.00	1.00	1.00	16
	(Class2)	0.92	1.00	0.96	12
	(Class3)	0.90	1.00	0.95	9
	(Class4)	0.0	0.0	0.0	2
Avg / total		0.91	0.95	0.93	41

Table 4. 11: Classification report of long-beach dataset.

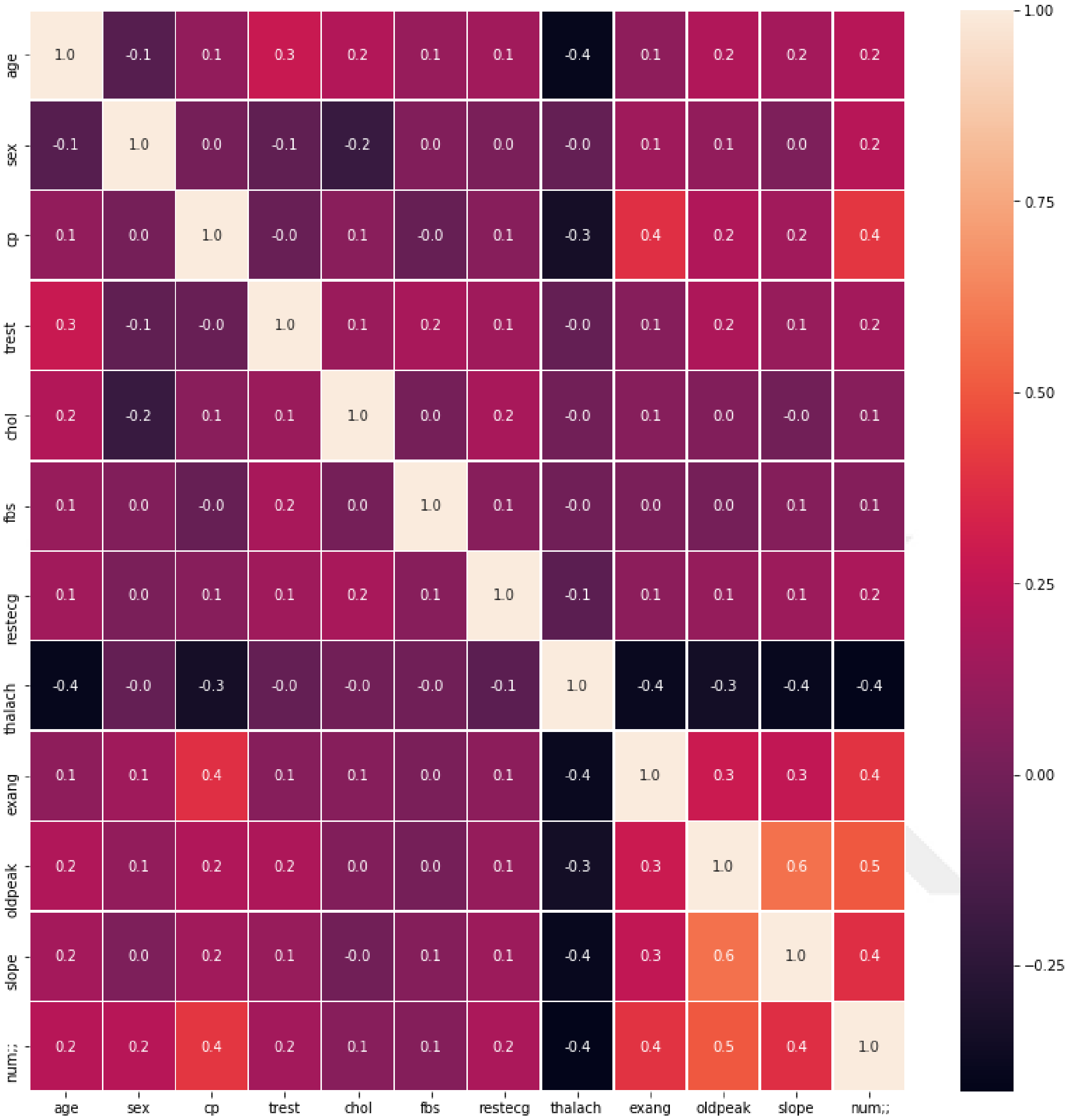
Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.92	0.92	0.92	12
distinguish CHD presence with four classes	(Class1)	0.92	0.96	0.94	23
	(Class2)	0.94	0.94	0.94	16
	(Class3)	0.83	0.91	0.87	11
	(Class4)	1.00	0.50	0.67	4
Avg / total		0.91	0.91	0.91	66

Table 4. 12: Classification report of heart.csv dataset.

Class NO.	prec	rec	f1. Scor	Supp
CHD Absence (class 0)	0.84	0.90	0.87	58
CHD presence (Class1)	0.84	0.76	0.80	42
Avg / total	0.84	0.84	0.84	100

Also in the figures below we show the description of our (CHD) datasets as a heat map and histogram show the distribution of classes as values in these datasets.

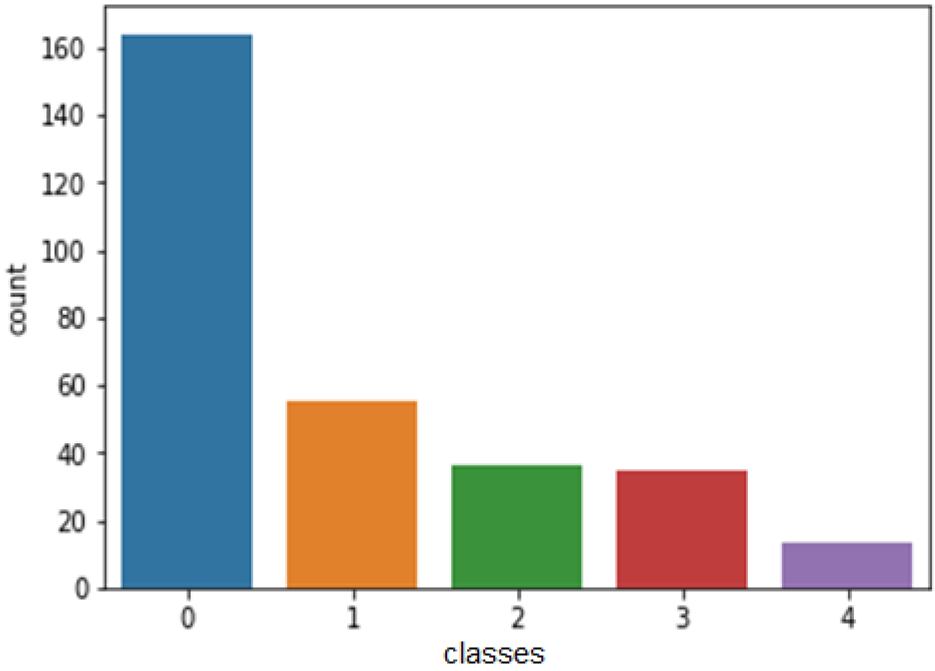
Figure 4. 1: Heat map of the Cleveland datasets.



	age	sex	cp	trest	chol	fbs	restecg	thalach	exang	oldpeak	slope	num;;
age	1.0	-0.1	0.1	0.3	0.2	0.1	0.1	-0.4	0.1	0.2	0.2	0.2
sex	-0.1	1.0	0.0	-0.1	-0.2	0.0	0.0	-0.0	0.1	0.1	0.0	0.2
cp	0.1	0.0	1.0	-0.0	0.1	-0.0	0.1	-0.3	0.4	0.2	0.2	0.4
trest	0.3	-0.1	-0.0	1.0	0.1	0.2	0.1	-0.0	0.1	0.2	0.1	0.2
chol	0.2	-0.2	0.1	0.1	1.0	0.0	0.2	-0.0	0.1	0.0	-0.0	0.1
fbs	0.1	0.0	-0.0	0.2	0.0	1.0	0.1	-0.0	0.0	0.0	0.1	0.1
restecg	0.1	0.0	0.1	0.1	0.2	0.1	1.0	-0.1	0.1	0.1	0.1	0.2
thalach	-0.4	-0.0	-0.3	-0.0	-0.0	-0.0	-0.1	1.0	-0.4	-0.3	-0.4	-0.4
exang	0.1	0.1	0.4	0.1	0.1	0.0	0.1	-0.4	1.0	0.3	0.3	0.4
oldpeak	0.2	0.1	0.2	0.2	0.0	0.0	0.1	-0.3	0.3	1.0	0.6	0.5
slope	0.2	0.0	0.2	0.1	-0.0	0.1	0.1	-0.4	0.3	0.6	1.0	0.4
num;;	0.2	0.2	0.4	0.2	0.1	0.1	0.2	-0.4	0.4	0.5	0.4	1.0

The different of colors in this heat map show the most effective attributes in the classification report the lowest number takes the darkness color while the brightest colors are the high numbers and the most effective between the other attributes. The multiple histograms below show different distributions of dataset values.

Figure 4. 2: Histogram of the Cleveland datasets.



In this figure with the Cleveland datasets, we present different ratios between the classes and the numerical ratio which distribute for each class. We found that the number of people who did not have CHD with class (0) was higher than the people who had that disease with other classes in multiple stages.

Figure 4. 3: Histogram of the Hungarian five class dataset.

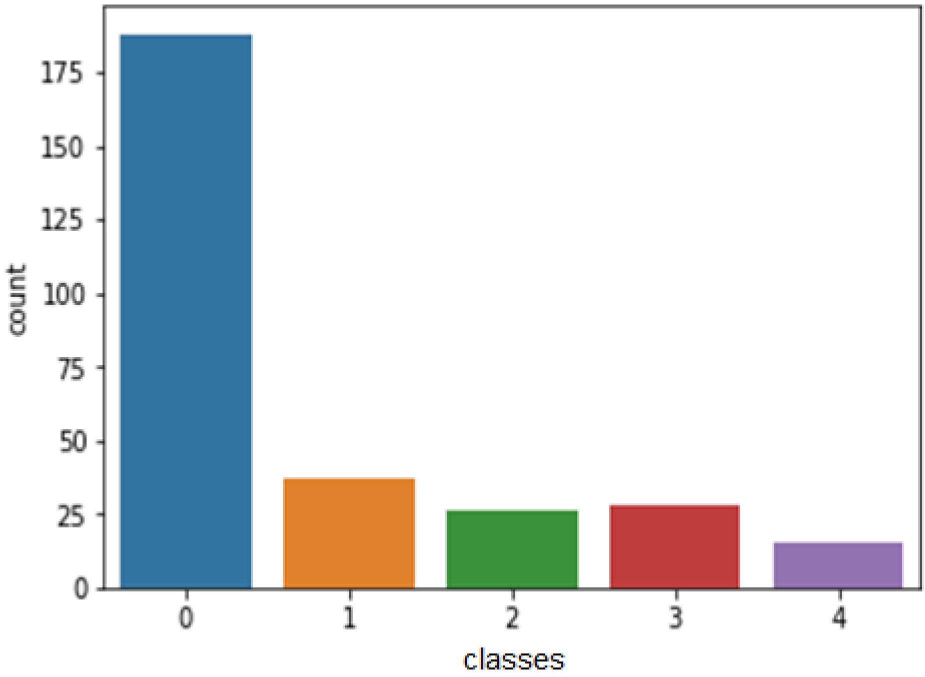
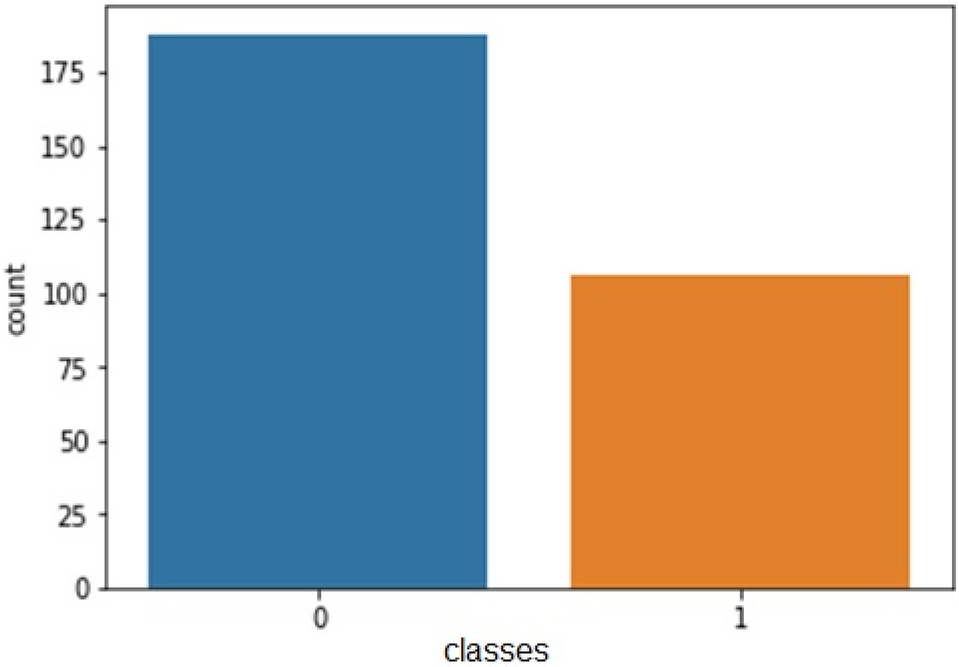
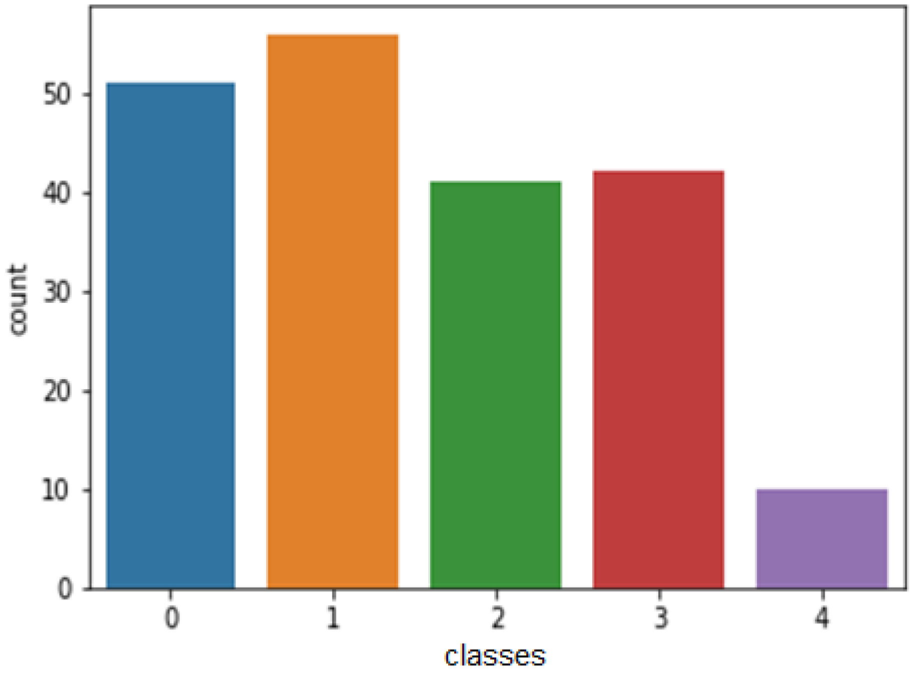


Figure 4. 4: Histogram of the Hungarian two class dataset.



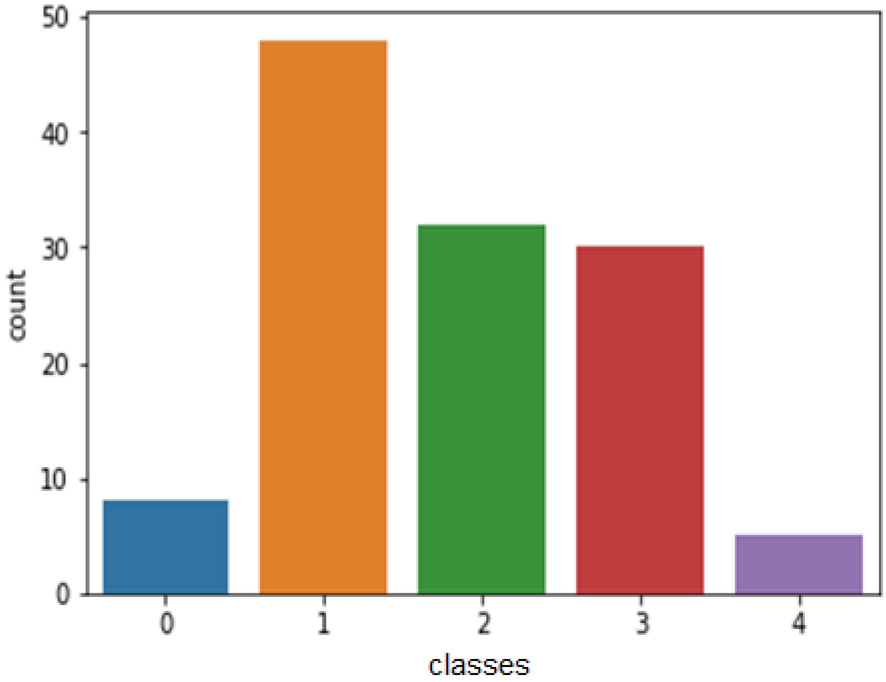
In the figure above with the Hungarian five and two class datasets, we present different ratios between the classes and the numerical ratio which distribute for each class. We found that the number of people who did not have CHD with class (0) was higher than the people who had that disease with other classes in multiple stages with the same distribution and numerical ratios for Hungarian two classes datasets.

Figure 4. 5: Histogram of long-beach dataset.



With the figure above, the distribution of data between classes in the Long-Beach showed that the number of people with CHD according to class (0) was less than those people unaffected by CHD with other classes in different stages.

Figure 4. 6: Histogram of switzerland dataset.



Also with the figures above, the distribution of data between classes in the Switzerland datasets showed that the number of people with CHD according to class (0) was less than those people unaffected by CHD with other classes.

4.2.3 KNN classifier

With the KNN classifier, we achieved 63% accuracy with the Cleveland dataset and with the Hungarian five class dataset we achieved 76% accuracy. Using the Hungarian two classes, we achieved 81% as a better accuracy. Long-Beach gave us 35% accuracy, the Switzerland dataset gave us 46% accuracy and the heart.csv dataset gave us 84% best accuracy and results that show in the tables below.

Table 4. 13: Classification report of cleveland datasets.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.83	0.85	0.84	67
distinguish CHD presence with four classes	**(Class1)**	0.19	0.40	0.26	10
	(Class2)	0.29	0.18	0.22	11
	(Class3)	0.0	0.0	0.0	10
	(Class4)	0.0	0.0	0.0	2
Avg / total		0.60	0.63	0.61	100

Table 4. 14: Classification report of hungarian five class dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.86	0.96	0.90	69
distinguish CHD presence with four classes	**(Class1)**	0.71	0.45	0.56	11
	(Class2)	0.33	0.22	0.27	9
	(Class3)	0.12	0.25	0.17	4
	(Class4)	0.0	0.0	0.0	5
Avg / total		0.72	0.76	0.73	98

Table 4. 15: Classification report of hungarian two class datasets.

Class NO.	Prec	rec	f1. scor	Supp
CHD Absence (class 0)	0.81	0.88	0.85	59
CHD presence (Class1)	0.79	0.69	0.74	39
Avg / total	0.81	0.81	0.80	98

Table 4. 16: Classification report of long- beach dataset.

Class NO.		prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.37	0.58	0.45	12
distinguish CHD presence with four classes	**(Class1)**	0.35	0.58	0.44	12
	(Class2)	0.42	0.29	0.34	17
	(Class3)	0.29	0.12	0.17	17
	(Class4)	0.0	0.0	0.0	2
Avg / total		0.34	0.35	0.32	60

Table 4. 17: Classification report of switzerland dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.50	1.00	0.67	1
distinguish CHD presence with four classes	**(Class1)**	0.44	0.80	0.57	15
	(Class2)	0.50	0.25	0.33	12
	(Class3)	0.50	0.25	0.33	12
	(Class4)	0.0	0.0	0.0	1
Avg / total		0.47	0.46	0.42	41

Table 4. 18: Classification report of heart.csv dataset.

Class NO.	Prec	rec	f1. scor	Supp
CHD Absence (class 0)	0.86	0.86	0.86	58
CHD presence (Class1)	0.81	0.81	0.81	42
Avg / total	0.84	0.84	0.84	100

4.2.4 Decision tree classifier

Our study results used the Decision Tree classifier with different CHD datasets. We achieved 52% accuracy using the Cleveland dataset and 64% accuracy with the Hungarian five classes dataset. The heart.csv dataset gave us 77% as a better accuracy and 77% accuracy was achieved using the Hungarian two class dataset. The Switzerland dataset gave us 49% accuracy and with the Long-Beach dataset we had 33% accuracy. The tables below show the results with different datasets by classification report with class number.

Table 4. 19: Classification report of cleveland dataset.

Class NO.		prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.72	0.78	0.75	49
distinguish CHD presence with four classes	**(Class1)**	0.37	0.30	0.33	23
	(Class2)	0.31	0.29	0.30	14
	(Class3)	0.27	0.30	0.29	10
	(Class4)	0.0	0.0	0.0	4
Avg / total		0.51	0.52	0.51	100

Table 4. 20: Classification report with hungarian five class dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.90	0.79	0.84	68
distinguish CHD presence with four classes	**(Class1)**	0.29	0.57	0.38	7
	(Class2)	0.22	0.20	0.21	10
	(Class3)	0.15	0.29	0.20	7
	(Class4)	0.50	0.17	0.25	6
Avg / total		0.71	0.64	0.66	98

Table 4. 21: Classification report of heart.csv dataset.

Class NO.	Prec	rec	f1. scor	Supp
CHD Absence (class 0)	0.84	0.70	0.76	53
CHD presence (Class1)	0.71	0.85	0.78	47
Avg / total	0.78	0.77	0.77	100

Table 4. 22: Classification report of hungarian two class dataset.

Class NO.	prec	rec	f1. scor	Supp
CHD Absence (class 0)	0.81	0.82	0.81	61
CHD presence (Class1)	0.69	0.68	0.68	37
Avg / total	0.76	0.77	0.76	98

Table 4. 23: Classification report of switzerland dataset.

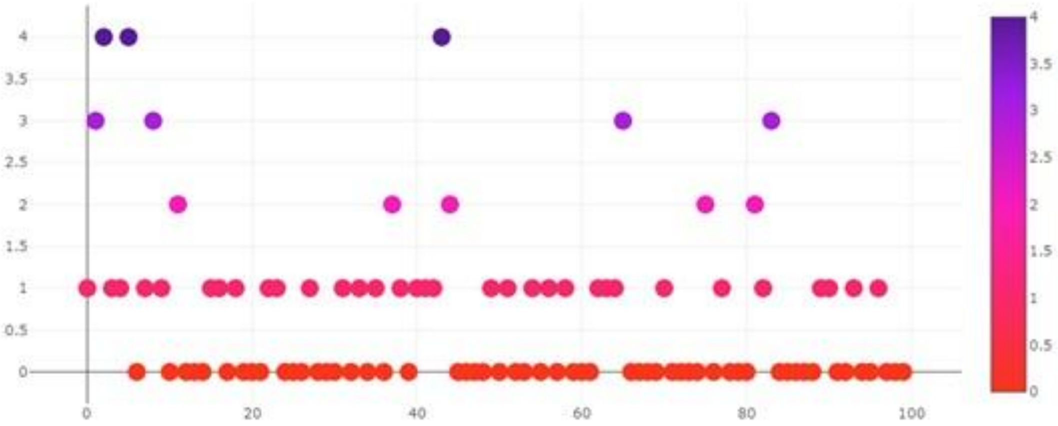
Class NO.		prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.33	1.00	0.50	1
distinguish CHD presence with four classes	**(Class1)**	0.48	0.67	0.56	15
	(Class2)	0.62	0.42	0.50	12
	(Class3)	0.50	0.33	0.40	12
	(Class4)	0.0	0.0	0.0	1
Avg / total		0.51	0.49	0.48	41

Table 4. 24: Classification report of long-beach dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.29	0.38	0.33	13
distinguish CHD presence with four classes	**(Class1)**	0.33	0.47	0.39	15
	(Class2)	0.29	0.22	0.25	18
	(Class3)	0.55	0.33	0.41	18
	(Class4)	0.0	0.0	0.0	2
Avg / total		0.36	0.33	0.34	66

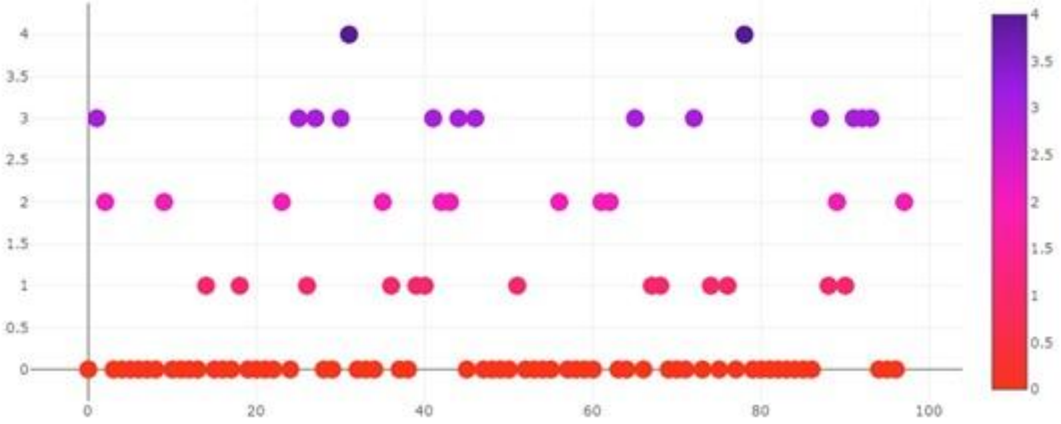
The figures below show the distribution of values of the Cleveland, Hungarian five classes and heart.csv datasets.

Figure 4. 7: Cleveland dataset is values distribution.



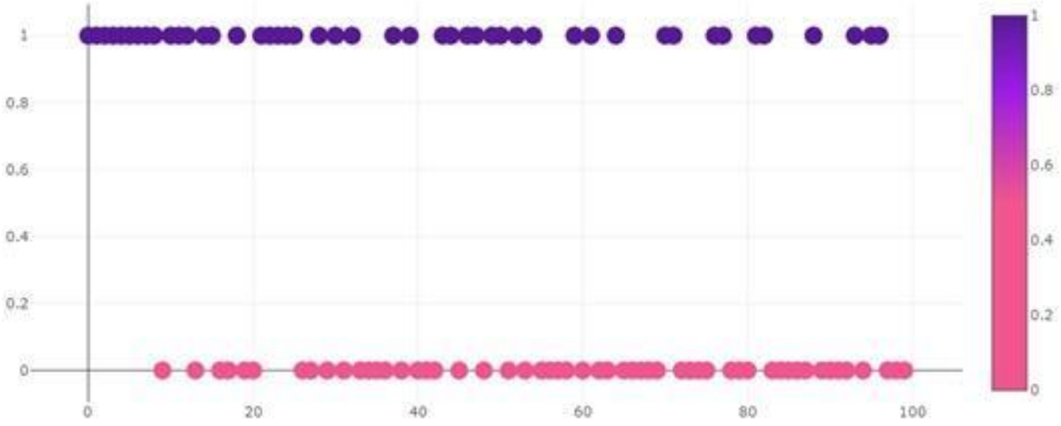
From the figure above we can see that the data distribution in Cleveland outcome or prediction attribute after the classification process this data consists of five classes and we see that classes (0) and (1) contains the largest proportion of this data the class zero refers to the cases with no CHD disease while the other classes refers to cases that indicate the presence of CHD with different color for each class.

Figure 4. 8: Hungarian five classes is values distribution.



Also in the Hungarian datasets figure above we can see that the distribution of the most data is in the class (0) which expresses the uninfected cases or the patients who do not have the CHD disease with different color ratios for each class.

Figure 4. 9: Heart.csv is values are distribution.



In this figure that expresses of Heart.csv dataset we can see that there are two classes, one for infected patients (1) and the other for non-infected (0) patients with different colors and data distribution method for each class after classification process.

4.2.5 SVM classifier

In this study, we used the SVM classifier with the heart.csv dataset and achieved 86% accuracy. With the Cleveland dataset, we achieved 69% accuracy. The Hungarian five classes dataset gave 69% accuracy and the Hungarian two class dataset gave 84% better accuracy. The Long-Beach dataset gave 35% accuracy and the Switzerland dataset gave 44% accuracy. The tables below show the results with different datasets by classification report with class number.

Table 4. 25: Classification report of heart.csv dataset.

Class NO.	Prec	rec	f1. scor	Supp
CHD Absence (class 0)	0.85	0.91	0.88	58
CHD presence (Class1)	0.87	0.79	0.82	42
Avg / total	0.86	0.86	0.86	100

Table 4. 26: Classification report of cleveland dataset.

Class NO.		prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.91	0.90	0.90	67
distinguish CHD presence with four classes	**(Class1)**	0.22	0.40	0.29	10
	(Class2)	0.31	0.36	0.33	11
	(Class3)	0.33	0.10	0.15	10
	(Class4)	0.0	0.0	0.0	2
Avg / total		0.70	0.69	0.69	100

Table 4. 27: Classification report with hungarian five class dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.86	0.87	0.86	68
distinguish CHD presence with four classes	**(Class1)**	0.19	0.43	0.26	7
	(Class2)	0.50	0.30	0.37	10
	(Class3)	0.25	0.14	0.18	7
	(Class4)	0.67	0.33	0.44	6
Avg / total		0.72	0.69	0.69	98

Table 4. 28: Classification report of hungarian two class dataset.

Class NO.	Prec	rec	f1. scor	supp
CHD Absence (class 0)	0.84	0.90	0.87	59
CHD presence (Class1)	0.83	0.74	0.78	39
Avg / total	0.84	0.84	0.83	98

Table 4. 29: Classification report of long-beach dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.58	0.48	0.52	23
distinguish CHD presence with four classes	(Class1)	0.20	0.29	0.24	17
	(Class2)	0.33	0.14	0.20	14
	(Class3)	0.38	0.45	0.42	11
	(Class4)	0.0	0.0	0.0	1
Avg / total		0.39	0.35	0.36	66

Table 4. 30: Classification report of switzerland dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.00	0.00	0.00	4
distinguish CHD presence with four classes	(Class1)	0.52	0.71	0.60	17
	(Class2)	0.25	0.33	0.29	9
	(Class3)	0.75	0.30	0.43	10
	(Class4)	0.0	0.0	0.0	1
Avg / total		0.45	0.44	0.42	41

4.2.6 Logistic regression classifier

With the logistic regression classifier, we achieved 87% accuracy using the heart.csv dataset and with the Cleveland dataset, we had 80% accuracy. Using the Hungarian five class dataset, we achieved 89% better accuracy and with the Hungarian two class dataset, we had 99% best accuracy. The Long-Beach dataset gave 76% accuracy and the Switzerland dataset gave 78% accuracy and different results show in the tables below.

Table 4. 31: Classification report of heart.csv dataset.

Class NO.	Prec	Rec	f1. scor	Supp
CHD Absence (class 0)	0.95	0.85	0.90	67
CHD presence (Class1)	0.75	0.91	0.82	33
Avg / total	0.88	0.87	0.87	100

Table 4. 32: Classification report of cleveland dataset.

Class NO.		prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.96	1.00	0.98	67
distinguish CHD presence with four classes	**(Class1)**	0.46	0.60	0.52	10
	(Class2)	0.33	0.18	0.24	11
	(Class3)	0.50	0.50	0.50	10
	(Class4)	0.0	0.0	0.0	2
Avg / total		0.77	0.80	0.78	100

Table 4. 33: Classification report with hungarian five class dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.97	1.00	0.99	68
distinguish CHD presence with four classes	(Class1)	0.50	0.71	0.59	7
	(Class2)	1.00	0.40	0.57	10
	(Class3)	0.64	1.00	0.78	7
	(Class4)	1.00	0.50	0.67	6
Avg / total		0.92	0.89	0.88	98

Table 4. 34: Classification report of hungarian two class dataset.

Class NO.	Prec	Rec	f1. scor	supp
CHD Absence (class 0)	1.00	0.98	0.99	61
CHD presence (Class1)	0.97	1.00	0.99	37
Avg / total	0.99	0.99	0.99	98

Table 4. 35: Classification report of long-beach datasets.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.85	0.94	0.89	18
distinguish CHD presence with four classes	**(Class1)**	0.84	0.84	0.84	19
	(Class2)	0.67	0.73	0.70	11
	(Class3)	0.60	0.75	0.67	12
	(Class4)	0.0	0.0	0.0	6
Avg / total		0.69	0.76	0.72	66

Table 4. 36: Classification report of switzerland dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		0.0	0.0	0.0	1
distinguish CHD presence with four classes	**(Class1)**	0.75	1.00	0.86	15
	(Class2)	0.73	0.67	0.70	12
	(Class3)	0.90	0.75	0.82	12
	(Class4)	0.0	0.0	0.0	1
Avg / total		0.75	0.78	0.76	41

4.2.7 Ada Boost classifier

Using the Adaboost classifier, we achieved 82% accuracy and using the heart.csv dataset, we had 86% as a better accuracy. With the Hungarian five class dataset, we achieved 87% accuracy and the Hungarian two class dataset gave us 82% accuracy. We had 78% accuracy from the Switzerland dataset and the Long-Beach dataset gave 80% accuracy with different results show in the tables below.

Table 4. 37: Classification report of cleveland dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		1.00	1.00	1.00	58
distinguish CHD presence with four classes	**(Class1)**	0.51	1.00	0.68	19
	(Class2)	0.0	0.0	0.0	8
	(Class3)	0.0	0.0	0.0	10
	(Class4)	1.00	1.00	1.00	5
Avg / total		0.73	0.82	0.76	100

Table 4. 38: Classification report of heart.csv dataset.

Class NO.	Prec	Rec	f1. scor	Supp
CHD Absence (class 0)	0.79	0.96	0.86	46
CHD presence (Class1)	0.95	0.78	0.86	54
Avg / total	0.88	0.86	0.86	100

Table 4. 39: Classification report of hungarian two class dataset.

Class NO.	Prec	Rec	f1. scor	Supp
CHD Absence (class 0)	0.81	0.94	0.87	63
CHD presence (Class1)	0.84	0.60	0.70	35
Avg / total	0.82	0.82	0.81	98

Table 4. 40: Classification report with hungarian five class dataset.

Class NO.		prec	Rec	f1. scor	supp
CHD Absence (class 0)		1.00	1.00	1.00	69
distinguish CHD presence with four classes	**(Class1)**	0.46	1.00	0.63	11
	(Class2)	0.0	0.0	0.0	9
	(Class3)	0.0	0.0	0.0	4
	(Class4)	1.00	1.00	1.00	5
Avg / total		0.81	0.87	0.83	98

Table 4. 41: Classification report of switzerland dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		1.00	1.00	1.00	4
distinguish CHD presence with four classes	**(Class1)**	1.00	1.00	1.00	16
	(Class2)	0.50	1.00	0.67	9
	(Class3)	0.0	0.0	0.0	9
	(Class4)	1.00	1.00	1.00	3
Avg / total		0.67	0.78	0.71	41

Table 4. 42: Classification report of long-beach dataset.

Class NO.		Prec	Rec	f1. scor	Supp
CHD Absence (class 0)		1.00	1.00	1.00	22
distinguish CHD presence with four classes	**(Class1)**	1.00	1.00	1.00	18
	(Class2)	0.46	1.00	0.63	11
	(Class3)	0.0	0.0	0.0	13
	(Class4)	1.00	1.00	1.00	2
Avg / total		0.71	0.80	0.74	66

In addition when we applied the Xgboost classifier with the heart.csv dataset, we achieved 85% accuracy. The table below show and describes the results of the Xgboost classification report using the heart.csv dataset.

Table 4. 43: Classification report of heart.csv dataset.

Class NO.	prec	Rec	f1. scor	Supp
CHD Absence (class 0)	0.90	0.81	0.85	53
CHD presence (Class1)	0.81	0.89	0.85	47
Avg / total	0.85	0.85	0.85	100

CHAPTER 5

Conclusion

This study was performed to create and determine the best way to diagnose a most common disease (CHD). We used the Python programing language to apply data mining classification techniques to classify Cleveland , Hungarian with five classes and 10 features, Hungarian with two class and 10 features, Switzerland and Long-Beach on the heart.csv coronary heart disease datasets. We produced better results and accuracy through the use of the Random Forest (RF) algorithm at 99% with the Hungarian two class, and 94% with the Cleveland dataset. In addition to the Naive-Bayes Gausian algorithm, we obtained the best accuracy at 98% and 99% using ther Random Forest classifier with the Hungarian five class dataset, with the Switzerland dataset we obtained 95% better accuracy for this dataset using the Random Forest algorithm, 80% accuracy using the AdaBoost classifier, with the Long-Beach dataset, we achieved 91% better accuracy using the Random Forest algorithm and 78% accuracy using the logistic regression classifier. In addition, with the heart.csv dataset, we achieved 87% better accuracy using the logistic regression classifier and 86% accuracy using the AdaBoost classifier. The best results and highest accuracy in this study were due to the trained test split and preprocessing for the CHD dataset in a different manner to produce a simple system in a short time and at low cost for diagnoses of coronary heart disease. In Table 5.1,we show the comparison between this study and previous studies.

In future work, we will attempt to group or combine machine learning techniques working together to obtain better results with best high accuracy to improve the performance of this system with the CHD diagnosis problem.

Table 5. 1: Comparison with previous studies.

References number	Objective	Data mining techniques	Accuracy
C.Kalaiselvi [80]	Diagnosis of heart disease	NB	94.4%
		DT	96%
		KNN	96.5%
Randa El-Bialy[29]	Feature analysis of coronary artery disease datasets	C4.5	78.5%
		Fast DT	77.5%
Khushboo Chandel[81]	A comparative study on thyroid disease detection	KNN	93.4%
		NB	22.5%
Wiharto , Hari Kusnanto[31]	Diagnosis of Coronary Heart Disease	C4.5	82%
		mSMOTE+C4.5	88%
		mSMOTE+IG+ C4.5	90%
Priyanka N, Pushpa RaviKumar[79]	predicting the heart diseases	DT	89%
		NB	82%
Our study with best accuracy	Diagnoses of coronary heart disease	RF	99%
		Gausian	98%
		Adaboost	87%
		Logistic regression	89%

References

[1] R. Alizadehsani, J. Habibi, M. J. Hosseini, H. Mashayekhi, R. Boghrati, A. Ghandeharioun, *et al.*, "A data mining approach for diagnosis of coronary artery disease," *Computer methods and programs in biomedicine,* vol. 111, pp. 52-61, 2013.

[2] T. L. Assimes and R. Roberts, "Genetics: implications for prevention and management of coronary artery disease," *Journal of the American College of Cardiology,* vol. 68, pp. 2797-2818, 2016.

[3] S. Hulley, D. Grady, T. Bush, C. Furberg, D. Herrington, B. Riggs, *et al.*, "Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women," *Jama,* vol. 280, pp. 605-613, 1998.

[4] A. M. K. Kaul, *Acute and chronic rejection: Compartmentalization and kinetics of counterbalancing signals in cardiac transplants*: Cleveland State University, 2014.

[5] P. W. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation,* vol. 97, pp. 1837-1847, 1998.

[6] M. G. Marmot and P. Elliott, *Coronary heart disease epidemiology: from aetiology to public health*: Oxford Medical Publications, 2005.

[7] G. G. De Backer, "The global burden of coronary heart disease," *Medicographia,* vol. 31, pp. 343-8, 2009.

[8] M. Mirzaei, A. Truswell, R. Taylor, and S. R. Leeder, "Coronary heart disease epidemics: not all the same," *Heart,* vol. 95, pp. 740-746, 2009.

[9] Archer-Soft. (2018). *Data Mining in Healthcare*. Available: http://www.archer-soft.com/en/blog/data-mining-healthcare

[10] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, *Data mining: a knowledge discovery approach*: Springer Science & Business Media, 2007.

[11] T. S. Genders, E. W. Steyerberg, H. Alkadhi, S. Leschka, L. Desbiolles, K. Nieman, *et al.*, "A clinical prediction rule for the diagnosis of coronary artery disease: validation, updating, and extension," *European heart journal*, vol. 32, pp. 1316-1330, 2011.

[12] R. Gupta, S. Gupta, S. Sharma, D. N. Sinha, and R. Mehrotra, "Risk of coronary heart disease among smokeless tobacco users: results of systematic review and meta-analysis of global data," *Nicotine & Tobacco Research*, 2018.

[13] J. P. Ioannidis, "Diagnosis and Treatment of Hypertension in the 2017 ACC/AHA Guidelines and in the Real World," *Jama*, vol. 319, pp. 115-116, 2018.

[14] B. Antonny, J. Bigay, and B. Mesmin, "The Oxysterol-Binding Protein Cycle: Burning Off PI (4) P to Transport Cholesterol," *Annual review of biochemistry*, 2018.

[15] C. K.-S. Leung, "Big data analysis and mining," in *Encyclopedia of Information Science and Technology, Fourth Edition*, ed: IGI Global, 2018, pp. 338-348.

[16] B. Cohen and B. Hasselbring, *Coronary Heart Disease: A Guide to Diagnosis and Treatment*: Addicus Books, 2007.

[17] (2017). *Cronary Heart Disease*. Available: https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease

[18] J. T. Willerson and D. R. Holmes, *Coronary Artery Disease*: Springer London, 2015.

[19] D. Brown, H. Edwards, L. Seaton, and T. Buckley, *Lewis's Medical-Surgical Nursing: Assessment and Management of Clinical Problems*: Elsevier Health Sciences, 2017.

[20] P. Cerrito, *Cases on Health Outcomes and Clinical Data Mining: Studies and Frameworks: Studies and Frameworks*: IGI Global, 2010.

[21] K. Srinivas, B. K. Rani, and A. Govrdhan, "Applications of data mining techniques in healthcare and prediction of heart attacks," *International Journal on Computer Science and Engineering (IJCSE),* vol. 2, pp. 250-255, 2010.

[22] D.-Y. Yeh, C.-H. Cheng, and Y.-W. Chen, "A predictive model for cerebrovascular disease using data mining," *Expert Systems with Applications,* vol. 38, pp. 8970-8977, 2011.

[23] A. S. Abdullah and R. Rajalaxmi, "A data mining model for predicting the coronary heart disease using random forest classifier," in *International Conference in Recent Trends in Computational Methods, Communication and Controls*, 2012.

[24] B. Deekshatulu and P. Chandra, "Classification of heart disease using k-nearest neighbor and genetic algorithm," *Procedia Technology,* vol. 10, pp. 85-94, 2013.

[25] T. Santhanam and E. Ephzibah, "Heart disease classification using PCA and feed forward neural networks," in *Mining Intelligence and Knowledge Exploration*, ed: Springer, 2013, pp. 90-99.

[26] S. B. Patel, P. K. Yadav, and D. D. Shukla, "Predict the diagnosis of heart disease patients using classification mining techniques," *IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS),* vol. 4, pp. 61-64, 2013.

[27] Z. Mahmoodabadi and M. S. Abadeh, "CADICA: Diagnosis of Coronary Artery Disease Using the Imperialist Competitive Algorithm," *Journal of Computing Science and Engineering,* vol. 8, pp. 87-93, 2014.

[28] B. Venkatalakshmi and M. Shivsankar, "Heart disease diagnosis using predictive data mining," *International Journal of Innovative Research in Science, Engineering and Technology,* vol. 3, pp. 1873-7, 2014.

[29] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature analysis of coronary artery heart disease data sets," *Procedia Computer Science,* vol. 65, pp. 459-468, 2015.

[30] B. Bahrami and M. H. Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques," *Journal of Multidisciplinary Engineering Science and Technology (JMEST),* vol. 2, pp. 164-168, 2015.

[31] W. Wiharto, H. Kusnanto, and H. Herianto, "Interpretation of clinical data based on C4. 5 algorithm for the diagnosis of coronary heart disease," *Healthcare informatics research,* vol. 22, pp. 186-195, 2016.

[32] M. S. Lakshmi, D. Haritha, and V. SRKIT, "Heart disease diagnosis using predictive data mining," *International Journal of Computer Science and Information Security,* 2016.

[33] M. Tayefi, M. Tajfard, S. Saffar, P. Hanachi, A. R. Amirabadizadeh, H. Esmaeily, *et al.,* "hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm," *Computer methods and programs in biomedicine,* vol. 141, pp. 105-109, 2017.

[34] S. S. Varghese and L. Devadas, "CORONARY HEART DISEASE PEDICTIONS USING EXPERT SYSTEM AND DEEP LEARNING," 2017.

[35] T. Puyalnithi and M. Vankadara, "Performance Analysis of Classification Algorithms on a Novel Unified Clinical Decision Support Model for Predicting Coronary Heart Disease Risks," 2017.

[36] S. Wikipedia, *Classification Algorithms: Artificial Neural Network, Naive Bayes Classifier, Support Vector MacHine, Boosting, Linear Classifier, Case-Based Reasonin*: General Books, 2013.

[37] N. Deng, Y. Tian, and C. Zhang, *Support Vector Machines: Optimization Based Theory, Algorithms, and Extensions*: CRC Press, 2012.

[38] C. C. Aggarwal, *Data Classification: Algorithms and Applications*: CRC Press, 2015.

[39] R. C. Barros, A. C. P. L. F. de Carvalho, and A. A. Freitas, *Automatic Design of Decision-Tree Induction Algorithms*: Springer International Publishing, 2015.

[40] H. Rajaguru and S. K. Prabhakar, *KNN Classifier and K-Means Clustering for Robust Classification of Epilepsy from EEG Signals. A Detailed Analysis*: Anchor Academic Publishing, 2017.

[41] T. Cao, E. P. Lim, Z. H. Zhou, T. B. Ho, D. Cheung, and H. Motoda, *Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings*: Springer International Publishing, 2015.

[42] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*: Springer New York, 2012.

[43] P. Kashyap, *Machine Learning for Decision Makers: Cognitive Computing Fundamentals for Better Decision Making*: Apress, 2018.

[44] S. P. Rahayu, *Logistic Regression Methods for Classification of Imbalanced Data Sets*: UMP, 2012.

[45] P. Chandrayan, "machine learning logistic regression," 2017.

[46] R. Nisbet, G. Miner, and K. Yale, *Handbook of Statistical Analysis and Data Mining Applications*: Elsevier Science, 2017.

[47] K. Teknomo, "K-means clustering tutorial," *Medicine,* vol. 100, p. 3, 2006.

[48] J. H. Leung, Y.-L. Kuo, T.-W. Weng, and C.-L. Chin, "Hybrid-Neuro-Fuzzy System and Adaboost-Classifier for Classifying Breast Calcification," *Journal of Computers*, vol. 28, pp. 29-42, 2017.

[49] B. Marsh, "Multivariate Analysis of the Vector Boson Fusion Higgs Boson," 2016.

[50] B. Steele, J. Chandler, and S. Reddy, *Algorithms for Data Science*: Springer International Publishing, 2016.

[51] J. MSV, "Machine Learning and Linear Regression for Mere Mortals," 2017.

[52] A. Azzalini and B. Scarpa, *Data Analysis and Data Mining: An Introduction*: Oxford University Press, 2012.

[53] M. Lichman. (2013). *{UCI} Machine Learning Repository*. Available: https://data.world/uci/heart-disease

[54] (2017). *Heart.CSV Datase*. Available: https://www.kaggle.com/zhaoyingzhu/heartcsv/data

[55] G. Svolba, *Data Preparation for Analytics Using SAS*: SAS Institute, 2015.

[56] Y. Wang and L. Ma, "Zheng classification with missing feature values using local-validity approach," *Evidence-Based Complementary and Alternative Medicine*, vol. 2013, 2013.

[57] H. Liu and H. Motoda, *Computational Methods of Feature Selection*: CRC Press, 2007.

[58] G. Dougherty, *Pattern Recognition and Classification: An Introduction*: Springer New York, 2012.

[59] D. Singh, B. Raman, A. K. Luhach, and P. Lingras, *Advanced Informatics for Computing Research: First International Conference, ICAICR 2017, Jalandhar, India, March 17–18, 2017, Revised Selected Papers*: Springer Singapore, 2017.

[60] T. R. Patil and S. Sherekar, "Performance analysis of Naive Bayes and J48 classification algorithm for data classification," *International Journal of Computer Science and Applications,* vol. 6, pp. 256-261, 2013.

[61] K. T. I. Nguyen, "Using Intel® Data Analytics Acceleration Library to Improve the Performance of Naïve Bayes Algorithm in Python," 2016.

[62] C. Chen, *Computer Vision in Medical Imaging*: World Scientific Publishing Company, 2014.

[63] E. Alickovic and A. Subasi, "Medical decision support system for diagnosis of heart arrhythmia using DWT and random forests classifier," *Journal of medical systems,* vol. 40, p. 108, 2016.

[64] "High-Resolution Remote Sensing Data Classification over Urban Areas Using Random Forest Ensemble and Fully Connected Conditional Random Field," 2017.

[65] R. Ani, J. Jose, M. Wilson, and O. Deepa, "Modified Rotation Forest Ensemble Classifier for Medical Diagnosis in Decision Support Systems," in *Progress in Advanced Computing and Intelligent Engineering*, ed: Springer, 2018, pp. 137-146.

[66] J. Calvo-Zaragoza, J. J. Valero-Mas, and J. R. Rico-Juan, "Improving kNN multi-label classification in Prototype Selection scenarios using class proposals," *Pattern Recognition,* vol. 48, pp. 1608-1622, 2015.

[67] M. O. Z and R. Lior, *Data Mining With Decision Trees: Theory And Applications (2nd Edition)*: World Scientific Publishing Company, 2014.

[68] "The Shape of Data," 2017.

[69] C. Campbell and Y. Ying, *Learning with Support Vector Machines*: Morgan & Claypool, 2011.

[70] M. Sharma, "Text Classification using SVM," 2017.

[71] H. Guo, B. Liu, D. Cai, and T. Lu, "Predicting protein–protein interaction sites using modified support vector machine," *International Journal of Machine Learning and Cybernetics,* vol. 9, pp. 393-398, 2018.

[72] *Data Science: Questions and Answers*: George Duckett, 2018.

[73] V. Jha, "Discover Extreme Gardient Boosting(XGBoost) for Applied Machine Learning," 2017.

[74] C. Kraetzer, Y. Q. Shi, J. Dittmann, and H. J. Kim, *Digital Forensics and Watermarking: 16th International Workshop , IWDW 2017, Magdeburg, Germany, August 23-25, 2017, Proceedings*: Springer International Publishing, 2017.

[75] S. A. Mokeddem, "A fuzzy classification model for myocardial infarction risk assessment," *Applied Intelligence,* vol. 48, pp. 1233-1250, 2018.

[76] R. Dhanaseelan and M. J. Sutha, "Diagnosis of coronary artery disease using an efficient hash table based closed frequent itemsets mining," *Medical & biological engineering & computing,* pp. 1-11, 2017.

[77] D. Vadicherla and S. Sonawane, "Classification Of Heart Disease Using Svm And Ann," *Ijrcct,* vol. 2, pp. 693-701, 2013.

[78] D. Chutia, D. K. Bhattacharyya, J. Sarma, and P. N. L. Raju, "An effective ensemble classification framework using random forests and a correlation based feature selection technique," *Transactions in GIS,* vol. 21, pp. 1165-1178, 2017.

[79] N. Priyanka and P. RaviKumar, "Usage of data mining techniques in predicting the heart diseases—Naïve Bayes & decision tree," in *Circuit, Power and Computing Technologies (ICCPCT), 2017 International Conference on*, 2017, pp. 1-7.

[80] C. Kalaiselvi, "Diagnosing of heart diseases using average k-nearest neighbor algorithm of data mining," in *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*, 2016, pp. 3099-3103.

[81] K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, and S. Mukherjee, "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques," *CSI transactions on ICT*, vol. 4, pp. 313-319, 2016.

APPENDIX

PYTHON CODES

```
import numpy as np
from sklearn.cross_validation import train_test_split
from sklearn import metrics
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
import plotly.plotly as py
import plotly.tools as tls
import plotly.offline as offline
import scipy
import seaborn as sb
import pandas as pd
import plotly.graph_objs as go
from sklearn import preprocessing
from sklearn.metrics import classification_report,confusion_matrix
import seaborn as sns # data visualization library
from subprocess import check_output
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import BernoulliNB
from sklearn.naive_bayes import GaussianNB
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import AdaBoostClassifier
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.metrics import f1_score,confusion_matrix
from sklearn.metrics import accuracy_score
dataset= pd.read_csv('the path of the datasets')
```

```
x =dataset.iloc[:,0:13]

y =dataset.iloc[:,-1]

X_train, X_test, y_train, y_test = train_test_split(features,labels, test_size=0.33,
random_state= 5)

#random forest classifier with n_estimators=10 (default)

clf_rf = RandomForestClassifier(random_state=5)

clr_rf = clf_rf.fit(X_train,y_train)

ac = accuracy_score(y_test,clf_rf.predict(X_test))

print('Accuracy is: ',ac)

cm = confusion_matrix(y_test,clf_rf.predict(X_test))

sns.heatmap(cm,annot=True,fmt="d")

y_pred=clf_rf.predict(X_test)

print(classification_report(y_test,y_pred))

# naïve bayes classifier gaussian technique

GaussianNB=GaussianNB()

scaler=preprocessing.MinMaxScaler()

scaler.fit(x_train)

x_train=scaler.transform(x_train)

x_test=scaler.transform(x_test)

x_train,x_test,y_train,y_test = train_test_split(X,y,test_size= 0.33, random_state =11)

y_expect=y_test

GaussianNB.fit(x_train,y_train)

print(GaussianNB)

y_pred=GaussianNB.predict(x_test)

print(accuracy_score(y_expect,y_pred))

print(metrics.classification_report(y_expect,y_pred))

# AdaBoostClassifier

X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.33,
random_state= 9)

# Instantiate

abc = AdaBoostClassifier()
```

```
# Fit
abc.fit(X_train, y_train)
# Predict
y_pred = abc.predict(X_test)
# Accuracy
accuracy_score(y_pred, y_test)
print(confusion_matrix(y_test,y_pred))
print('\n')
print(classification_report(y_test,y_pred))
# LogisticRegression classifier
X_train, X_test, y_train, y_test = train_test_split(features,labels, test_size=0.33, random_state= 15)
# all parameters not specified are set to their defaults
logisticRegr = LogisticRegression()
logisticRegr.fit(X_train, y_train)
y_pred = logisticRegr.predict(X_test)
score = logisticRegr.score(X_test, y_test)
print(score)
print(classification_report(y_test,y_pred))
# XGBClassifier
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size = 0.33, random_state = 15)
classifier = XGBClassifier()
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
cm = confusion_matrix(y_test, y_pred)
ac = accuracy_score(y_test,y_pred)
print('Accuracy is: ',ac)
print(classification_report(y_test,y_pred))
```