HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

A PROJECT REPORT

for

DATA MINING TECHNIQUES (ITE2006)

in

B.Tech – Information Technology and Engineering

by

PORITOSH BARDHAN (19BIT0079)

AZEEM ULLAH KHAN (19BIT0131)

PRANCHAL SIHARE (19BIT0144)

Under the Guidance of

Dr. SENTHILKUMAR N C

Associate Professor, SITE

School of Information Technology and Engineering

June, 2021

DECLARATION BY THE CANDIDATE

We hereby declare that the project report entitled **"HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES"** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide project work carried out by us under the guidance of **Dr. Senthilkumar N C.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore Signature

Date : 21st May 2021

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology & Engineering [SITE]

CERTIFICATE

This is to certify that the project report entitled **"HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES"** submitted by **Pranchal Sihare (19BIT0144)** to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Data Mining Techniques (ITE2006)** is a record of bonafide work carried out by them under my guidance.

Dr. Senthilkumar N C

GUIDE

Associate Professor, SITE

HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

Abstract

According to various studies of causes of death all across the world, heart diseases are found to be one of the most common causes of death among humans. Quoting from WHO, around 31% of deaths in the world are caused by cardiovascular diseases and more than 75% of deaths occur in developing countries. It is a topic which needs immediate attention due to its impact on human health. The diagnosis of the heart diseases is a very important and is itself the most complicated task in the medical field. All the factors are to be taken into consideration when analysing and understanding the patients by the doctor through manual check-ups at regular intervals of time. This is where data mining techniques can be used for predicting heart diseases at an early stage which can save many lives. Data mining in simple words means taking out and refining of useful information from an enormous amount of data. It is a basic process in defining and discovering useful information and hidden patterns from different databases. We have examined various research papers in this topic to find out how different data mining classification techniques are being used for predicting heart ailments. Based on our study we will try to find out the most accurate classification technique which is currently being used for predicting heart diseases, find out its laggings and suggest improvements to it.

Keyword – WHO, Heart Disease, Data Mining, Classification, Accuracy.

I. INTRODUCTION

In day-to-day life many factors that affect a human heart. Many problems are occurring at a rapid pace and new heart diseases are rapidly being identified. In today's world of stress Heart, being an essential organ in a human body which pumps blood through the body for the blood circulation is essential and its health is to be conserved for a healthy living. The health of a human heart is based on the experiences in a person's life and is completely dependent on professional and personal behaviors of a person. There may also be several genetic factors through which a type of heart disease is passed down from generations. According to the World Health Organization, every year more than 12 million deaths are occurring worldwide due to the various types of heart diseases which is also known by the term cardiovascular disease. The term Heart disease includes many diseases that are diverse and specifically affect the heart and the arteries of a human being. Even young aged people around their 20-30 years of lifespan are getting affected by heart diseases. The increase in the possibility of heart disease among young may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behavior, family history, smoking and hypertension.

In this research work, the supervised machine learning concept is utilized for making the predictions. A comparative analysis of the three data mining classification algorithms namely **Naïve Bayes and KNN algorithm** are used to make predictions. The Stat Log dataset from UCI machine learning repository is utilized for making heart disease predictions in this research work. The predictions are made using the classification model that is built from the classification algorithms when the heart disease dataset is used for training.

The model aims to be researched and advance in further to become robust and end to end reliable research tool. We will discuss about the classical methods and algorithms implemented on CVD prediction, gradual advancements, draw comparison of performance among the existing systems and propose an enhanced multi-module system performing better in terms of accuracy and feasibility. Implementation, training and testing of the modules have been done on datasets obtained from UCI and Physio net data repositories. Data format have been modified in case of the ECG report data for betterment of action by the convolutional neural network used in our research and in the risk prediction module we have chosen attributes for training and implementing the multi-layered neural network developed by us.

II. BACKGROUND

A risk of a heart attack or the possibility of the heart disease if identified early, can help the patients take precautions and take regulatory measures. Recently, the healthcare industry has been generating huge amounts of data about patients and their disease diagnosis reports are being especially taken for the prediction of heart attacks worldwide. When the data about heart disease is huge, the machine learning techniques can be implemented for the analysis.

Data Mining is a task of extracting the vital decision-making information from a collective of past records for future analysis or prediction. The information may be hidden and is not identifiable without the use of data mining. The classification is one data mining technique through which the future outcome or predictions can be made based on the historical data that is available. The medical data mining made a possible solution to integrate the classification techniques and provide computerized training on the dataset that further leads to exploring the hidden patterns in the medical data sets which is used for the prediction of the patient's future state. Thus, by using medical data mining it is possible to provide insights on a patient's history and is able to provide clinical support through the analysis. For clinical analysis of the patients, these patterns are very much essential. In simple English, the medical data mining uses classification algorithms that is a vital part for identifying the possibility of heart attack before the occurrence. The classification algorithms can be trained and tested to make the predictions that determine the person's nature of being affected by heart disease.

The two classification techniques used in our project are **Naïve Bayes** and **KNN algorithms.**

Naïve Bayes

Bayes' Theorem is stated as:

$$P(Ci|X) = (P(X|Ci) * P(Ci)) / P(X)$$

Informally, this can be written as

posteriori = likelihood x prior/evidence

Here, Ci indicates the different class labels target attribute can take.

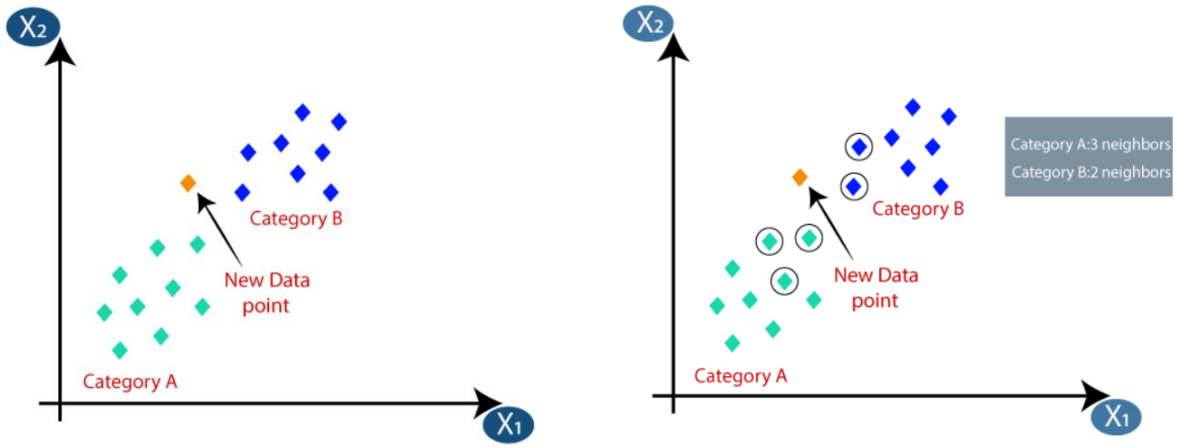
X is the set of vectors having predictor attributes. X=(x1,x2,…,xn)

For final classification we derive the maximum posteriori, i.e., the maximal P(Ci|X)

K-Nearest Neighbor

KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. Firstly the Euclidian distance is computed between the new data and all the existing data across all categories. It then selects k number of data points which have the least Euclidian distance values. The new data point is thus classified in the category which has the most number of data points.

For e.g. – In the below example if we have k value as 5 the since 3 points are there in A and 2 are there in B so we classify the new data in category A



III. LITERATURE REVIEW

Numerous research papers were examined and their observations are as follows.

[1]. In 2018, Kanika Khera and Dr. Neelam Duhan conducted a research on Prediction of heart disease using data mining techniques. There are different data mining techniques for classification. Performance analysis on different classification algorithms such as Decision tree, Naïve Bayes (NB), K-Nearest Neighbour (KNN), and Neural Networks (NN) are carried out. Accuracy of Naïve Bayes is 85% and SVM is 82%.

[2]. In May 2018 R. Sharmila and S. Chemmamal conducted a research on creation of conceptual model to enhance the prediction of heart diseases using big data and data mining techniques. Based on some existing literature related to the prediction of heart diseases using data mining techniques they have drawn inferences about which attributes to consider and which models to be used. They have carries out a survey to find applicability of classification techniques such as Decision Tree, Naive Bayes algorithm, Neural Network and Support Vector Machine (SVM) for Heart disease prediction. Based on this they concluded that six attributes are sufficient for predicting heart diseases correctly and SVM gives the best accuracy and efficiency. It has been proposed to use big data tools such as Hadoop Distributed File System (HDFS) along with SVM for prediction of heart disease with optimized attribute set. It has been proposed to use data set UCI repository as it is the bench mark dataset. Accuracy of SVM is 85%.

[3]. In November 2018, H. Benjamin Fredrick David and S. Antony Belcy developed a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using three classification algorithms namely Naïve Bayes, Decision Tree, and Random Forest are used at different levels of evaluations. The database for this research work has been taken from the StatLog dataset in UCI repository. It includes 13 attributes. The heart disease dataset included in this research work consists of total 270 instances with no missing values. For the validation of the results, several ranges of experiments are carried out using Cross validation and Percentage split methods. Finally, the results show that the Random Forest is best suited for the prediction of heart disease, then the Decision Tree and Naïve Bayes classification algorithms. Accuracy of Naïve Bayes is 80.45% and Decision Tree is 77.4%.

[4]. In November 2018, Chaithra N and Madhu B worked on various data mining techniques introduced in recent years to design a predictive model for cardiovascular diseases from the data obtained by transthoracic echocardiography. A total of 336 records with 24 attributes were highly relevant in predicting heart disease from echocardiography dataset were analysed by applying techniques prospectively. This study investigates three different classification models: J48 Decision Tree, Naive Bayes and Neural Network on cardiovascular disease prediction and the same has been justified with the results of different experiments conducted and the performance of the models was evaluated using the standard metrics of Accuracy, Precision, Recall and F-measure. The experimental results have shown that Neural Network outperformed J48 Decision tree and Naïve Bayes in the domain of predicting heart diseases cases. The results of all the three algorithms performed best in true negative rate which makes it a handy tool to train medical students and junior cardiologists to diagnose patients with heart disease. Accuracy of Naïve Bayes is 79.9%, Neural Networks is 97.9% and Decision Tree is 94.1%

[5]. In 2018 Sarangam Kodati and Dr R. Vivekanandam carried out a research on Analysis of Heart Disease using in Data Mining Tools Orange and Weka. Data mining methods like Naïve Bayes, SVM and KNN algorithm have been considered for the diagnosis of heart disease patients. This paper analysed few parameters to predicts heart diseases and suggested a heart diseases prediction system (HDPS) based total on the data mining approaches. 14 attributes and 303 instances were considered in the dataset for heart disease prediction. Precision and recall values were calculated using different tools for each methodology. WEKA and ORANGE tools were used which produced the best results. Accuracy of Naïve Bayes is 83.1%, SVM is 82.8% and KNN is 66.6%.

[6]. In 2018 Siddharth Joshi, Ashish Sasanapuri, Shreyash Anand, Saurav Nandi and Varsha Nemade conducted research on predictive analysis using data mining techniques for heart disease diagnosis. The paper uses Naïve bayes approach for the prediction. It has demonstrated the use of Data Mining algorithm by using python programming language for creating a desktop application for prediction of heart diseases. They have used 14 attributes to predict data and compare it across four databases namely Cleveland, Hungarian, Switzerland and Long Beach VA consisting of 303, 294, 123 and 200 records respectively. It has claimed to have achieved better accuracy in all the databases as compared to existing studies. Accuracy of Naïve Bayes is 82%.

[7]. In May 2018 Uma N Dulhare conducted a research on Prediction system for heart disease using Naive Bayes and particle swarm optimization. The Naïve Bayes algorithm is relatively stable with respect to small variation or changes in training data. It aims at improving the accuracy of prediction of Naïve Bayes algorithm using particle swarm optimization. The dataset consists of 14 attributes and 270 instances for prediction. The results showed that the proposed model improved the performance of existing Naïve Bayes approach by 8%. The system also removed irrelevant features which further enhanced the performance. Accuracy of Naïve Bayes is 87.9%.

[8]. In 2018, Mr. Chala Beyene and Prof. Pooja Kamat came up with a methodology whose main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within retrieve result in short time. This system uses data mining techniques and machine learning algorithms J48, Naïve Bayes and Support Vector Machine, with k-fold cross-validation to predict the occurrence of heart disease. It uses different medical attributes that are more relevant such age, sex, blood pressure, cholesterol, blood sugar and heart rate are some of the attributes are included to identify if the person has heart disease or not. Analyses of data set are computing (implementing) using WEKA software. StatLog heart disease dataset taken from UCI machine learning laboratory for this research paper. This database contains 13 attributes. Accuracy of Naïve Bayes is 87.7%, Decision Tree is 86% and SVM of 83.8%.

[9]. In 2018, Charu V. Verma and Dr. S. M. Ghosh developed a model to predict the risk of heart disease in diabetic patients. In this research paper we are applying Naive Bayes data mining classification technique which is a probabilistic classifier based on Bayes theorem with strong (naive) independence assumptions between the features. The patient data set is compiled from UCI data repositories as combined data from Statlog data set and Cleveland Clinic Foundation for heart patients. Here we are taken 14 attributes. From the system we get confusion matrix from which we can predict accuracy of the applied Naïve Bayes algorithm. The result shows that accuracy of our applied algorithm is 89.41% in the prediction of risk of heart disease in diabetic patient. As a future work, further data analysis has been planned to perform other data mining algorithms to improve the classification accuracy. Accuracy of Naïve Bayes is 89.41%

[10]. In 2018, thie study by Le Minh Hung, Tran Dinh Toan and Tran Van Lang we have conducted 2 experiments to investigate the performance of HD prediction using different

classification and feature selection methods. Although the HD dataset can be considered as linear separable, a hard-margin SVM hardly separates the two classes. Soft-margin kernel SVM is selected as the classifier to compare the effectiveness of 4 different feature selection methods including ILFS, CFS, LLCFS and PCA. Our experiments results show that PCA could generate a competitive result when the number of PCs used is less than 31 while CFS and LLCFS perform well with over 31 attributes. ILFS generates the best performance and maintains stable when the number of attributes used is over 31. Accuracy of Naïve Bayes is 77.7% and SVM is 81.4%.

[11]. In 2018 Md. Fazle Rabbi, Md. Palash Uddin, Md. Arshad Alithis paper, the Cleveland standard heart disease dataset is gathered from the UCI machine learning repository. Although there are a total 270 records of 76 different attributes along with the true sample label in the dataset, most of the published experiments has referred to using a subset of attributes. The used 13 attributes with respective explanation. In this experiment, approximately half of the data are used for training and the rest is for the testing. Accuracy of Neural Networks is 73.3% and SVM is 85.2%.

[12]. In 2018, Noreen Akhtar, Muhammad Ramzan Talib, Nosheen Kanwal studied the data mining is the region that reviews which implies that data and knowledge are helpful from past information. There are various strategies for information mining. Data mining can be utilized as a part of various regions including medical utilize. Khemphila and Boonjing explained that given meaning tree “which can be used to divide a large number of structures through over-application of simple sequence records gathered to decrease continuously record set decision-making rules. The KDD procedure demonstrate embraced in this examination along these lines as indicated by Hanand Kamber, sub-class is to locate a work of art (or process reason) depict and recognize information projects or thoughts keep in mind that end goal to foresee motivation behind the question class of the model can be utilized Its class tag is unidentified. Accuracy of Naïve Bayes is 79.91% and Decision Tree is 77.1%.

[13]. In this study by Poornima Singh, Sanjay Singh, Gayatri S Pandijain in 2018, an EHDPS has been presented using data mining techniques. From ANN, an MLPNN together with BP algorithm is used to develop the system. The MLPNN model proves the better results and assists the domain experts and even the person related to the medical field to plan for a better and early diagnosis for the patient. This system performs realistically well even without retraining. Furthermore, the experimental results show that the system predicts heart disease

with ~100% accuracy by using neural networks. The experiment was carried out on a publicly available database for heart disease. The dataset contains a total of 303 records that were divided into two sets, training set (40%) and testing set (60%). A data mining tool named Weka 3.6.11 was used for the experiment. Additionally, multilayer perceptron neural network (MLPNN) with backpropagation (BP) was used as the training algorithm. Accuracy of Neural Networks is 77.3%

[14]. In 2018 Navdeep Singh, Sonika Jindal proposed a system that uses a neural network for prediction of cardiovascular disease, blood pressure, and sugar. a collection of seventy-eight records with thirteen attributes are used for training and testing. He urged supervised network for diagnosing of cardiovascular disease and trained it using back propagation formula. On the idea of unknown information is entered by a doctor the system can notice that unknown information from training data and generate a list of possible illness from that patient will suffer. Accuracy of Naïve Bayes is 97.1%

[15]. In January 2019 Wan Muhamad Taufik Wan Ahmad, Nur Laila Ab Ghani, and Sulfeeza Mohammad Drus carried out a research on Data Mining Techniques for Disease Risk Prediction Model. This review focuses on decision tree, neural network and support vector machines where heart-related disease is commonly studied. It studies different research papers which use the UCI database and different number of attributes out of the 76 attributes. The paper finds out that artificial neural networks are most frequently for heart related diseases prediction and also have the highest accuracy. Accuracy of Neural Networks is 86%, Decision Tree is 53.3%, SVM is 67.7%

[16]. In January 2019, Ajit Solanki, Mehul P. Barot provided insights on heart disease risk diagnosis using various classification techniques, such as support vector machine (SVM), decision tree, k-mean, naive Bayes and MLP. In this paper different kind of classification methods which are applied in the forecasting of heart disease has been talked and also comparison made on the classifiers to justify which algorithm achieve the high accuracy. Decision Tree performs poor with large datasets. Naïve Bayes showed highest accuracy of 81.25% for 14 attribute dataset of heart disease and SVM showed the accuracy of 83% when applied on dataset of National Health and Nutrition Exam Survey. From this study they concluded that Multilayer Perceptron (MLP) achieved the highest accuracy, but drawback of MLP is that it is very slow in performance and it can only be applied for linear data set. Accuracy of Naïve Bayes is 81.2%, SVM is 82.5% and KNN is 87%

[17]. In July 2019 S. Anitha and N. Shridevi conducted a research on heart disease prediction using data mining techniques. In this work, supervised machine learning algorithms namely SVM, KNN and Naïve Bayes were used to predict the heart diseases. The machine learning algorithms were implemented using R programming language. The dataset consisted of 14 attributes out of the original 76 and 302 records were considered for the research. For analysis a confusion matrix, sometimes also called error matrix was used. From the experimental results the concluded that Naïve Bayes algorithm predicts the heart disease with the highest accuracy among the three methods. Accuracy of Naïve Bayes is 86.6%, SVM is 77.7% and KNN is 76.7%

[18]. In April 2019 Dr Yasemin Gultepe and Sabah Rashed carried out a research on The Use of Data Mining Techniques in Heart Disease Prediction. In this article they have used Weka software as one of Data Mining techniques in heart disease prediction by testing a dataset obtained from UCI repository. The classification technique used is Naïve Bayes algorithm for the prediction of heart diseases. The dataset consists of 14 attributes and 303 instances. The results were calculated over many iterations. The research concluded that Naïve bayes approach can be further improved by newer methods. Accuracy of Naïve Bayes is 85.7%.

[19]. In April 2019, Adil Hussain Seh, Dr. Pawan Kumar Chaurasia researched on different types of data mining techniques and machine learning techniques such as Classification, Association, Clustering, Decision Tree, Naive Bayes, Artificial Neural Networks, Genetic Algorithm and Cross Validation. The main objective of this research paper is to summarise the recent research with comparative results that has been done on heart disease prediction and also make analytical conclusions. The dataset consists of 303 records with 14 essential attributes (total attributes 75) with some missing values also. Accuracy of research is directly proportional to the selection of research tools and procedures. So, Choice of appropriate experimental tool (WEKA, METLAB etc.) for implementation of techniques is also an important parameter. From the study, it is observed Naïve Bayes with Genetic algorithm; Decision Trees and Artificial Neural Networks techniques improve the accuracy of the heart disease prediction system in different scenarios. Accuracy of Naïve Bayes is 81.5% and Neural Networks is 89%

[20]. In June 2019, Monther Tarawneh and Ossama Embarak conducted a two-phase experiment to understand how machine learning techniques can help in comprehending the level of risk associated with heart disease using information gain and gain ratio feature

selection techniques. The proposed Method contains three phases. Starts with pre-processing phase where data filtered and classified before any processing. The output of this phase goes into number of classification techniques where these techniques evaluated tom eliminate low performance one. Then we combine the result and look at the patient history to give a decision (negative/ positive) of heart attack. Accuracy of Naïve Bayes is 89.2%.

[21]. In August 2019, Akansha Jain, Manish Ahirwar, Rajeev Pandey provided overview on intuitive prediction of heart Disease using various data mining techniques. The aims of the researchers were, to find the best classification algorithm applied for providing best accuracy whether person is suffered with heart disease or not using Decision Tree and Naive Bayes. Discovery of unfamiliar patterns in cardiovascular diseases, easily available algorithms of classification are referred to dataset and their accuracies are compared. They analysed the accuracy of each algorithm by taking different number of attributes and also different datasets. After evaluating various surveys and research papers, picking diverse data mining strategies and actualising them on the chose dataset, highest accuracy results in SVM technique. The performance of Naive Bayes shows high level compare with other classifiers except SVM technique. Accuracy of Naïve Bayes is 83.7%, Decision Tree is 79% and SVM is 84.4%.

[22]. In July 2019 Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava conducted research on effective heart disease prediction using hybrid machine learning techniques. In this paper a novel method has been proposed that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. It analyses Naïve Bayes, Decision tree and SVM among other methods for heart disease prediction. It considers 13 attributes and 297 records as the dataset for the research. It proposes a Hybrid Random Forest with Linear Model using all the different models like decision tree, SVM and KNN for different tasks. It achieves a better accuracy than any of the individual method, that of 88.4%. Accuracy of Naïve Bayes is 75.8%, Decision Tree is 85% and SVM is 86.1%

[23]. In August 2019 T. R. Stella Mary and Shoney Sebastian carried out a research on Predicting heart ailment in patients with varying number of features using data mining techniques. The research uses Naïve Bayes algorithm among other techniques for predicting heart diseases. Also, as a part of this research the prediction is carried out using three different set of attributes. First set comprises of 7 attributes and 300 records, second set is of

10 attributes and 290 records and the third set has 14 attributes and uses 303 records. They concluded that with the increase in number of attributes the accuracy increases irrespective of the data mining technique used. Also, for different dataset different techniques give different results, so it cannot be concluded which technique is better. Accuracy of Naïve Bayes is 81.4%.

[24]. In December 2019, GS Mallikarjuna Rao, K Anitha conducted a research on Usage of Data Mining Techniques in Predicting the Heart Diseases Decision Tree & Random Forest Algorithm. In the Proposed system, we are using Decision tree and Random forest algorithms to predict heart diseases. As the Naïve Bayes classifier requires a small dataset to predict there will be a loss of accuracy which is the disadvantage of the naïve Bayes classifier. Accuracy of Decision Tree is 98.1%

[25]. In 2019, Haleh Ayatollahi, Leila Gholamhosseini and Masoud Salehi aimed at developing a model using data mining algorithms to predict coronary artery diseases. Therefore, the present study aimed to compare the positive predictive value (PPV) of CAD using artificial neural network (ANN) and SVM algorithms and their distinction in terms of predicting CAD in the selected hospitals. The research sample was the medical records of the patients with coronary artery disease who were hospitalised in three hospitals affiliated to AJA University of Medical Sciences between March 2016 and March 2017 (n = 1324). Totally, 25 variables affecting CAD were selected and related data were extracted. According to the results, the SVM algorithm presented higher accuracy and better performance than the ANN model and was characterised with higher power and sensitivity. Overall, it provided a better classification for the prediction of CAD. The use of other data mining algorithms is suggested to improve the positive predictive value of the disease prediction. Accuracy of Neural Networks is 88% and SVM is 92.3%

[26]. In May 2020, Meenu Shukla, Sandeep Kumar, Rishabh Sharma, Saurabh Sharma and Rishabh Tyagi conducted a research on Heart disease prediction using data mining techniques. Frequent research related to disease forecast model has been performed using numerous mining methods and learning algorithms in health centres. Suggested cardiovascular disease Forecasting using linear regression, and K shows that multiple regression analysis is appropriate for forecasting risk of cardiac attack. The analysis is done by means of planning data collection consisting of 3000 instances with 13 specific attributes previously specified. Consequently, efficiency of the precision is enhanced further to offer

better diagnostic disease judgment. S. Seema et al, focus to strategies which can forecast chronic illness with the help of Support Vector Machine, Artificial Neural Network, Decision Tree and naive Bayes, extracting the data found in historic health documents. Accuracy of Naïve Bayes is 82.2%, Decision Tree is 75.4% and SVM is 88.5%

[27]. In August 2020, Basma Jumaa Saleh, Ahmed Yousif Falih Saedi, Ali Talib Qasim al Aqbi, Lamees abdalhasan Salman aimed at predicting future heart problems successfully from the medical data collection. A model has been established using a prediction technique to assess the cardiac disease's features by certain features, Waikato Environment for Knowledge Study (WEKA) was used for prediction. Weka uses classification, and simulation algorithms and also the clusters on the re-processed datasets to provide a visualisation of the trend. Only by taking into account characteristics is it possible to predict heart disease. This approach can be accomplished by integrating data extraction methods. Data mining method includes K-star, J48(Decision Tree), SMO, Naïve Bayes, MLP, Bayes Net REPTREE, etc. Accuracy of Naïve Bayes is 85.4% and Decision Tree is 86.3%

[28]. In August 2020, Edy Irwansyah, Ebiet Salim Pratama and Margaretha Ohyver proposed a various stages clustering model to predict cardiovascular disease in patients. This study uses secondary data obtained from a private hospital in Jakarta. The data obtained are 644 observations. This study uses age variables as well as 8 variables of the results of blood tests of patients with cardiovascular disease. Data reduction technique with PCA from eight variable data of blood test of patients with cardiovascular disease and age variables obtained from 644 observations, can produce new five components with adequate data diversity. The combination of data reduction techniques with PCA and the application of the K-Menoids clustering algorithm is a new way of data mining to group data of patients with cardiovascular disease to see the level of patient complications in each different data cluster. There is still a lack of cluster evaluation results shown by the value of the Silhouette coefficient (SC) which has a weak structure with a value of 0.35 therefore it is necessary to develop further research methods to produce clusters with stronger structures. Accuracy of Clustering is 83.4%

[29]. In September 2020, Najmu Nissa, Sanjay Jamwal, Shahid Mohammad conducted Copious work for heart Disease prediction using Machine Learning, Deep Learning, Data mining tools and techniques. Different Datasets, Algorithms and methods used by the researchers and observed results alongside the future work is carried out in finding efficient

methods of medical diagnosis for Cardiovascular disease. Accuracy of Naïve Bayes is 83.8%, Neural Networks is 81.7%, Decision Tree is 94.2% and SVM is 87.1%

[30]. In February 2021 Pratiksha Shetgaonkar and Dr Shailendra Aswale conducted a research on the use of three AI-based methods namely Decision Tree, Naïve Bayes, & Neural Network for forecasting cardiovascular or heart disease. All of these methods have been evaluated based on different unique & parameters with optimizations for better accuracy. They have analysed many methodologies and have decided to use these three methodologies for the research. Their datasets consisted of 14 attributes with 668 records. From their research they found that on increasing hidden layers, the result becomes less accurate and it also consumes more time. They also found that upon changing the attributes the results changed. They also suggested that the system can be improvised by using a hybrid approach rather than individual methods. Accuracy of Naïve Bayes is 85%, Neural Networks is 81.8%, and Decision Tree is 98.5%.

IV. DATASET DESCRIPTION & SAMPLE DATA

The dataset that we will be using being used in the process contains the following 14 features including 1 class label and 270 instances with no missing values data collected from UCI repository which is balanced. Here presence or absence of heart disease is predicted on the basis of age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina and old peak etc.

The following attributes are used

Attribute	Name	Description
1	Age	Age in years
2	Sex	Male or female
3	Trestbps	Resting blood pressure (in mmHg on admission to the hospital)
4	Cp	Chest pain type (typical angina, atypical angina, non-anginal pain, asymptomatic)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl (true, false)
7	Restecg	Resting electrocardiographic results (normal, abnormal, LVH)
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina (yes, no)
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment (unsloping, flat, downsloping)
12	Ca	Number of major vessels (0-3) coloured by fluoroscopy
13	Thal	Normal, fixed defect, reversible
14	Num	Diagnosis of heart disease (yes, no)

* - http://archive.ics.uci.edu/ml/datasets/heart+disease

For our code the attribute values will be in the following ranges

- age: The person's age in years
- sex: The person's sex (1 = male, 0 = female)

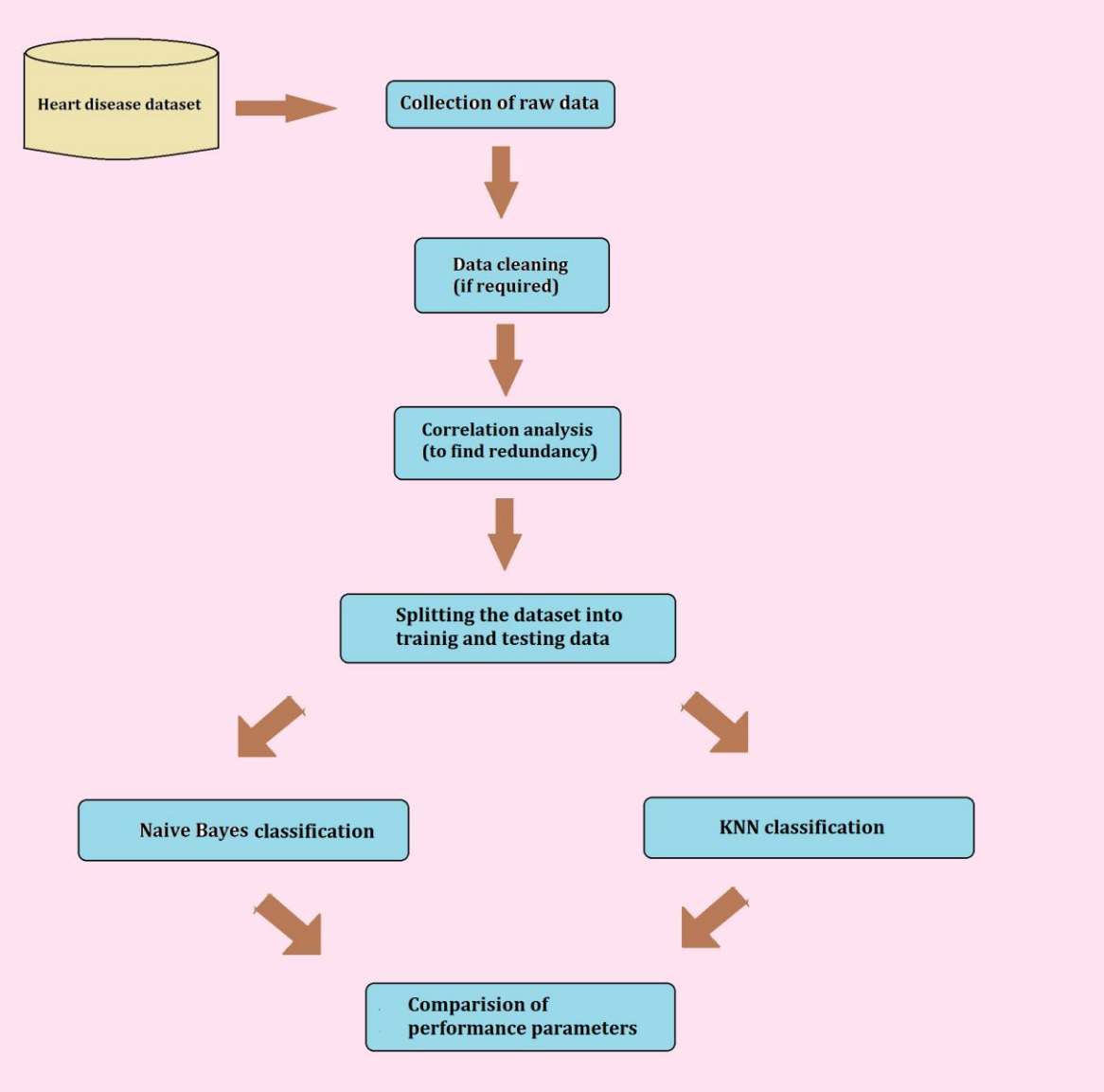
- chest_pain_type: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- resting_blood_pressure: The person's resting blood pressure (mm Hg on admission to the hospital)
- cholestrol: The person's cholesterol measurement in mg/dl
- fasting_blood_sugar: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
- rest_ecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- max_heart_rate_achieved: The person's maximum heart rate achieved
- exercise_induced_angina: Exercise induced angina (1 = yes; 0 = no)
- st_depression: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
- st_slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
- num_major_vessels: The number of major vessels (0-3)
- thalassemia: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)
- target: Heart disease (0 = no, 1 = yes)

The sample data is as follows: -

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

V. PROPOSED ALGORITHM WITH FLOWCHART

1. Import the dataset and check the type of the dataset.
2. Check for missing values in the dataset, if found then apply data cleaning by filling in the most probable value.
3. Check for redundancy in data among attributes using correlation analysis.
4. Apply the two classification techniques, namely, Naïve Bayes and KNN.
   4.1. Create the training model for the classifiers.
   4.2. Fit the training dataset into the model.
   4.3. Using the model predict the values of testing dataset and generate accuracy scores.
5. Now generate the bar-charts for comparing accuracies of both models.



VI. EXPERIMENTS RESULTS

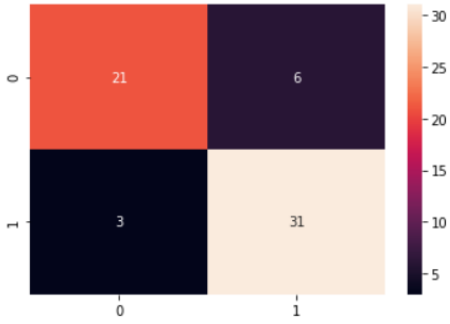
Accuracy

It is the ratio of correctly predicted target values to total number of target values.

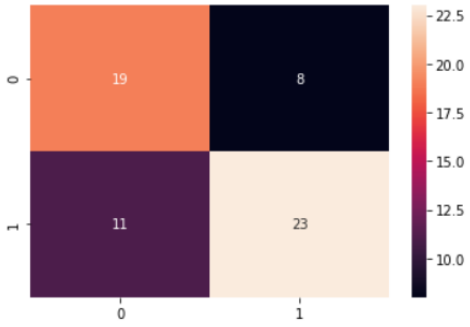
The accuracy of Naïve Bayes is **85.25%** and of KNN is **68.85%** (for optimal value of k=8)**.**

Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data. Each row of the matrix represents the instances in the actual target class while each column represents the instances in the predicted target class.



Naïve Bayes



K Nearest Neighbor

Precision Score

Precision score is the ratio of correctly predicted positive target value to the total predicted positive target values.

The precision scores of Naïve Bayes is **0.8378** and of KNN is **0.7419.**

Recall

It is the ratio of correctly predicted positive target value to the all observations in actual class.

The recall values of Naïve Bayes is **0.9117** and of KNN is **0.6764.**

F scores

It is the weighted average of precision and recall. Sometimes itt is better to compare f scores than comparing accuracy in case of uneven class distribution i.e. - false positive and false negative do not have similar cost.

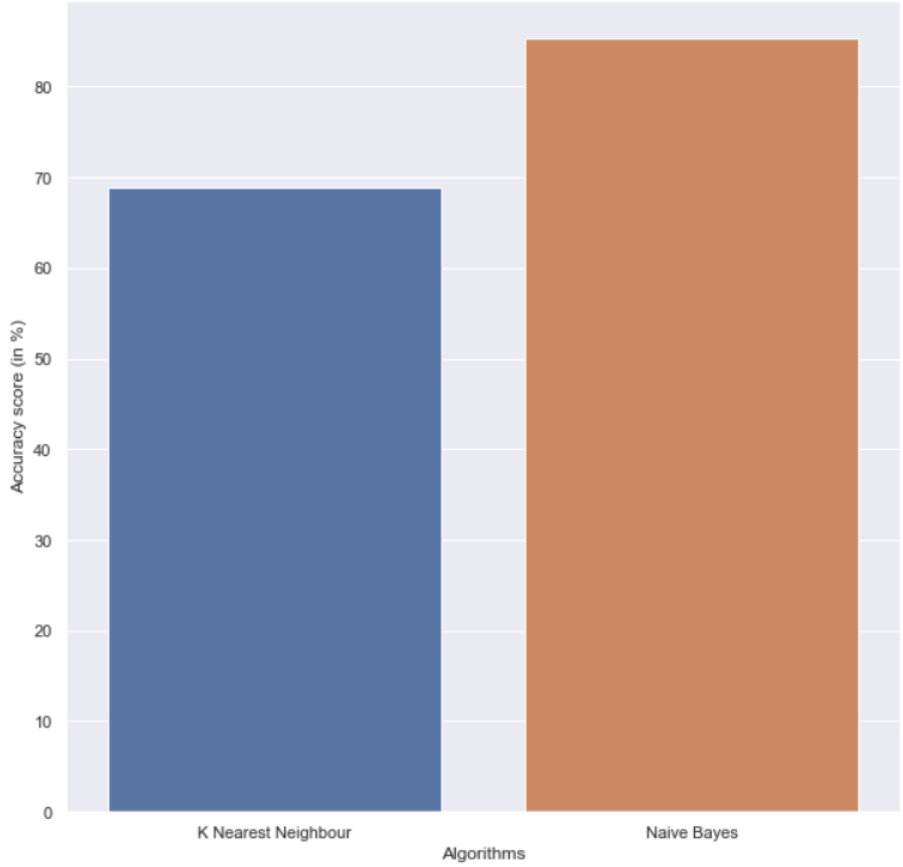
The f score of Naïve Bayes is **0.8732** and of KNN is **0.7076.**

VII. COMPARATIVE STUDY / RESULTS AND DISCUSSION

We get the accuracies for both classification techniques as follows

	accuracy
K Nearest Neighbour	68.85
Naive Bayes	85.25

This is the graphical representation

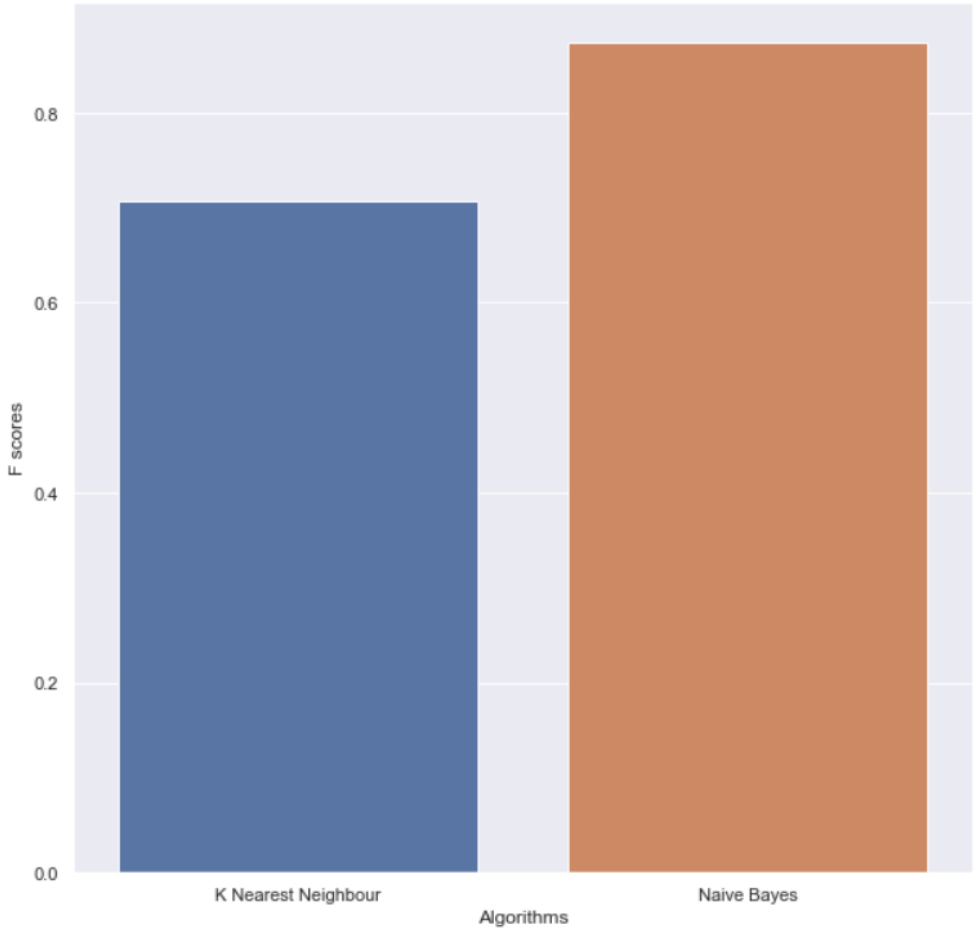


As observed in the graph the accuracy of naïve bayes is better than knn.

The F-scores calculated for both techniques are

	f-scores
K Nearest Neighbour	0.707692
Naive Bayes	0.873239

Its graphical representation looks like



This graph shows that f-score of naïve bayes is better than knn.

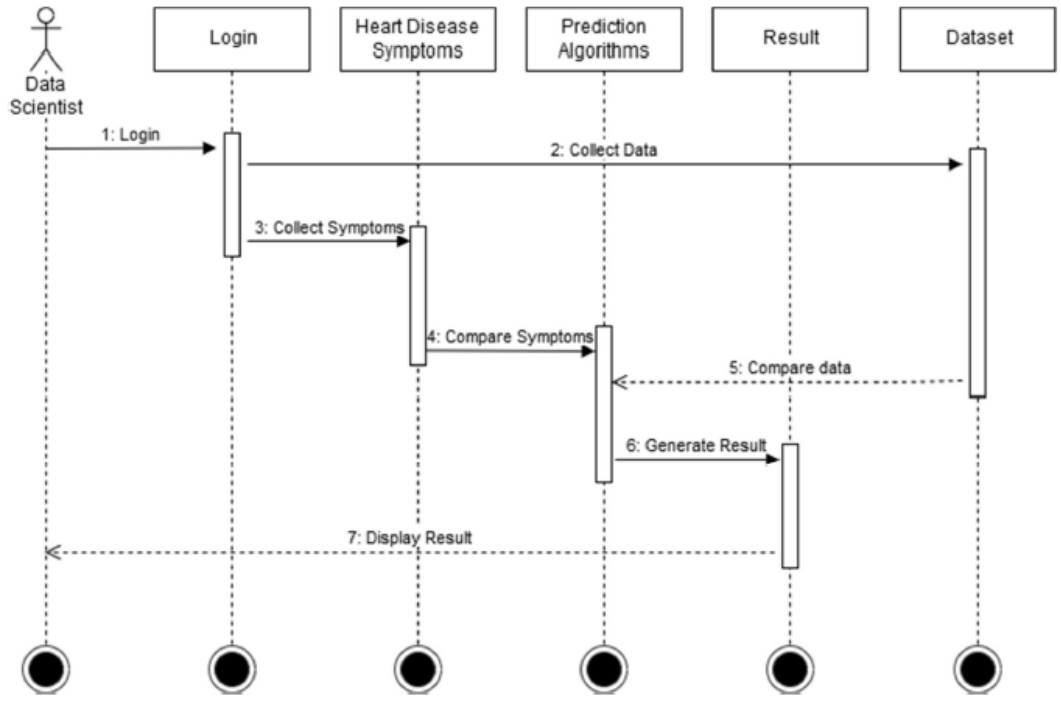
VIII. CONCLUSION AND FUTURE WORK

The overall objective of this work is to predict more exactly the occurrence of heart disease using data mining techniques. In this research work, the UCI data repository is used for classifying testing data using classification technique based on three algorithms such as KNN and Naive Bayes. Then the results of the two algorithms were compared to find out which among the three provides the best results. From the research work, it has been experimentally proven that Naïve Bayes provides better results as compared to KNN.

The Future work of this research work can be made to produce an impact in the accuracy of the Bayesian Classification by applying Particle Swarm Optimization algorithm. A novel algorithm can be developed for maximizing the classification performance and minimizing the number of features.

Our project can be used to develop a web platform to predict the occurrences of disease based on various symptoms. The user can select various symptoms and can find the diseases with their probabilistic figures. This can be used as a training tool for Nurses and Doctors who are freshly introduced in the field related to heart diseases. Moreover, if implemented on a large scale it can be used in medical facilities like hospital, clinics where a patient wouldn't have to wait in long queues for treatment if he is feeling symptoms related to heart disease.

This the flow chart of the proposed system that can be developed in the future scope



IX. REFERENCES

1. Kanika Khera and Dr. Neelam Duhan, *A Study on prediction of Heart Disease using data mining techniques*, International Journal of Creative Research Thoughts (IJCRT), Volume 6, Issue 2 April 2018.
2. R. Sharmila and S. Chemmamal, *A conceptual model to enhance the prediction of heart diseases using big data techniques*, International Journal of Computer Sciences and Engineering (IJCSE), Vol-6, Special Issue 4, May 2018.
3. H. Benjamin Fredrick David and S. Antony Belcy, *Heart disease prediction using data mining technique*, International Journal of Soft Computing (IJSC), November 2018.
4. Chaithra N and Madhu B, *Classification Models on Cardiovascular Disease Prediction using Data Mining Techniques*, J Cardiovasc Dis Diagn, Vol. 6, Issue-6, November 2018.
5. Sarangam Kodati and Dr R. Vivekanandam, *Analysis of Heart Disease using in Data Mining Tools Orange and Weka*, Global Journal of Computer Science and Technology: C Software & Data Engineering, Volume 18, Issue 1, Version 1.0, 2018.
6. Siddharth Joshi, Ashish Sasanapuri, Shreyash Anand, Saurav Nandi and Varsha Nemade, *Predictive Analysis using Data Mining techniques for Heart disease diagnosis*, International Journal of Engineering & Technology (IJET), Volume 7, Version 3.1, 2018.
7. Uma N Dulhare, *Prediction system for heart disease using Naive Bayes and particle swarm optimization*, Biomedical Research 29 (12), 2018.
8. Mr. Chala Beyene and Prof. Pooja Kamat, *Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques*, International Journal of Pure and Applied Mathematics (IJPAM), Vol. 118, No. 8, 2018.
9. Charu V.Verma and Dr. S. M. Ghosh, *Prediction of Heart Disease in Diabetic patients using Naive Bayes Classification Technique*, International Journal of Computer Applications Technology and Research (IJCATR), Vol. 7, Issue-7, 2018.
10. Le Minh Hung, Tran Dinh Toan and Tran Van Lang, *Automatic heart disease prediction using feature selection and data mining technique*, Journal of Computer Science and Cybernetics, V.34, N.1 2018.
11. Md. Fazle Rabbi, Md. Palash Uddin, Md. Arshad Ali, Md. Faruk Kibria, Masud Ibn Afjal and Md. Safiqul Islam and Adiba Mahjabin Nitu, *Performance Evaluation of Data Mining Classification Techniques for Heart Disease Prediction*, American Journal of Engineering Research (AJER), Volume-7, Issue-2 2018.

12. Noreen Akhtar, Muhammad Ramzan Talib and Nosheen Kanwal, *Data Mining Techniques to Construct a Model: Cardiac Diseases*, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, No. 1, 2018.
13. Poornima Singh, Sanjay Singh and Gayatri S pandi-Jain, *Effective heart disease prediction system using data mining techniques*, International Journal of Nanomedicine, 2018.
14. Navdeep Singh and Sonika Jindal, *Heart Disease Prediction System using Hybrid Technique of Data Mining Algorithms*, International Journal of Advance Research, Ideas and Innovations in Technology, Volume 4, Issue 2, 2018.
15. Wan Muhamad Taufik Wan Ahmad, Nur Laila Ab Ghani and Sulfeeza Mohd Drus, *Data Mining Techniques for Disease Risk Prediction Model: A Systematic Literature Review*, Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018), DOI: 10.1007/978-3-319-99007-1_4, January 2019.
16. Ajit Solanki, Mehul P. Barot, *Study of Heart Disease Diagnosis by Comparing Various Classification Algorithms*, International Journal of Engineering and Advanced Technology (IJEAT), Vol. 8, Issue-2S2, January 2019.
17. S. Anitha and N. Shridevi, *Heart disease prediction using data mining techniques*, Journal of Analysis and Computation (JAC), Volume 8, Issue 2, February 2019.
18. Dr Yasemin Gultepe and Sabah Rashed, *The Use of Data Mining Techniques in Heart Disease Prediction*, International Journal of Computer Science and Mobile Computing (IJCSMC), Volume 8, Issue. 4, April 2019.
19. Adil Hussain Seh, Dr. Pawan Chaurasia, *A Review on Heart Disease Prediction Using Machine Learning Techniques*, International Journals of Multidisciplinary Research Academy (IJMRA), Vol. 9 Issue 4, April 2019.
20. Monther Tarawneh and Ossama Embarak, *Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques*, Acta Scientific Nutritional Health, Volume 3, Issue 7, July 2019.
21. Akansha Jain, Manish Ahirwar, Rajeev Pandey, *A Review on Intuitive Prediction of Heart Disease Using Data Mining Techniques*, International Journal of Computer Sciences and Engineering (IJCSE), Vol. 7, Issue-7, July 2019.
22. Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, *Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques*, IEEE Access Digital Object Identifier 10.1109/ACCESS.2019.2923707, July 2019.

23. T. R. Stella Mary and Shoney Sebastian, *Predicting heart ailment in patients with varying number of features using data mining techniques*, International Journal of Electrical and Computer Engineering (IJECE) Volume 9, No 4, August 2019.
24. G S Mallikarjuna Rao and K Anitha, *Usage of Data Mining Techniques in Predicting the Heart Diseases Decision Tree & Random Forest Algorithm,* International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume 9 Issue 2, December 2019.
25. Haleh Ayatollahi, Leila Gholamhosseini and Masoud Salehi, *Predicting coronary artery disease: a comparison between two data mining algorithms*, Ayatollahi et al. BMC Public Health, 2019.
26. Meenu Shukla, Sandeep Kumar, Rishab Sharma, Saurabh Sharma and Rishab Tyagi, *A Review on Heart Disease Prediction using Data Mining Techniques,* Journal of Emerging Technologies and Innovative Research (JETIR), Volume 7, Issue 5, May 2020.
27. Basma Jumaa Saleh, Ahmed Yousif Falih Saedi, Ali Talib Qasim al-Aqbi, Lamees abdalhasan Salman, *A review paper: Analysis of WEKA data mining techniques for heart disease prediction system*, University of Nebraska - Lincoln, Library Philosophy and Practice (e-journal) - 4032, August 2020.
28. Edy Irwansyah, Ebiet Salim Pratama and Margaretha Ohyver, *Clustering of Cardiovascular Disease Patients Using Data Mining Techniques with Principal Component Analysis and K-Medoids*, doi:10.20944/preprints202008.0074.v1, August 2020.
29. Najmu Nissa, Sanjay Jamwal and Shahid Mohammad, *Early Detection of Cardiovascular Disease using Machine learning Techniques an Experimental Study*, International Journal of Recent Technology and Engineering (IJRTE), Volume-9 Issue-3, September 2020.
30. Pratiksha Shetgaonkar and Dr Shailendra Aswale, *Heart Disease Prediction using Data Mining Techniques*, International Journal of Engineering Research & Technology (IJERT), Vol. 10, Issue 02, February-2021.

Appendix

```
import numpy as np
import pandas as pd  #used for operations on dataset
import matplotlib.pyplot as plt  #used for defining graphs
import seaborn as sns  #used for actual plotting of graphs
import os
import warnings  #used to handle warnings if any
warnings.filterwarnings('ignore')

data = pd.read_csv("heart.csv")
data.columns = ['age', 'sex', 'chest_pain_type', 'resting_blood_pressure', 'cholesterol',
'fasting_blood_sugar', 'rest_ecg', 'max_heart_rate_achieved', 'exercise_induced_angina',
'st_depression', 'st_slope', 'num_major_vessels', 'thalassemia', 'target']

data.shape
```

```
Out[4]: (303, 14)
```

data.head() *#takes default value as 5*

Out[5]:

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	rest_ecg	max_heart_rate_achieved	exercise_induced_angina	st_depr
0	63	1	3	145	233	1	0	150	0	2.3
1	37	1	2	130	250	0	1	187	0	3.5
2	41	0	1	130	204	0	0	172	0	1.4
3	56	1	1	120	236	0	1	178	0	0.8
4	57	0	0	120	354	0	1	163	1	0.6

data.describe()

Out[6]:

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	rest_ecg	max_heart_rate_achieved	exercise_inc
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000

data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   age                       303 non-null    int64  
 1   sex                       303 non-null    int64  
 2   chest_pain_type           303 non-null    int64  
 3   resting_blood_pressure    303 non-null    int64  
 4   cholesterol               303 non-null    int64  
 5   fasting_blood_sugar       303 non-null    int64  
 6   rest_ecg                  303 non-null    int64  
 7   max_heart_rate_achieved   303 non-null    int64  
 8   exercise_induced_angina   303 non-null    int64  
 9   st_depression             303 non-null    float64
 10  st_slope                  303 non-null    int64  
 11  num_major_vessels         303 non-null    int64  
 12  thalassemia               303 non-null    int64  
 13  target                    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

data.sample(5)

Out[8]:

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	rest_ecg	max_heart_rate_achieved	exercise_induced_angina	st_de
117	56	1	3	120	193	0	0	162	0	1.9
163	38	1	2	138	175	0	1	173	0	0.0
262	53	1	0	123	282	0	1	95	1	2.0
81	45	1	1	128	308	0	0	170	0	0.0
52	62	1	2	130	231	0	1	146	0	1.8

data.isnull().sum()

```
Out[9]: age                       0
        sex                       0
        chest_pain_type           0
        resting_blood_pressure    0
        cholesterol               0
        fasting_blood_sugar       0
        rest_ecg                  0
        max_heart_rate_achieved   0
        exercise_induced_angina   0
        st_depression             0
        st_slope                  0
        num_major_vessels         0
        thalassemia               0
        target                    0
        dtype: int64
```

data.isnull().sum().sum()

```
Out[10]: 0
```

So, we have no missing values, therefore data cleaning is not required

#Correlation plot

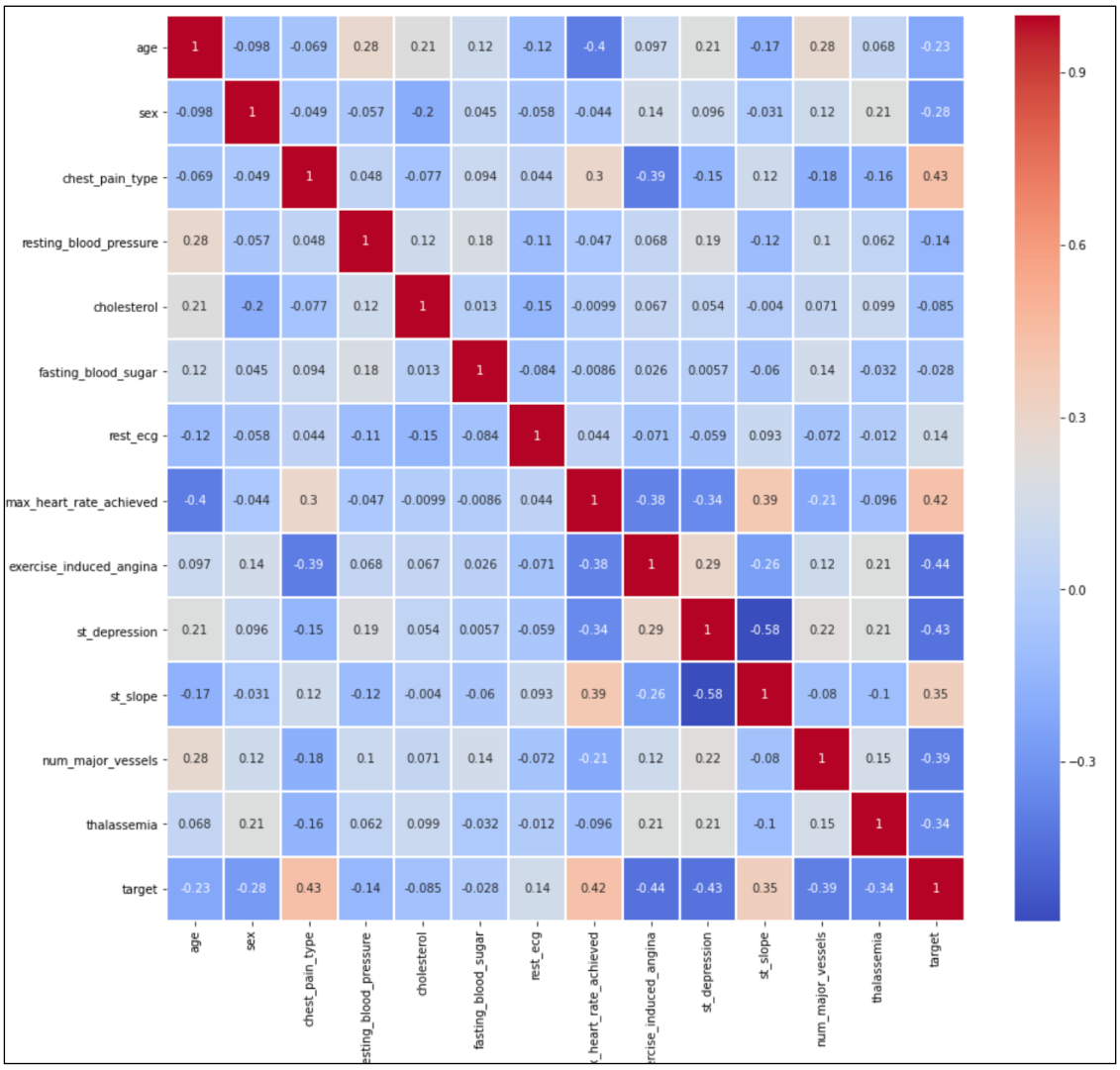
cnames=['age', 'sex', 'chest_pain_type', 'resting_blood_pressure', 'cholesterol', 'fasting_blood_sugar', 'rest_ecg', 'max_heart_rate_achieved', 'exercise_induced_angina', 'st_depression', 'st_slope', 'num_major_vessels', 'thalassemia','target']

plt.subplots(figsize=(15, 14)) *#Set the width and height of the plot*

corr = data.corr() *#Generate correlation matrix*

sns.heatmap(corr, annot = True, cmap='coolwarm',linewidths=.2) *#Plot using seaborn library*

plt.show()



#Splitting the dataset to Train and Test

```
from sklearn.model_selection import train_test_split  #used to split the dataset
predictors = data.drop("target",axis=1)
target = data["target"]
X_train,X_test,Y_train,Y_test = train_test_split(predictors, target, test_size=0.20,
random_state=0)
print("Training features have {0} records and Testing features have {1} records.".\
format(X_train.shape[0], X_test.shape[0]))
```

```
Training features have 242 records and Testing features have 61 records.
```

X_train.shape *#training model with predictor attributes*

```
Out[14]: (242, 13)
```

X_test.shape *#testing model with predictor attributes*

```
Out[15]: (61, 13)
```

Y_train.shape *#training model target attribute*

```
Out[16]: (242,)
```

Y_test.shape *#testing model target attribute*

```
Out[17]: (61,)
```

X_train.head()

Out[18]:

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	rest_ecg	max_heart_rate_achieved	exercise_induced_angina	st_de
74	43	0	2	122	213	0	1	165	0	0.2
153	66	0	2	146	278	0	0	152	0	0.0
64	58	1	2	140	211	1	0	165	0	0.0
296	63	0	0	124	197	0	1	136	1	0.0
287	57	1	1	154	232	0	0	164	0	0.0

X_test.head()

Out[19]:

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	rest_ecg	max_heart_rate_achieved	exercise_induced_angina	st_d
225	70	1	0	145	174	0	1	125	1	2.6
152	64	1	3	170	227	0	0	155	0	0.6
228	59	1	3	170	288	0	0	159	0	0.2
201	60	1	0	125	258	0	0	141	1	2.8
52	62	1	2	130	231	0	1	146	0	1.8

Y_train.head()

```
Out[20]: 74     1
         153    1
         64     1
         296    0
         287    0
         Name: target, dtype: int64
```

Y_test.head()

```
Out[21]: 225    0
         152    1
         228    0
         201    0
         52     1
         Name: target, dtype: int64
```

#Creating training model

```
from sklearn.metrics import accuracy_score
def train_model(X_train, y_train, X_test, y_test, classifier,**kwargs):
    model = classifier(**kwargs)
    model.fit(X_train,y_train)  #for fitting the training dataset
    fit_accuracy = model.score(X_train, y_train)
    test_accuracy = model.score(X_test, y_test)
    print(f"Train accuracy: {fit_accuracy:0.2%}")
    print(f"Test accuracy: {test_accuracy:0.2%}")
    return model
```

#Naïve Bayes

```
from sklearn.naive_bayes import GaussianNB
nb = train_model(X_train, Y_train, X_test, Y_test, GaussianNB)
nb.fit(X_train, Y_train)
y_pred_nb = nb.predict(X_test)
print(y_pred_nb)
```

```
Train accuracy: 83.47%
Test accuracy: 85.25%
[0 1 1 0 0 1 0 0 0 0 1 1 0 1 1 1 0 1 0 1 1 1 0 0 1 0 0 1 1 1 0 0 1 1 1 0 0
 1 0 0 1 1 0 0 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1]
```

score_nb = round(accuracy_score(y_pred_nb,Y_test)*100,2)

print("The accuracy score achieved using Naive Bayes is: "+str(score_nb)+" %")

```
The accuracy score achieved using Naive Bayes is: 85.25 %
```

#Confusion matrix

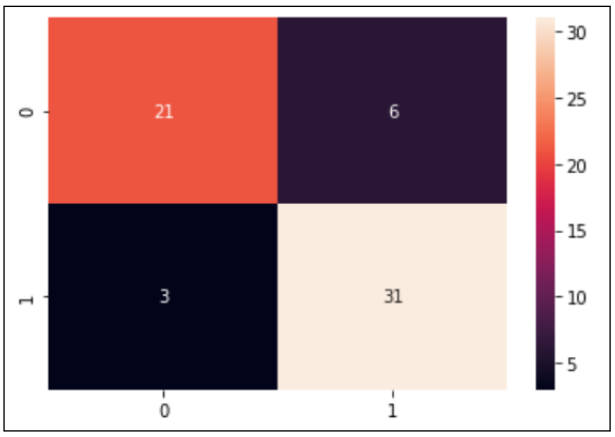
from sklearn.metrics import confusion_matrix

matrix= confusion_matrix(Y_test, y_pred_nb) *#Predicted values in x axis, actual values in y axis*

plt.subplots(figsize=(6,4))

sns.heatmap(matrix,annot = True)

plt.show()



#Precision score

from sklearn.metrics import precision_score

precision = precision_score(Y_test, y_pred_nb) *#ratio of correctly predicted positive to total predicted positive*

print("Precision: ",precision)

```
Precision:  0.8378378378378378
```

#Recall

from sklearn.metrics import recall_score

recall = recall_score(Y_test, y_pred_nb) *#ratio of correctly predicted positive to total actual positive*

print("Recall is: ",recall)

```
Recall is:  0.9117647058823529
```

#F-score

```
fs_nb=(2*precision*recall)/(precision+recall)
print(fs_nb)
```

```
0.8732394366197184
```

#KNN

```
from sklearn.neighbors import KNeighborsClassifier
model = train_model(X_train, Y_train, X_test, Y_test, KNeighborsClassifier)  #default k value is 5
```

```
Train accuracy: 78.10%
Test accuracy: 63.93%
```

```
from sklearn.neighbors import KNeighborsClassifier
for i in range(1,10):
    print("n_neigbors = "+str(i))
    train_model(X_train, Y_train, X_test, Y_test, KNeighborsClassifier, n_neighbors=i)
```

```
n_neigbors = 1
Train accuracy: 100.00%
Test accuracy: 52.46%
n_neigbors = 2
Train accuracy: 79.75%
Test accuracy: 59.02%
n_neigbors = 3
Train accuracy: 78.10%
Test accuracy: 63.93%
n_neigbors = 4
Train accuracy: 76.03%
Test accuracy: 63.93%
n_neigbors = 5
Train accuracy: 78.10%
Test accuracy: 63.93%
n_neigbors = 6
Train accuracy: 74.38%
Test accuracy: 65.57%
n_neigbors = 7
Train accuracy: 72.31%
Test accuracy: 67.21%
n_neigbors = 8
Train accuracy: 71.90%
Test accuracy: 68.85%
n_neigbors = 9
Train accuracy: 73.14%
Test accuracy: 67.21%
```

```
knn = train_model(X_train, Y_train, X_test, Y_test, KNeighborsClassifier, n_neighbors=8)
```

```
knn.fit(X_train, Y_train)
y_pred_knn = knn.predict(X_test)
print(y_pred_knn)
```

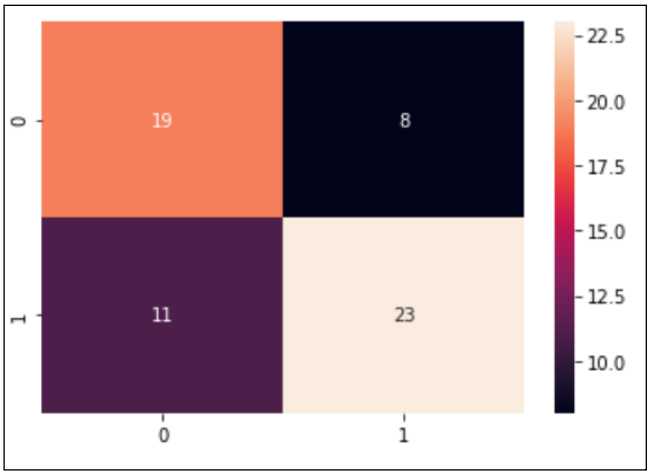
```
Train accuracy: 71.90%
Test accuracy: 68.85%
[0 0 1 0 1 1 0 0 0 0 1 1 0 1 1 1 0 1 0 1 1 1 0 0 0 0 1 0 1 1 0 0 1 0 1 0 0
 1 0 1 0 1 1 0 0 1 1 1 1 1 1 0 1 0 1 0 0 1 0 1 0]
```

```
score_knn = round(accuracy_score(y_pred_knn,Y_test)*100,2)
print("The accuracy score achieved using KNN is: "+str(score_knn)+" %")
```

```
The accuracy score achieved using KNN is: 68.85 %
```

#Confusion matrix

```
from sklearn.metrics import confusion_matrix
matrix= confusion_matrix(Y_test, y_pred_knn)  #Predicted values in x axis, actual values in y axis
plt.subplots(figsize=(6,4))
sns.heatmap(matrix,annot = True, fmt = "d")
plt.show()
```



#Precision score

```
from sklearn.metrics import precision_score
precision = precision_score(Y_test, y_pred_knn)  #ratio of correctly predicted positive to total predicted positive
print("Precision: ",precision)
```

```
Precision:  0.7419354838709677
```

#Recall

```
from sklearn.metrics import recall_score
recall = recall_score(Y_test, y_pred_knn)  #ratio of correctly predicted positive to total actual positive
print("Recall is: ",recall)
```

```
Recall is:  0.6764705882352942
```

#F-score

```
fs_knn=(2*precision*recall)/(precision+recall)
print(fs_knn)
```

```
0.7076923076923077
```

#Comparison of accuracies

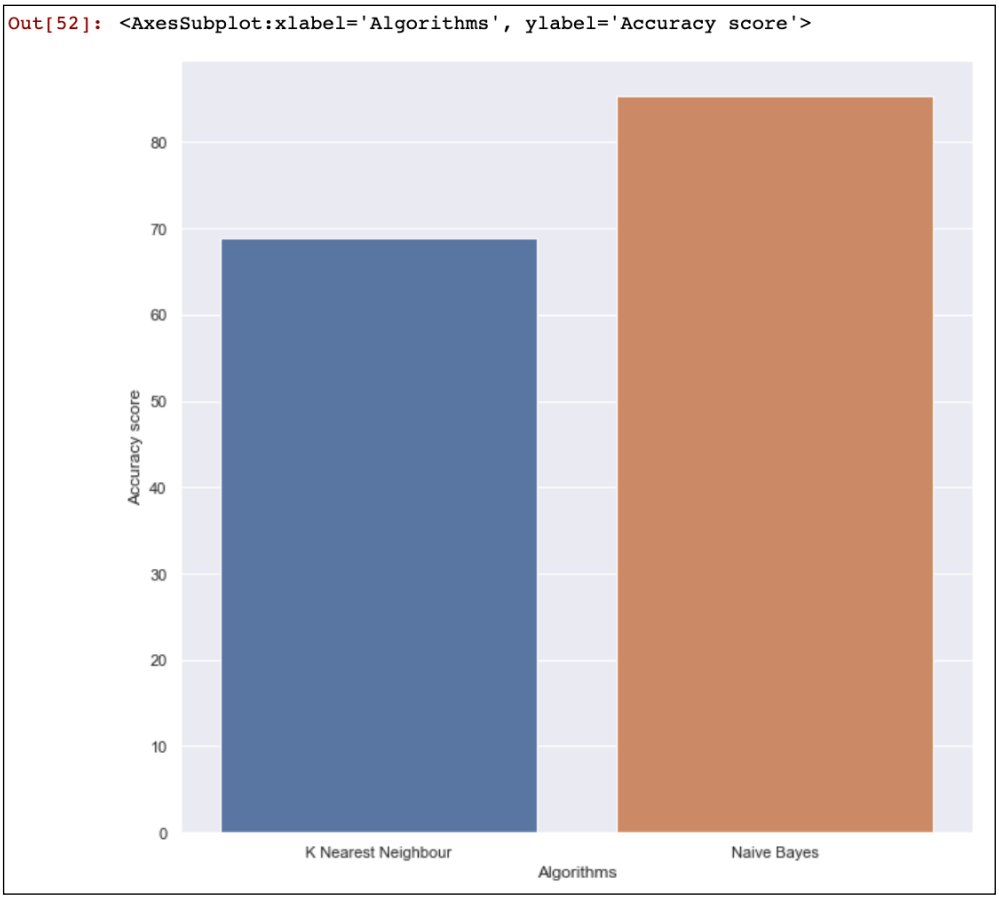
```
accuracy = [score_knn,score_nb]
classifiers = ['K Nearest Neighbour', 'Naive Bayes']
summary = pd.DataFrame({'accuracy':accuracy}, index=classifiers)  # create a dataframe from accuracy results
summary
```

Out[51]:

	accuracy
K Nearest Neighbour	68.85
Naive Bayes	85.25

```
scores = [score_knn,score_nb]
algorithms = ["K Nearest Neighbour","Naive Bayes"]
sns.set(rc={'figure.figsize':(10,10)})
plt.xlabel("Algorithms")
plt.ylabel("Accuracy score")
sns.barplot(algorithms,scores)
```

Out[52]: <AxesSubplot:xlabel='Algorithms', ylabel='Accuracy score'>



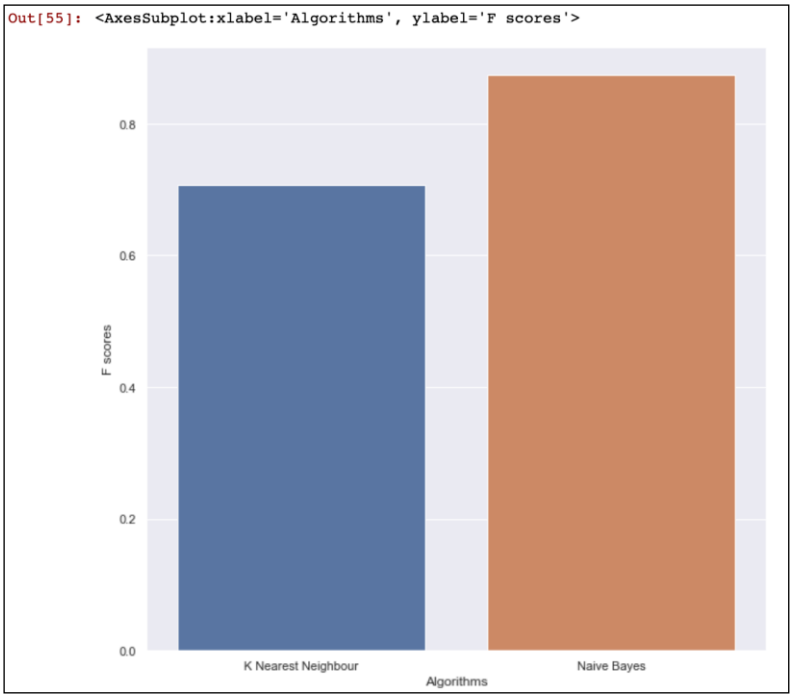
#Comparison of F-scores

```
f_scores=[fs_knn,fs_nb]
classifiers = ['K Nearest Neighbour', 'Naive Bayes']
fs_summary = pd.DataFrame({'f-scores':f_scores}, index=classifiers)
fs_summary
```

Out[54]:

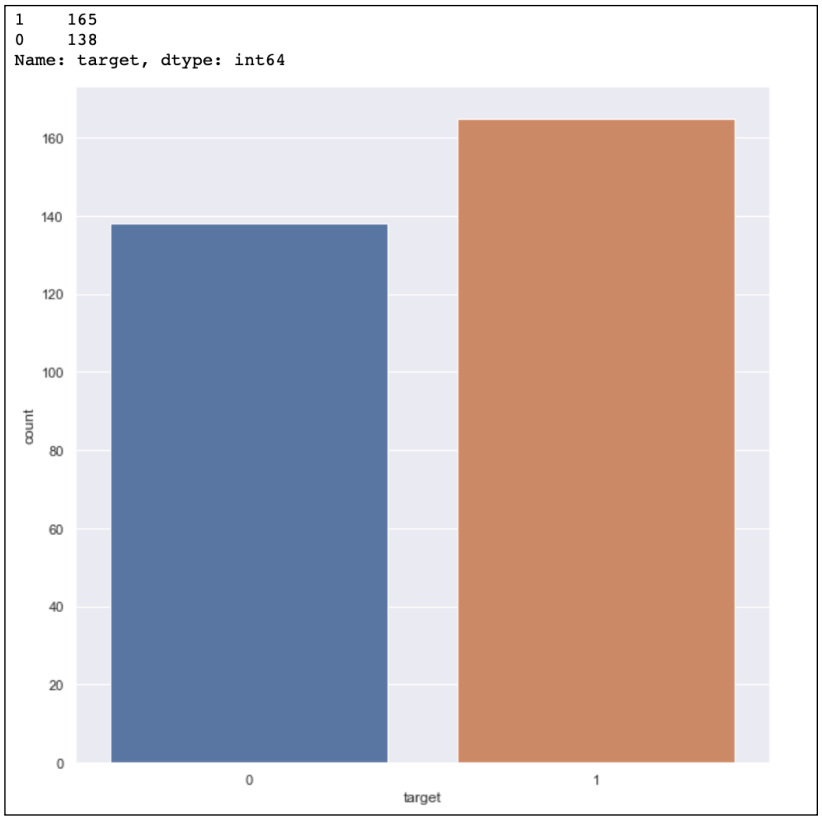
	f-scores
K Nearest Neighbour	0.707692
Naive Bayes	0.873239

```
fs_scores = [fs_knn,fs_nb]
algorithms = ["K Nearest Neighbour","Naive Bayes"]
sns.set(rc={'figure.figsize':(10,10)})
plt.xlabel("Algorithms")
plt.ylabel("F scores")
sns.barplot(algorithms,fs_scores)
```



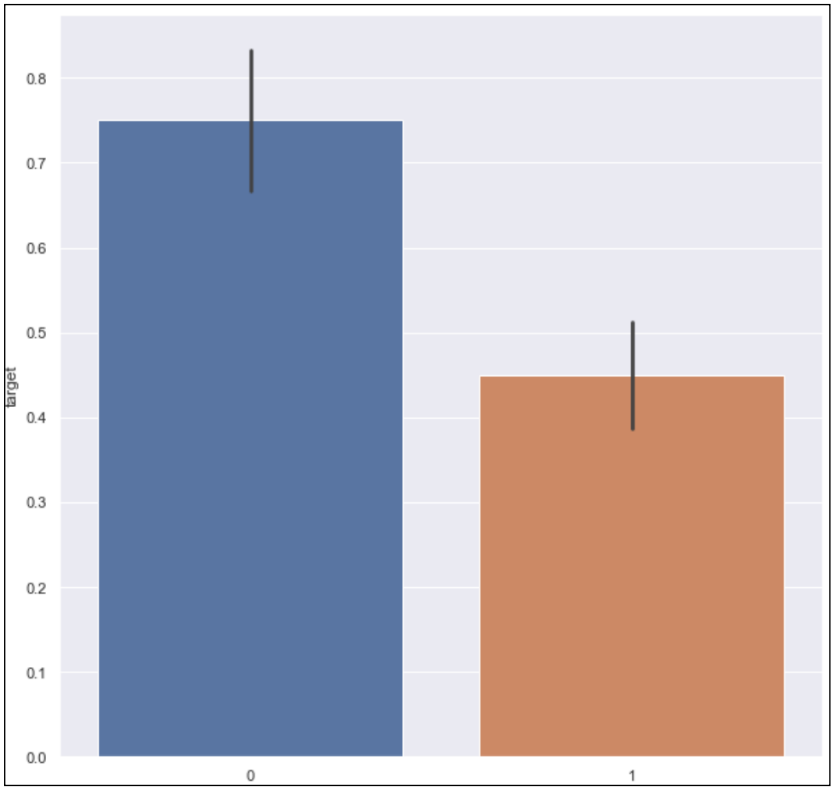
#Data Representation (count vs target)

```
plt.subplots(figsize=(10,10))
sns.countplot(data["target"])
target_temp = data.target.value_counts()
print(target_temp)
plt.show()
```



#Data Representation (target vs sex)

```
plt.subplots(figsize=(10,10))
sns.barplot(data["sex"],data["target"])
plt.show()
```



#Predicting values for custom input (using Naïve Bayes)

```
from sklearn.naive_bayes import GaussianNB
nb = train_model(X_train, Y_train, X_test, Y_test, GaussianNB)
nb.fit(X_train, Y_train)
pateint1=[[44,1,3,150,240,1,0,160,1,0.6,2,1,6]]
pateint2=[[37,1,2,130,250,0,1,187,0,3.5,0,0,2]]
result1 = nb.predict(pateint1)
result2 = nb.predict(pateint2)
print("For pateint 1 the predicted value of target is: ",result1[0])
print("For pateint 2 the predicted value of target is: ",result2[0])
```

```
Train accuracy: 83.47%
Test accuracy: 85.25%
For pateint 1 the predicted value of target is:  0
For pateint 2 the predicted value of target is:  1
```