HEART DISEASE PREDICTION USING TRADITIONAL MACHINE LEARNING

BY

PINKY PAUL
ID: 171-15-9479
AND

TONMOY RUDRA
ID: 171-15-9499

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

01 JUNE 2021

APPROVAL

This Project titled **"Heart Disease Prediction Using Traditional Machine Learning"**, submitted by Tonmoy Rudra, ID No: 171-15-9499 and Pinky Paul, ID No: 171-15-9479 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 01 June 2021.

BOARD OF EXAMINERS

Chairman

____

Dr. Touhid Bhuiyan

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

______________________________ **Internal Examiner**

Abdus Sattar

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

______________________ **Internal Examiner**

Md. Jueal Mia

Senior Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

External Examiner

Farid

____

Dr. Dewan Md. Farid

Associate Professor

Department of Computer Science and Engineering

United International University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Tarek Habib, Assistant Professor, Department of CSE,** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

SUPERVISED BY:

Md. Tarek Habib
Assistant Professor
Department of CSE
Daffodil International University

CO-SUPERVISED BY:

Mr. Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:

Tonmoy Rudra

Tonmoy Rudra
ID: 171-15-9499
Department of CSE
Daffodil International University

Pinky Paul

Pinky Paul
ID: 171-15-9479
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project successfully.

We really grateful and wish our profound our indebtedness to **Md. Tarek Habib**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Machine Learning” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan**, Professor and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

Heart disease is the main reason for death in the world in the course of the most recent decade. As per late study by WHO (World health organization) 17.9 million people die every year because of these type of diseases and it is expending quickly. With the expending populace and disease, it is become a challenge to diagnosing sickness and giving the appropriate therapy at the ideal time. But early prediction of heart disease may save numerous lives. However utilizing data mining methods can lessen the quantity of test that are required. In order to diminish number of death from heart disease there must be speedy and proficient detection procedure. This paper targets analyzing the different data mining procedures in particular Naive Bayes, Random Forest Classification, Decision tree and Support Vector Machine by utilizing a certified data set for heart disease prediction which is comprise of different features like sex, age, chest pain type, blood pressure, glucose and so forth. The research incorporates finding the correlations between the different features of the data set by using the standard data mining methods and hence utilizing the features appropriately to anticipate the possibility of a heart disease. These machine learning methods take least time for the prediction of the disease with more exactness which will reduce the dispose of valuable lives all over the world.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Introduction

A human body is comprised of various organs, all of which their own activities. Heart is one such organ which pumps blood throughout the body and on the off chance that it does not do as such, the human body have lethal conditions. One of the fundamental reasons of mortality today is having a heart disease. So, it gets important to ensure that the cardiovascular system or any other system in the human body besides should stay sound. Unfortunately, all over the world have been facing cardiovascular infections.

According to world Health Organization, 31% of all global deaths occur every year because of cardiovascular diseases. The huge number of deaths is regular among low and middle-income countries. In Bangladesh too, heart related diseases have become the main cause of mortality.

Many predisposing factors like individual and professional habits and hereditary predisposition accounts for heart disease. Different accustomed rick factors such as smoking, excessive use of alcohol, stress and physical inactivity alongside other physiological factors such as obesity, hypertension, high blood cholesterol and previous heart states are predisposing factors of coronary illness. A big challenge facing medical care (hospitals, medical centers) association is the arrangement of quality services at reasonable expenses. For this reason, effective and proper prediction of heart related disease is very necessary. Data mining techniques can be helpful in predicting these types of disease. Data mining is the way to take out valuable data and information from enormous data sets. Several data mining techniques such as regression, clustering, association rule and classification techniques are used to classify different heart disease features in prediction heart diseases. In our research we use the classification techniques like Decision tree, K-Nearest Neighbor and Support Vector Machine to predict heart related diseases using an effective data set. The classification model is advanced utilizing classification algorithms for prediction of heart disease. By using various type of classification algorithms for predicting heart disease made comparison among the existing systems. In this paper, we also mention future scopes of this research and progression possibilities. This paper can boost to significantly enhance the quality of clinical decisions.

1.2 Motivation

Like other South Asians, Bangladesh is unduly in client to create CAD (Coronary Artery Disease) which is mostly untimely in beginning follows a quickly dynamic course and angiographically more intense. Of all South Asians nations, Bangladesh presumably has the most elevated paces of TVD Cardiovascular disease and is the least studied in the worldwide battle against. Bangladesh is a country 'missing in action'. We notice that a lots of people around us have to face heart related problem but they do not find the main reason behind this heart related problems. Besides the people of our country have less income. This affected people do not have the ability to diagnose these type of heart related disease for its huge amount of cost. The cost of diagnosis of the heart disease is very high because it is the most difficult task in the medical field. All the factors such as bad eating habit, absence of rest, depression, obesity, poor diet, family history, hyper tension, high blood cholesterol inactive conduct, smoking are taken into consideration when analyzing and understanding the patients by the specialist through manual check-ups at normal time frames. After consider this situations, we thought that data mining techniques can be very effective in healthcare management. For this reason, we took decision to use data mining techniques like classification predict heart disease by analyzing an effective data set.

1.3 Objectives

- Make a complete database by correcting essential data.
- Study the cause of heart related problems.
- Detect the heart disease quickly by following different techniques.
- Reducing the cost to detect heart disease.
- Find a solution to save our lives from heart disease.

1.4 Expected Outcome

- Increase the accuracy rate of early prediction of heart disease.
- Development of health sector we can detect the heart disease.
- Decease the early death for heart disease.
- General people can take proper steps at appropriate time.
- We develop a system that predict heart disease at a short time.
- We make our system feasible to all users.

1.5 Report Layout

The report is decorated with five chapters. Different aspects of each chapter are discussed about “Heart Disease Prediction Using Traditional Machine Learning”. Various parts of each chapter are explaining here in detail.

- **Chapter 1: Introduction**
  Important theoretical concepts behind the project are discussed in this chapter. We also discuss about motivation, objectives and expected outcomes of this project here.
- **Chapter 2: Background**
  Related works, research synopsis, extent of the issues and challenges that are faced were discussed in this chapter.
- **Chapter 3: Research Methodology**
  Research subject and instrumentation, process of data collection, statistical analysis of data, choosing methodology and requirements for implementation are discussed in this chapter.
- **Chapter 4: Experimental Result & Discussion**
  Experimental setup and outcomes, analysis of the outcomes are discussed in this chapter.
- **Chapter 5: Summary, Conclusion and Implication for Future Research**
  Summary of the full study, conclusion and further work are discussed in this last chapter.

CHAPTER 2

BACKGROUND

2.1 Introduction

Heart disease affects millions of people and remains the leading cause of death in world wide. Data mining techniques that helps our technology to classify different features. In our research paper we have used classification techniques to predict heart disease. For this detection process at first, we need an efficient data set which have a good number of features. Then we need to follow some essential criteria like data prepossessing, data selection and accuracy checking. Different types of machine learning algorithm analysis will be discussed.

In this chapter, we give the details about the recent work, related work, research summary and details about the scope of our work. Our goals and challenges that we have face are also represented here briefly.

2.2 Related Work

There are vast use of data mining for prediction process. In medical section prediction process is very important. So data mining is extremely helpful in this sector. A good collection of recorded data can be helpful in predictions. Many diseases can be predicted quickly and efficiently by the use of different techniques of data mining. Researcher are using data mining techniques for the more colossal disease such as Heart Disease, Cancer etc. Uses of data mining techniques for prediction breast cancer indicates that, it can be given more accurate result of predicting this disease. Cancer sells at different states help to know the present condition and the changes after taking particular drug. Diabetes prediction also been predicted by this techniques.

Jyoti Soni showed in her paper decision tree and Bayesian classification's performance can be improved subsequent to utilizing genetic algorithm. On this paper she explain that what is the analysis before in using genetic algorithm and the analysis after. Clustering classification was used for that method. Clustering base performance was seen in this classifications.

Poornima Singh, Sanjay Singh and Jayatri S Pandi-Jain developed an effective heart disease prediction system (ESDPS) using neural network which can use for predicting the risk level of heart disease. Using 15 parameters of medical such as age sex, blood pressure, cholesterol and obesity the system can predict the likelihood of patient getting heart disease. It can process significant knowledge about relationship between medical factors related to heart disease and patterns that to be established. The results which are obtained from it represented that system can effectively predict the risk level of heart diseases. The system predicts heart disease with~100% accuracy by using neural networks.

2.3 Research Summery

Nowadays most hospitals take some steps to manage their medical care or information of patients. Using this process huge amounts of data are collected which appear as numbers, text, diagrams and pictures. But unfortunately this data are infrequently used in decision making process of medical sector.

In this research we use classification technique and show the analysis of various machine learning algorithm for predicting heart disease at early stage. The algorithms that are used in our paper are Naive Bayes, Random Forest Classification, Decision tree and Support Vector Machine. These algorithm can be very useful for specialist needs or clinical experts for precisely analysis heart disease. This paper work we include current times data set of cardio vascular disease. Then we select our ideal algorithm based on the accuracy and execution utilizing different execution measurement. By using this procedure heart disease prediction will become easy to help specialist.

2.4 Scope of the problem

- As this research is prediction type research, so data analysis was very essential part for this research and this data analysis process was very difficult and time consuming.
- The dataset was sparse which was difficult for prediction purpose and decreased the accuracy score if the scale factor was set to get a refine.
- As we didn’t have enough knowledge about various classifier algorithm, so it took a lot of time to acquire the proper concept of these algorithms.

2.5 Challenges

- First of all for this pandemic situation we could not collect data physically because data collection from different hospitals was too much risk for us on this situation.
- Although there are enormous data set find in online but the selection of an effective data set was a tough task for us.
- As we selected a big data set, it was difficult to manage it for further working procedure.
- As we collect our data set from online, it was not easy to check the effectiveness of the features of our data set.
- The preprocessing process procedure of our big data set was not simple because the data set contains 11 medical features.
- Choosing effective machine learning algorithms was took huge amount of time because we have needed to gain enough knowledge to implement these algorithms.
- As there were much previous documentation and research paper on this topic, we have needed much time to acquire proper knowledge about this.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This research aims to find the cause of having heart disease with the help of computerized prediction process. This is helpful for saving heart disease patients live. To achieve the aim we analysis the data set by using various machine learning algorithms and which is mentioned in this research paper. Classification technique is very useful for predicting colossal disease. So we use this technique of data mining to achieve our goal. Here we discuss the full procedure of data analysis using classification technique step by step and also describe the implementation of the choosing algorithms.

3.2 Research subject and instrumentation

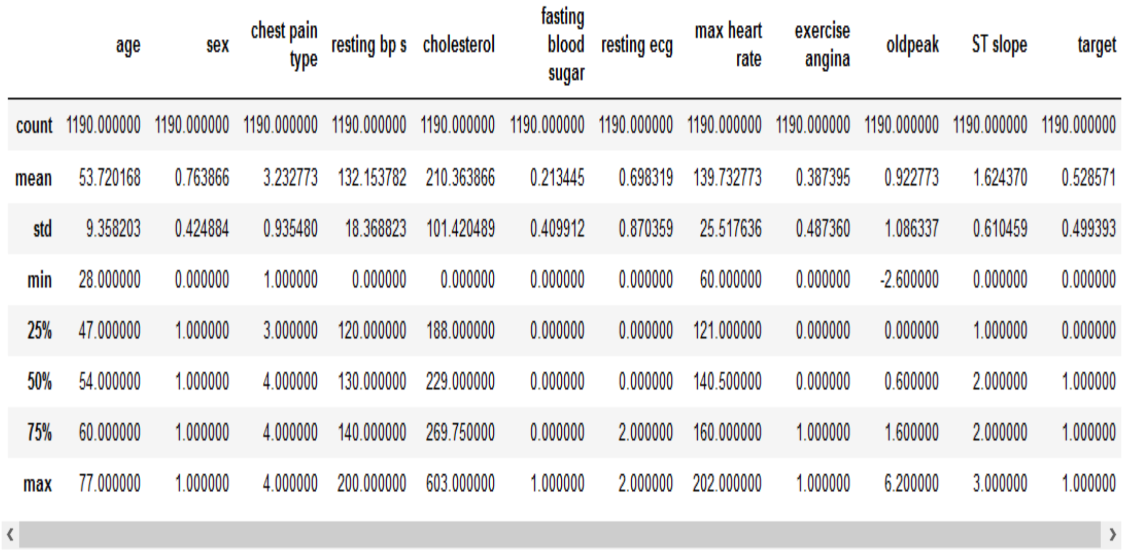
Through our research work we try to give a clear concept of our research procedure. As we collect the data set from online so the research procedure is fully based on computer. For completing our work we used windows platform. The implementation of our work is fully python programming language based and we also used some library packages- pandas, nampy matplotlib: pyplot, sklearn, cufflinks, seaborn, os, plotly, csv. We have used **Jupyter Notebook** as IDE. We have used python because of its reliability in fast testing of any complex algorithm. It is also very useful for making machine learning application.

3.3 Data collection procedure

For this pandemic situation the physically data collection procedure was very risky to us. So we have collected the data set from online. The data collection procedure was not so easy for us. We collected our data set from kaggle. First of all we have searched a data set with a good number of features and less missing value. Then we found an effective data set from kaggle which consists of 1190 patient’s records and 11 features with target value. This data set was created by Manu Siddhartha. We selected this data set for having a good number of patient’s data and enough features for better prediction.

3.4 Statistical analysis

This is a tough work to select an effective data set from online. After selecting a dataset the process of analysis the dataset is useful for gathering more information about the dataset for better understanding. Our dataset contains 11 features and using some analysis technique we can get better understanding about the features of the dataset.



	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
count	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000	1190.000000
mean	53.720168	0.763866	3.232773	132.153782	210.363866	0.213445	0.698319	139.732773	0.387395	0.922773	1.624370	0.528571
std	9.358203	0.424884	0.935480	18.368823	101.420489	0.409912	0.870359	25.517636	0.487360	1.086337	0.610459	0.499393
min	28.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	60.000000	0.000000	-2.600000	0.000000	0.000000
25%	47.000000	1.000000	3.000000	120.000000	188.000000	0.000000	0.000000	121.000000	0.000000	0.000000	1.000000	0.000000
50%	54.000000	1.000000	4.000000	130.000000	229.000000	0.000000	0.000000	140.500000	0.000000	0.600000	2.000000	1.000000
75%	60.000000	1.000000	4.000000	140.000000	269.750000	0.000000	2.000000	160.000000	1.000000	1.600000	2.000000	1.000000
max	77.000000	1.000000	4.000000	200.000000	603.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	1.000000

Figure 3.1: Statistical information of features

The describe() function is one of the functions provided by pandas library. This function is mainly used for retrieving some statistical information of the dataset. Here, statistical information means standard, mean, min of the values of the using attributes. The statistical information of 11 attributes of our dataset have easily counted by using this functions. The statistical analysis of the features is so much beneficial for understanding the dataset.

3.5 Proposed methodology

Different machine learning algorithm known as classifier that can help us for predicting in our project. Through our project we are looking forward to predict the heart disease. For this predicting purpose we are choosing 3 algorithms that it will give us better and reliable prediction. We have chosen more than one algorithm for comparing them with one another. If the accuracy of one algorithm varies at a large scale with others accuracy we can understand that the algorithm is not suitable for this data or we had a mistake in our coding.

So it is so much effective to use more than one algorithm for any prediction based project. We have taken the decision to use Decision Tree, K-Nearest Neighbor and Support Vector Machine in our project. The algorithms we are using are as follows:

3.5.1 Decision Tree

We are using Decision Tree classifier as our first algorithm. It is mainly used for classification but it is useful for both classification and regression problems. Decision Tree works with various types of data such as categorical and numerical. But a question might arise why we are choosing this algorithm over other algorithms. To give the answer of that question we mention here two reasons. Decision Tree mostly follows the same way human brain thinks. For this reason it is easy to understand the data and easily find the conclusion of the problems. The other reason, we handle the medical data set in a simple way and most widely.

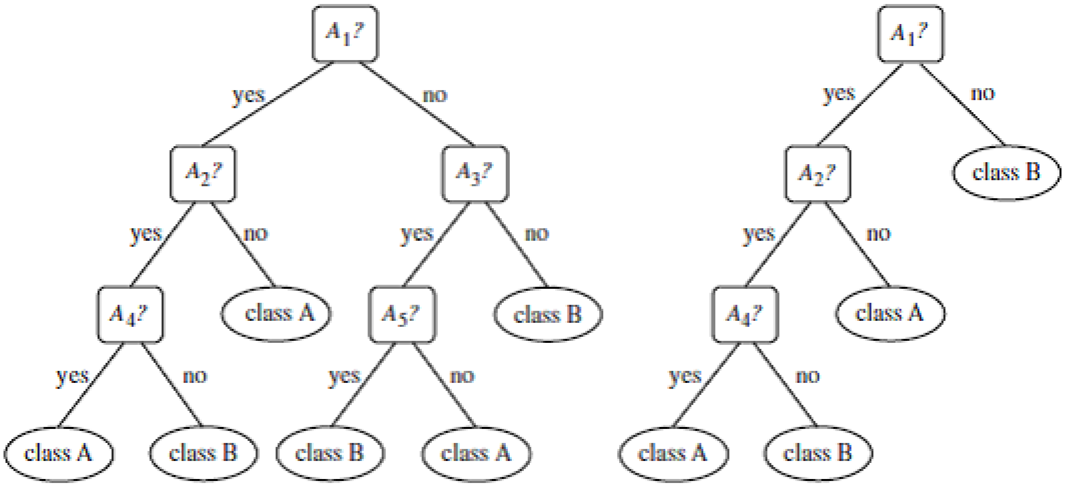


Figure 3.2: Basic decision tree and its pruned version

Here, there are three types of note in decision tree

- Root node: it is mane node of the tree and it’s based on others node functions.
- Interior node: various features are handled through it.
- Leaf node: each test result is found in leaf node.

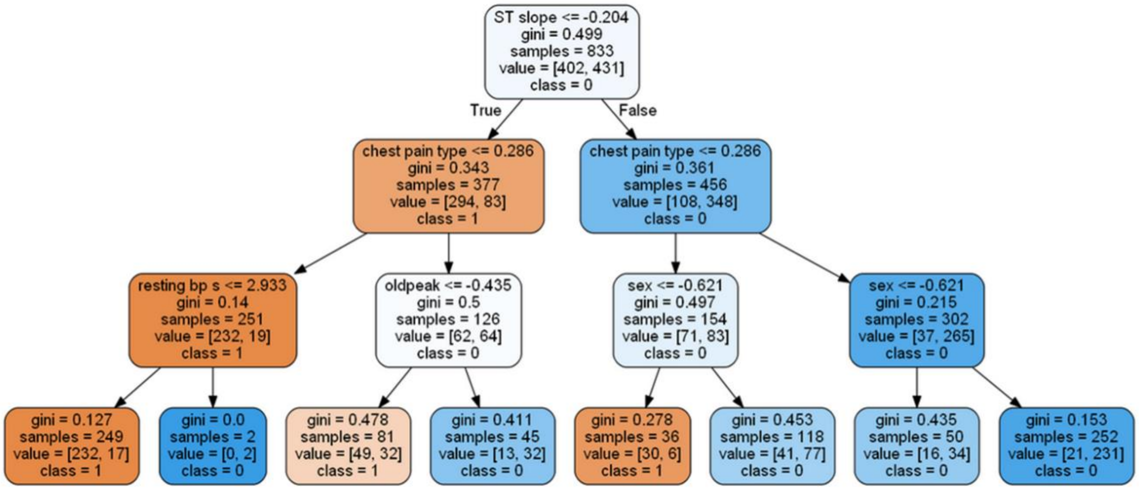
Each link of the tree represent a decision.

DecisionTreeClassifier function is required for using decision tree algorithm which provided by Sklearn module. The train part of the dataset is used for training the algorithm. Then, the decision tree algorithm works with these data in its own process. First of all, it select the best feature using Attributes Selection Measure (ASN) through splitting the data. The feature which has the most influence on the value of Y is selected as the root node. The two methods Gini index and information gain both of these are used to select the root node of the decision tree.

$$\textbf{Gini Index} = 1 - \sum_{j=1}^{c} p_j^2$$

Here, Pj is proportion of the samples that belongs to class c for a particular node.

Gini Index is used to measure how often a randomly chosen element would be incorrectly identified. So we understand easily that the feature with less Gini Index should be preferred for node. “gini” criteria is selected by default for Gini Index in Sklearn and it takes “gini” value.



Type equation here.

Figure 3.3: Making of decission tree using Gini Index

Here, the png graph of decision tree shows that Gini Index method is used for selecting the decision nodes in different levels. For this classification method “gini” criteria passed as a parameter of DecisionTreeClassifier function.

$$\textbf{Entropy} = - \sum_{j=1}^{c} p_j \, log_2(p_j)$$

Here, Pj is proportion of the samples that belongs to class c for a particular node.

Entropy is another method of selecting the features in different nodes of decision tree. The measure of uncertainly of a random variable is entropy. The decrease of entropy is mainly Information Gain. When we are partitioning the training instances into smaller subsets with the use of a node of a decision tree, the entropy is typically changed. Information gain computes the difference of changing entropy. Decision tree algorithm uses this information gain. “Entropy” is used as criteria for information gain which supported by Sklearn. For using Information Gain method in Sklearn we have to set “entropy” criteria explicitly.

Decision tree algorithm use one of the method from them to make the feature in decision nodes and breaks the data set into smaller subsets. This algorithm starts to make the tree by recursively repeating the process for every child. When the same feature value has found in all the tuples or no more features are remaining or no more instances have found the algorithm stop its work.

After training the model the test dataset is used for prediction purpose. Then by comparing with test target value we can find the accuracy of the model.

3.5.2 k-Nearest Neighbor (KNN)

The k-Nearest Neighbor algorithm is one of the most supervised machine learning algorithm. The technique of classifying object is depended on nearest neighbor. All cases are stored in it and based on similarity manager it classifies the new cases. KNN is mostly use in estimation of statistical point and recognition of patterns. It calculates the distance between a feature and it neighbor using euclidean distance. It makes a group with marked point and uses this group to mark another point similarity among the data is used as base for clustering the data. This algorithm is also much effective so fill the null values of the data. KNN is broadly classified in two types. One is NN techniques with Structure and other is NN techniques without Structure. In our project we have used NN techniques without Structure. For this purpose we have to classify our data set into training and test part. Distance between training point and sample point is measured. The point which have lowest distance is called nearest neighbor. If the features are continues, KNN algorithm works more efficiently.

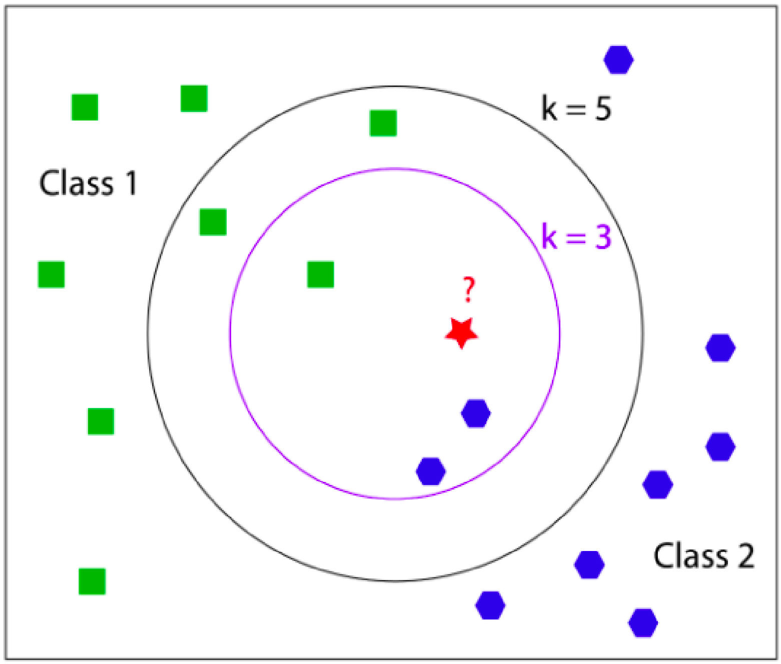


Figure 3.4: Choosing the value of K

The main reason behind the choosing this algorithm is it works well with numerical features. As the value of every features in our data set are numerical, so we have preferred this algorithm for our project.

First of all, KNeighborsClassifier function is need to be imported by using Scikit-Learn module. The process of training the algorithm is so much straight forward and can easily make prediction with it.

After importing KNeighborsClassifier function, it initializes by passing n_neighbors(K) as parameter and set the value of K for the parameter. No ideal value for K is found before testing and evaluation. Some data are used for training the model which are called as train data. Then the algorithm make prediction based on the test dataset and which is also used for finding the accuracy score. The mean error for the values of test set are plotted for predicting all the K values of a small range. Different accuracies are found from different K values in this algorithm.

3.5.3 Support Vector Machine (SVM)

Support Vector Machine is another machine learning algorithm for classification approach. The concepts of SVM are relatively simple. We have used the algorithm for classification related work but it can be handled both classification and regression problems. The data set having multiple continuous or categorical variables is easily handled by this algorithm. The working procedure of this algorithm is quite a bit different because it works by defining decision boundaries. First of all, plotting each data item of the data set as a point is the initial task of thus algorithm. We have plotted these data items in-dimensional (here, N is number of features we have used). The value of every feature bring the value of particular coordinate. The main part of SVM is construction of hyperplane in multidimensional space. SVM perform its classification by finding these hyperplane. The hyperplane can differentiate the two classes very well. In order to minimize and error, optimal hyperplane is need to generate through and iterative manner.

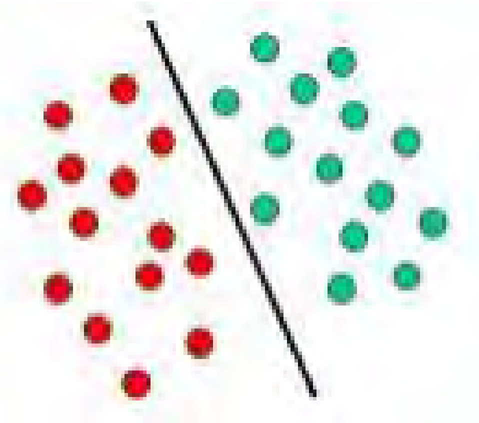


Figure 3.5: General SVM classification for linear hyper plane

The main target of SVM is making the best partitions of the data set into classes by finding a maximum marginal hyperplane. SVM finds this maximum marginal hyperplane through some steps. Firstly, generate some hyperplanes which separate the classes in the most IBL manner. Secondly, the selection process of the effective hyperplane with the best separation from the either nearest data points.

SVM is so much efficient in high dimensional spaces and it is also efficient when the number of dimension exceed the number of samples. In this situation a technique is used which called kernel trick. The kernel transforms the low dimensional input space into the higher dimensional space. This technique is most effective for curve hyperplane shows at figure 3.6.

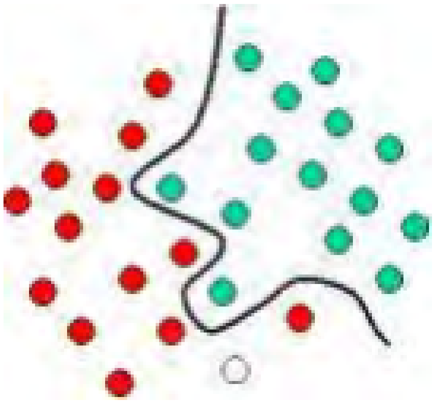


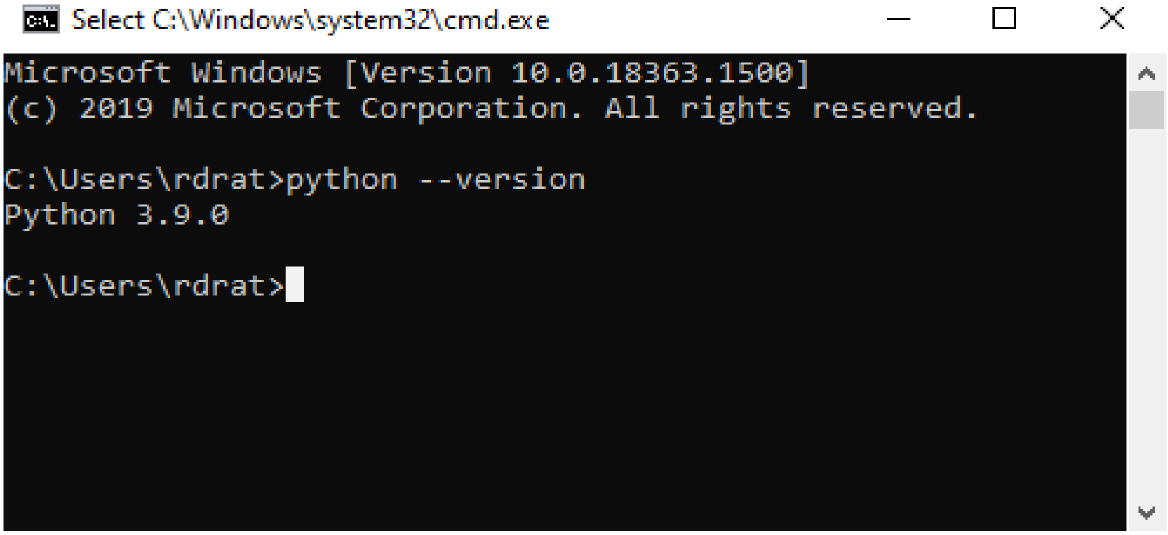
Figure 3.6: SVM classification for curve hyper plane

Different kernel functions can be fixed for the function of decision making which make its versatility. Some common kernels are provided for this, but we can also classify custom kernels.

Among the kernel functions the radial basis function kernel is highly populated for this classification. For this type kernel function gamma has to be passed as parameter of SVC() function. The train data has to be used for training the model. Then, test data is used for performing the prediction task of the model.

3.6 Implementation Requirements

Every step of our working process be discuss in this chapter one by one. First of all we have install the update version of **Python** application for implementation of our project.

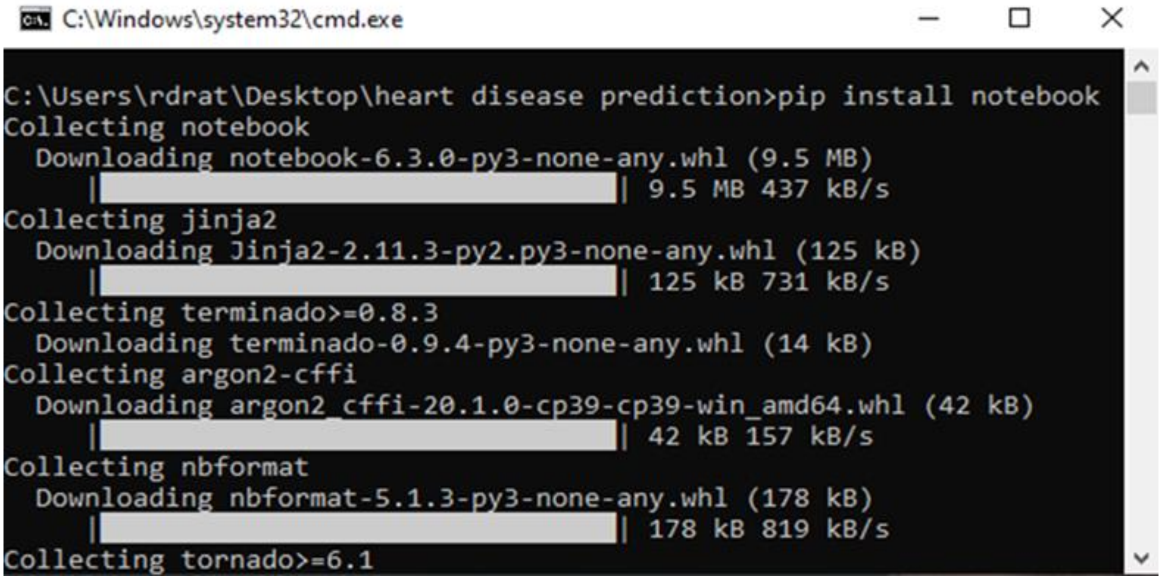


```
Microsoft Windows [Version 10.0.18363.1500]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\rdrat>python --version
Python 3.9.0

C:\Users\rdrat>
```

Figure 3.7: Python update version

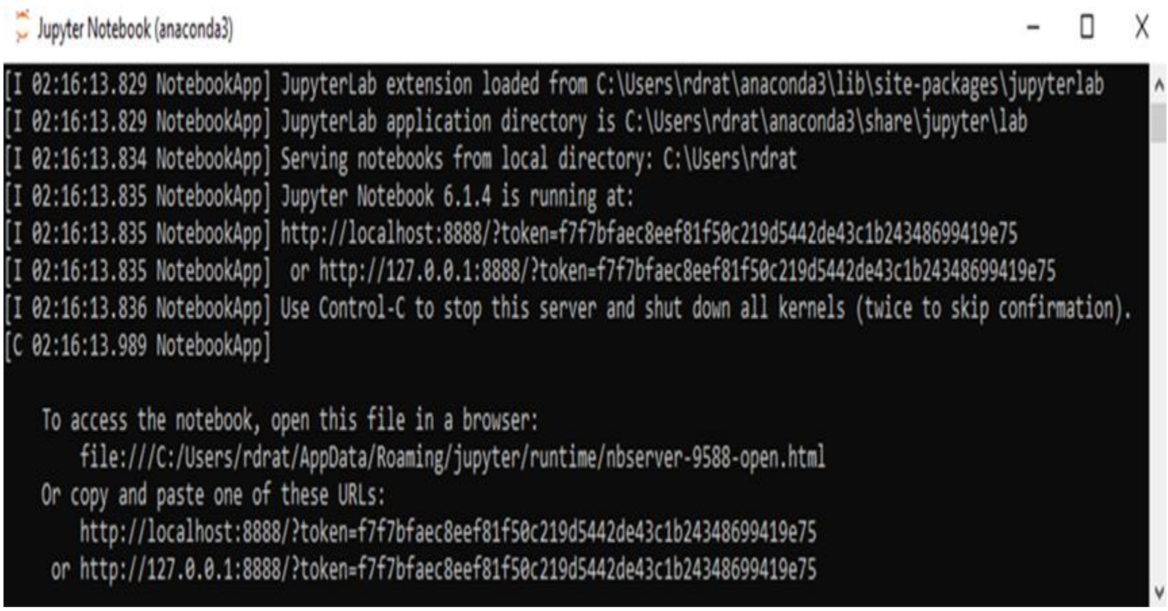
After installing the **Python** application, we install the **Jupyter Notebook** we have used it as IDE for our research work. The installing process are included here-

Jupyter Notebook installation-



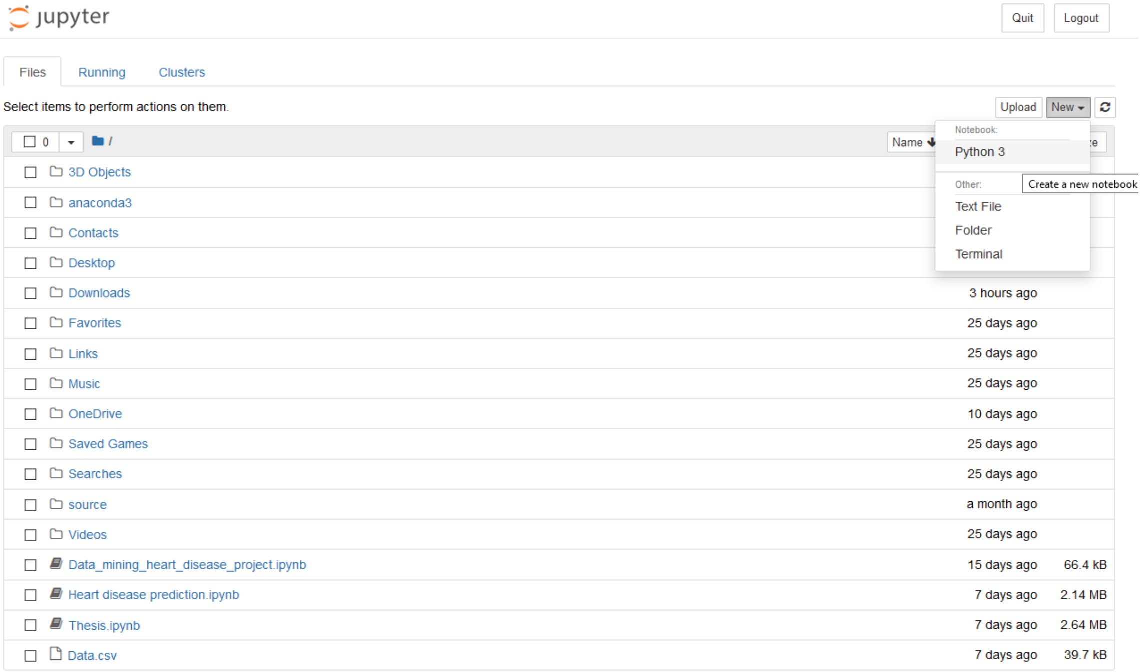
```
C:\Users\rdrat\Desktop\heart disease prediction>pip install notebook
Collecting notebook
  Downloading notebook-6.3.0-py3-none-any.whl (9.5 MB)
     |████████████████████████████████| 9.5 MB 437 kB/s
Collecting jinja2
  Downloading Jinja2-2.11.3-py2.py3-none-any.whl (125 kB)
     |████████████████████████████████| 125 kB 731 kB/s
Collecting terminado>=0.8.3
  Downloading terminado-0.9.4-py3-none-any.whl (14 kB)
Collecting argon2-cffi
  Downloading argon2_cffi-20.1.0-cp39-cp39-win_amd64.whl (42 kB)
     |████████████████████████████████| 42 kB 157 kB/s
Collecting nbformat
  Downloading nbformat-5.1.3-py3-none-any.whl (178 kB)
     |████████████████████████████████| 178 kB 819 kB/s
Collecting tornado>=6.1
```

Figure 3.8: Jupyter Notebook Installation



```
[I 02:16:13.829 NotebookApp] JupyterLab extension loaded from C:\Users\rdrat\anaconda3\lib\site-packages\jupyterlab
[I 02:16:13.829 NotebookApp] JupyterLab application directory is C:\Users\rdrat\anaconda3\share\jupyter\lab
[I 02:16:13.834 NotebookApp] Serving notebooks from local directory: C:\Users\rdrat
[I 02:16:13.835 NotebookApp] Jupyter Notebook 6.1.4 is running at:
[I 02:16:13.835 NotebookApp] http://localhost:8888/?token=f7f7bfaec8eef81f50c219d5442de43c1b24348699419e75
[I 02:16:13.835 NotebookApp]  or http://127.0.0.1:8888/?token=f7f7bfaec8eef81f50c219d5442de43c1b24348699419e75
[I 02:16:13.836 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
[C 02:16:13.989 NotebookApp]

    To access the notebook, open this file in a browser:
        file:///C:/Users/rdrat/AppData/Roaming/jupyter/runtime/nbserver-9588-open.html
    Or copy and paste one of these URLs:
        http://localhost:8888/?token=f7f7bfaec8eef81f50c219d5442de43c1b24348699419e75
     or http://127.0.0.1:8888/?token=f7f7bfaec8eef81f50c219d5442de43c1b24348699419e75
```

Figure 3.9: Opening Notebook



Select items to perform actions on them.

□ 3D Objects		
□ anaconda3		
□ Contacts		
□ Desktop		
□ Downloads	3 hours ago	
□ Favorites	25 days ago	
□ Links	25 days ago	
□ Music	25 days ago	
□ OneDrive	10 days ago	
□ Saved Games	25 days ago	
□ Searches	25 days ago	
□ source	a month ago	
□ Videos	25 days ago	
□ Data_mining_heart_disease_project.ipynb	15 days ago	66.4 kB
□ Heart disease prediction.ipynb	7 days ago	2.14 MB
□ Thesis.ipynb	7 days ago	2.64 MB
□ Data.csv	7 days ago	39.7 kB

Figure 3.10: Jupyter Notebook Interface

A sample interface of Jupyter Notebook is also added here. A Python file was created by using Jupyter Notebook to start the implementation of our research work.

3.7 Required Libraries Installation

In order to implement of our research work we have required some library installation. Without installation of this libraries the visualization process of data and some checking procedure were impossible. Here we try to give a brief description of our required libraries.

3.7.1 Pandas

Pandas is an open source Python package by which we can analysis the data and complete machine learning task. Numpy is the parent package of Pandas. It provides multi-dimensional arrays support. Python is included in its Python distribution because inside the Python ecosystem it works well. For analyzing our data set Pandas was used in our research.

3.7.2 Numpy

Numpy which stands for Numerical Python and it is an array processing package of Python. Multidimensional array objects performance can be high by using and it works with these array as a tool. Square brackets are used for accessing the elements in Numpy arrays. We used Numpy for achieve high performance while we have needed to work with array data.

3.7.3 Plotly-Python

The Plotly-Python is also an open source library which is interactive. Data visualization is very easy by using this library. And for this we can understand the data simply and easily. Scatter plots, line charts, bar charts, box plots, histograms, pie charts are the various types of graphs and charts which can be plotted through it. As we used different types of graph and charts for visualize our data so Plotly-Python library was essential for our research.

3.7.4 Matplotlib and typlot

The typlot is an API which is a stateful interface of MATLAB-style. It was originally written for the alternative of MATLAB. It mainly includes everything visualized in a plot which have one or more axes. For understanding how to work with plots the use of matplotlib and typlot was essential in our project.

3.7.5 Seaborn

Seabron is also used for data visualization. It is mainly built on top of matplotlib. It is closely incorporated with Pandas. Visualize the data is the main part of Seaborn which can help to explore and understand the data. Maximum data of our project were visualized with the help of Seaborn. It is a very essential Python library for our project.

3.7.6 CSV

CSV means Comma Separated Values which is a file format of storing tabular data such as spread sheet or data base. Tabular data of a CSV file are stored in plain text. One or more fields of its record are separated by commas. As our data set is in CSV format so we have needed to use the inbuilt module of Python called **csv**.

3.7.7 Sklearn (Scikit-learn)

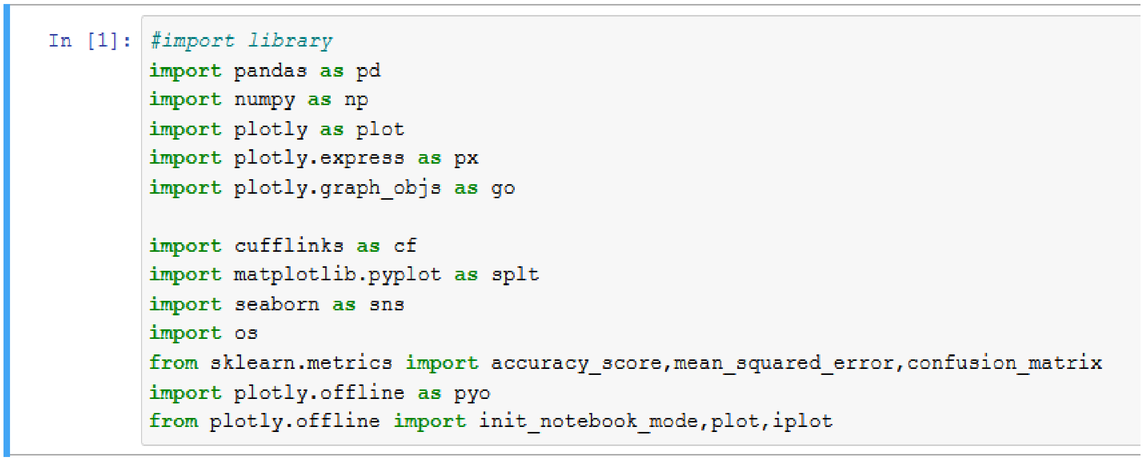
It is probably the most effective Python library for machine learning. A lot of efficient machine learning tools and statistical models such as classification, regression, plastering are included in sklearn library. To build our machine learning model we use sklearn because classification technique was used for prediction.

3.7.8 Plotly Express

Plotly.express is one kin of module which contains function to create full figures at ones, and is mention to as Plotly Express. It is a built-in part of **plotly** library. More than 30 functions are include in Plotly Express and can create different types of figures. We used Plotly Express for our various types of graphs and charts throughout a data exploration session.

3.7.9 Cufflink

Cufflink is mainly a connector which connects plotly with pandas for creating graphs and charts of data frames directly. There are various types plotlys such as boxplot, spreadplot which are used cufflink and plotly also. For creating boxplot in our project, we use cufflinks.



```
In [1]: #import library
        import pandas as pd
        import numpy as np
        import plotly as plot
        import plotly.express as px
        import plotly.graph_objs as go

        import cufflinks as cf
        import matplotlib.pyplot as splt
        import seaborn as sns
        import os
        from sklearn.metrics import accuracy_score,mean_squared_error,confusion_matrix
        import plotly.offline as pyo
        from plotly.offline import init_notebook_mode,plot,iplot
```

Figure 3.11: Import Required Library

Here, required library packages have been installed for start our project implementation. All the above library packages are used in respective task for implementation.

CHAPTER 4

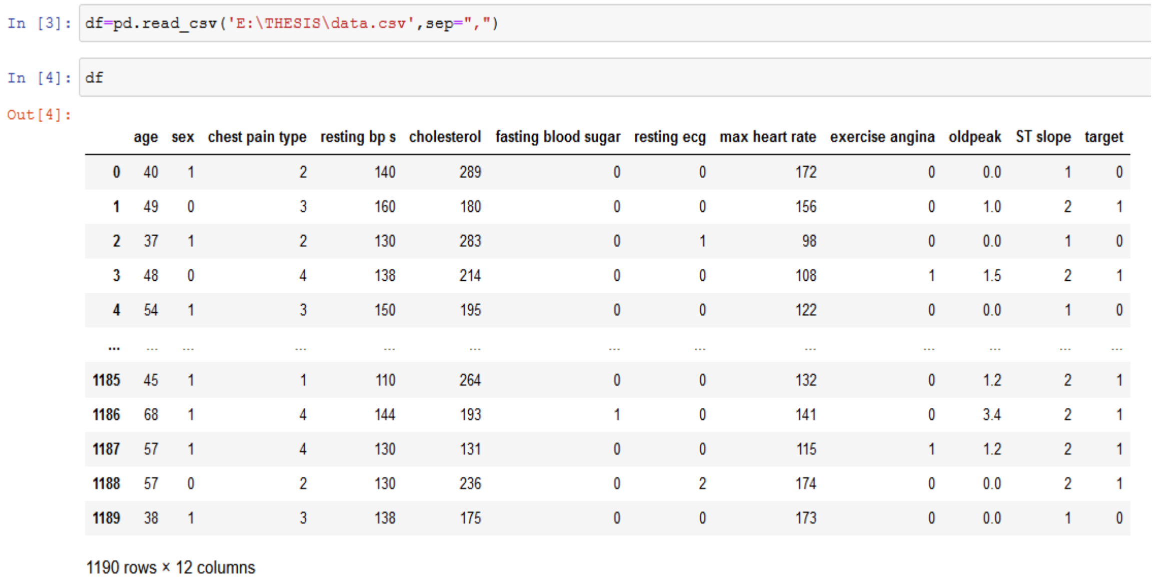
Experimental Result & Discussion

4.1 Introduction

Previous work in this field, the dataset that we used in our project and the selection of various classifier algorithms were the base of this chapter. Data preprocessing and the result that we found after applying the algorithms were discussed in this chapter and also analyzed them.

4.2 Experimental setup

First of all the dataset has to be obtained which contains the features of different people suffering from heart disease or not suffering from heart disease.



```
In [3]: df=pd.read_csv('E:\THESIS\data.csv',sep=",")

In [4]: df
Out[4]:
```

	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1185	45	1	1	110	264	0	0	132	0	1.2	2	1
1186	68	1	4	144	193	1	0	141	0	3.4	2	1
1187	57	1	4	130	131	0	0	115	1	1.2	2	1
1188	57	0	2	130	236	0	2	174	0	0.0	2	1
1189	38	1	3	138	175	0	0	173	0	0.0	1	0

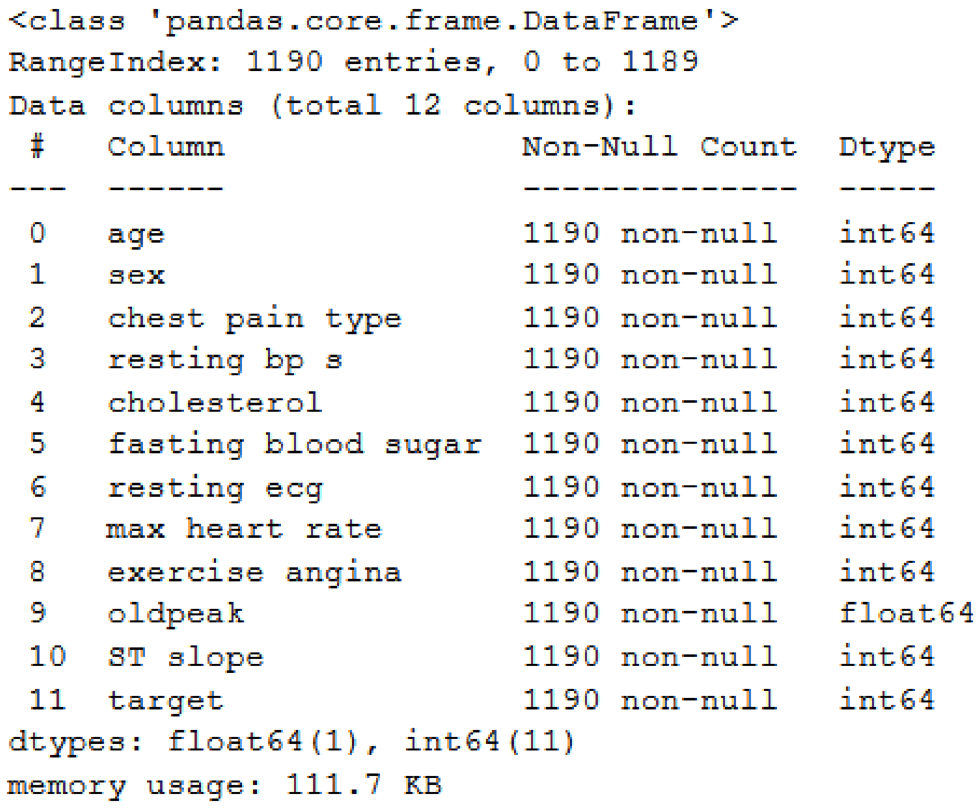
1190 rows × 12 columns

Figure 4.1: csv dataset

The system read the csv file of our dataset with the help of pandas library and gave the output with 1190 rows of individual patient’s data and 12 columns. The first 11 columns of this data set contains the value of various features for predicting heart diseases and the last column contains the target value which indicates that the patient had heart diseases or not.

4.2.1 Data Exploration & Visualization

The project which is used to predict purpose, data analysis is the most valuable portion of this. Through this data analysis process one can easily find the correlation between the features and the target value. So, after collecting our dataset we have started our data analysis process. For this process, the full information of the dataset is required.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1190 entries, 0 to 1189
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   age                  1190 non-null   int64
 1   sex                  1190 non-null   int64
 2   chest pain type      1190 non-null   int64
 3   resting bp s         1190 non-null   int64
 4   cholesterol          1190 non-null   int64
 5   fasting blood sugar  1190 non-null   int64
 6   resting ecg          1190 non-null   int64
 7   max heart rate       1190 non-null   int64
 8   exercise angina      1190 non-null   int64
 9   oldpeak              1190 non-null   float64
 10  ST slope             1190 non-null   int64
 11  target               1190 non-null   int64
dtypes: float64(1), int64(11)
memory usage: 111.7 KB
```

Figure 4.2: Concise summary of the DataFrame

Then we have used info() function on the dataset which is provided by the pandas library to summarize the full DataFrame.

The target feature contains binary value 0 or 1. The target feature with the value 0 indicates that the patient don't have heart disease and the value 1 indicates that the patient have heart disease. Now, we checked the data set is it balanced or not. We created count plot on a figure with the help of seaborn library based on target feature.

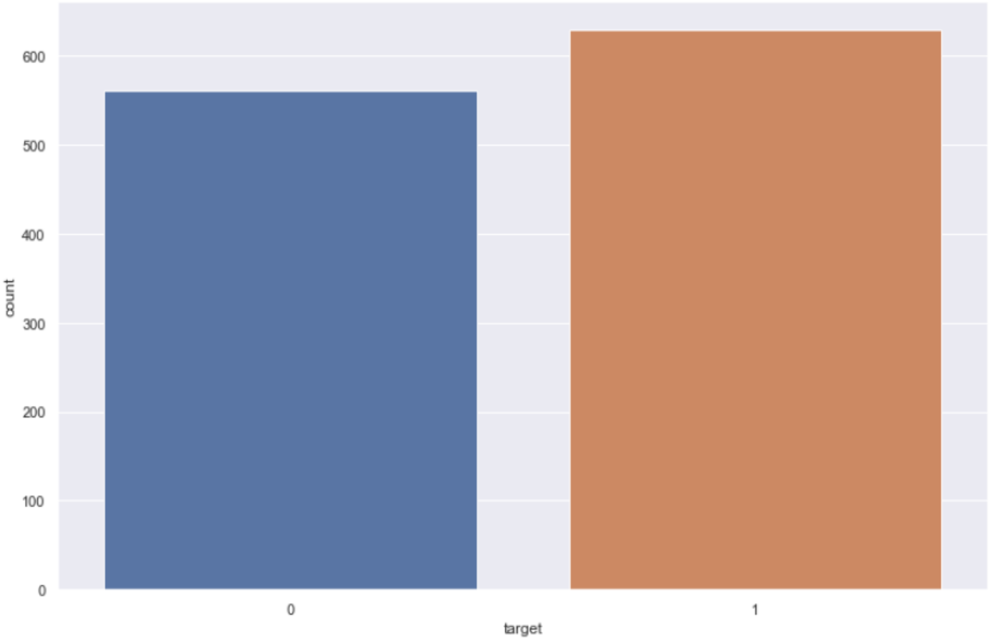


Figure 4.3: Count plot of target feature

From this plot figure we understand that the dataset is quite balanced. It also gives concept about the number of patients with or without heart disease.

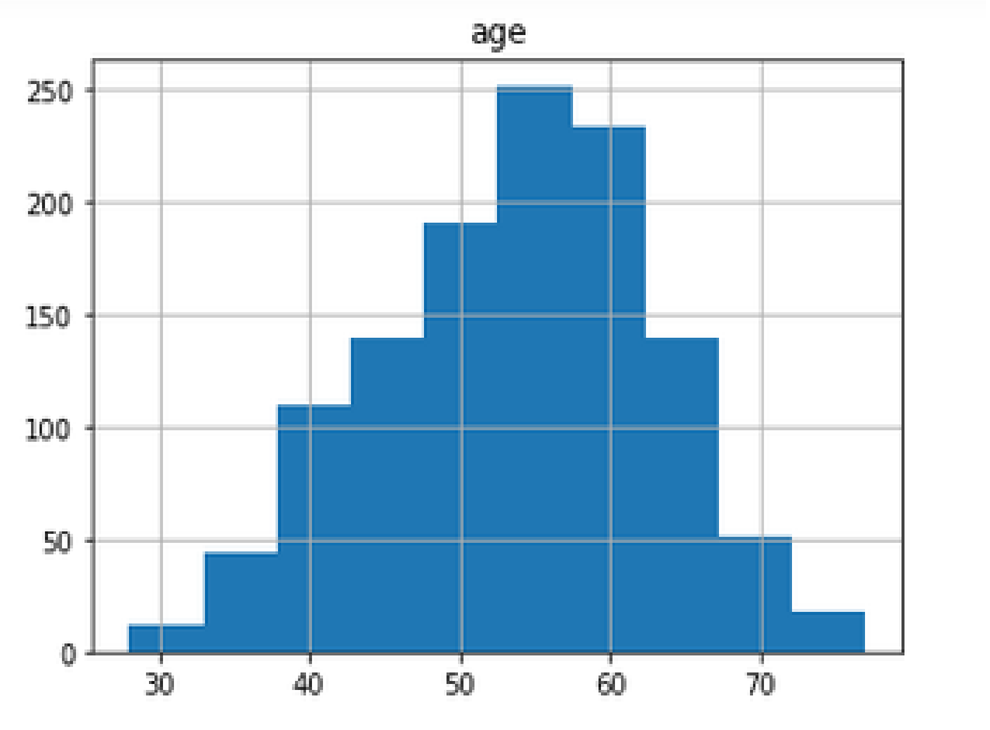


Figure 4.4: Histogram of age feature

Here, we also used another graphical display technique of the data by using histogram. This graphical form of features is helpful to give proper idea about the features that were used.

After analyzing the dataset another process can be used to represent the data graphically which is data visualization. Through visualization the data can be easily understand and explore the data quickly. So the visualization part is very essential for any predicting purpose project. The tools of visualizing provide an accessible way to see and understand trends, outliers, and patterns in data.

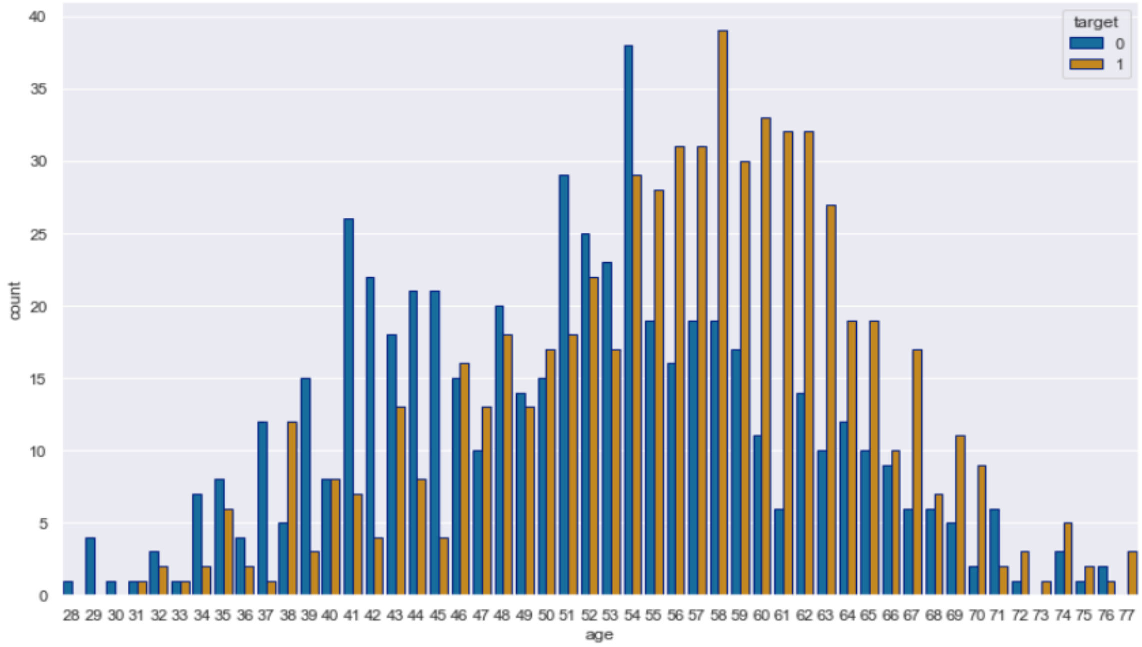


Figure 4.5: Count plot of relation between age and target feature

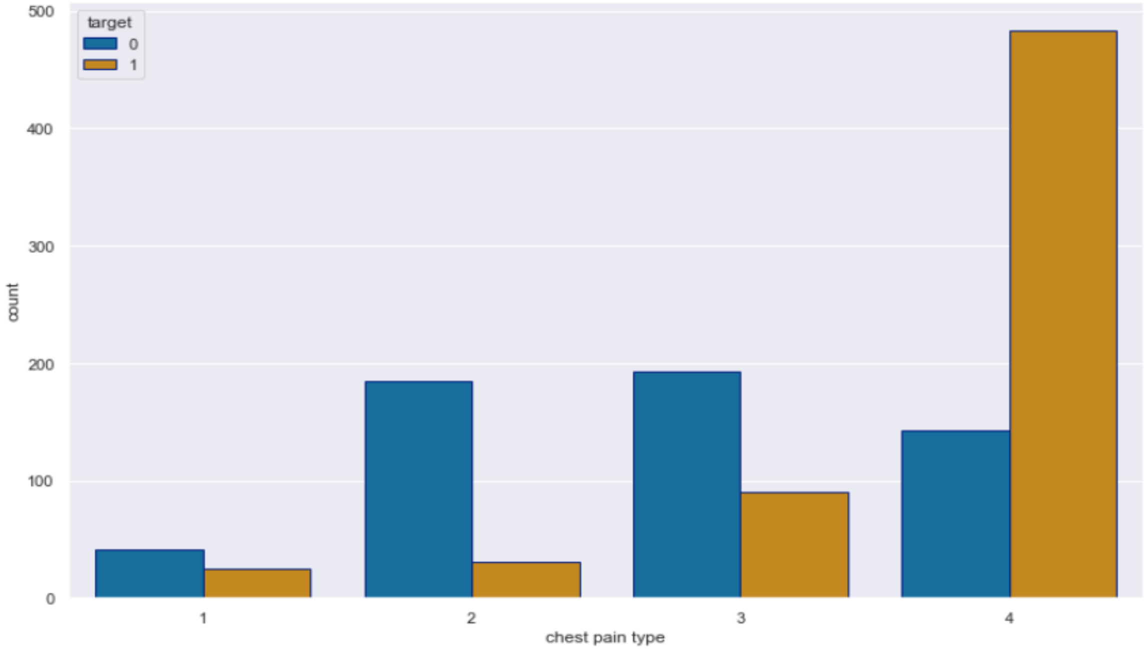


Figure 4.6: Count plot of relation between chest pain type and target feature

Here, the count plots provided by seaborn library represents the possibility of having heart disease according to the various features. From these graphical representation we can easily explore the data.

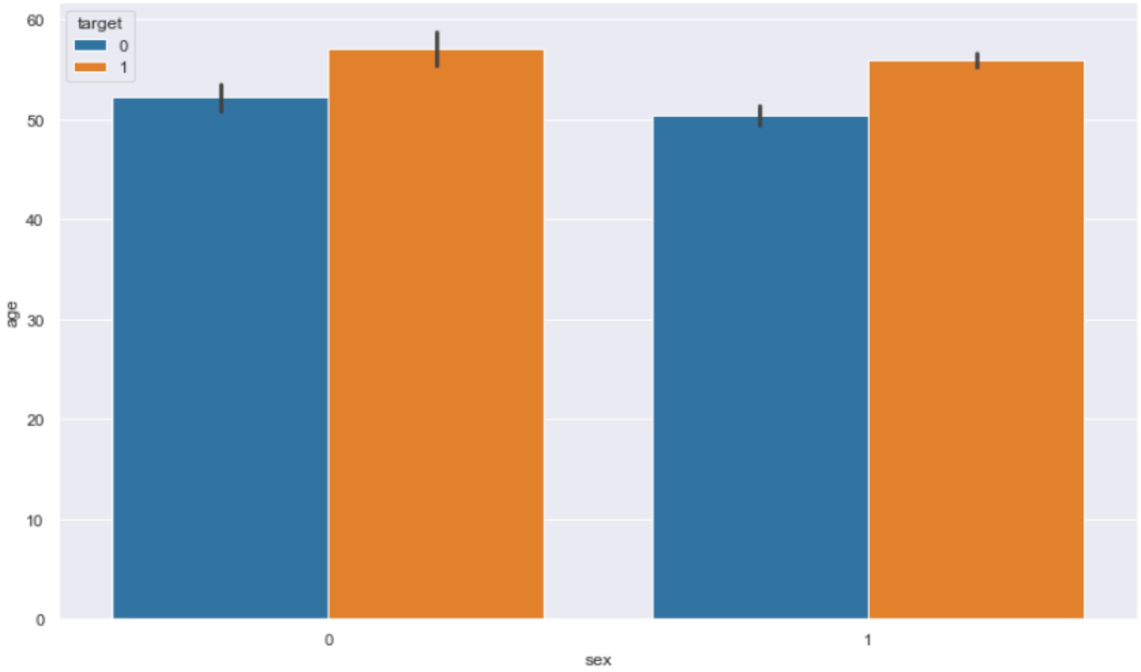


Figure 4.7: Bar plot of relation among age, sex and target features

Another data visualizing process is bar plot and it is also work with the help of seaborn library. By using these we have showed the relation among two features with target feature of the data set. Here the relation of two features (sex, age) values with the values of target feature is clearly understand through this bar plot.

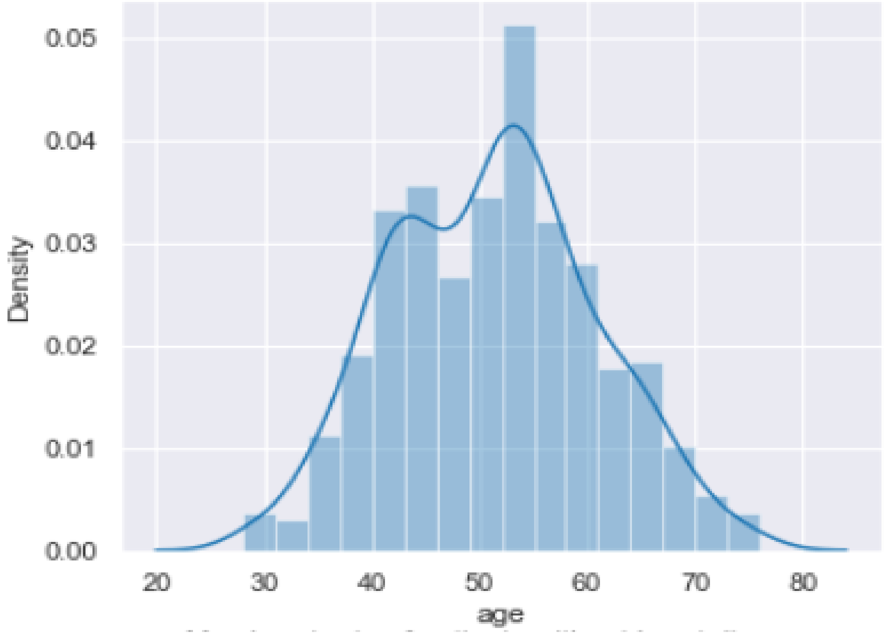


Figure 4.8: Age of patient without heart disease

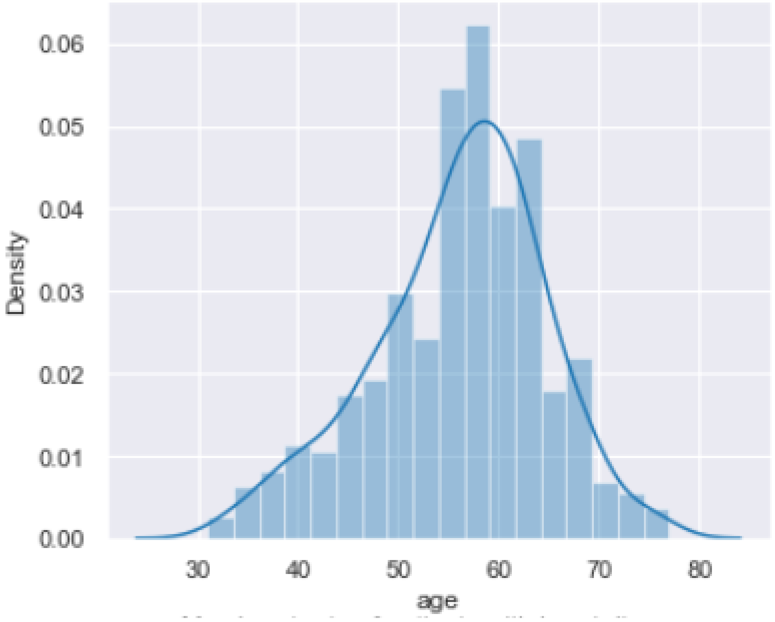


Figure 4.9: Age of patient with heart disease

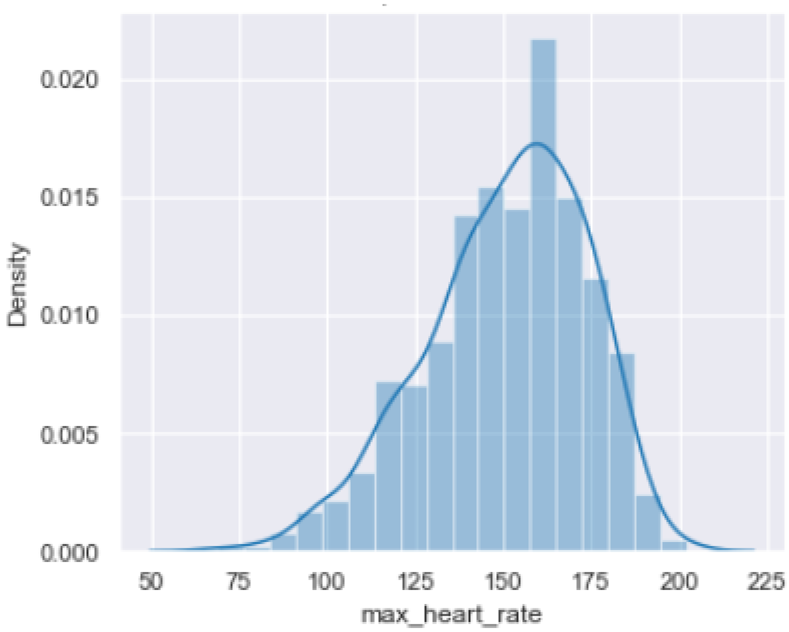


Figure 4.10: Max heart rate of patient without heart disease

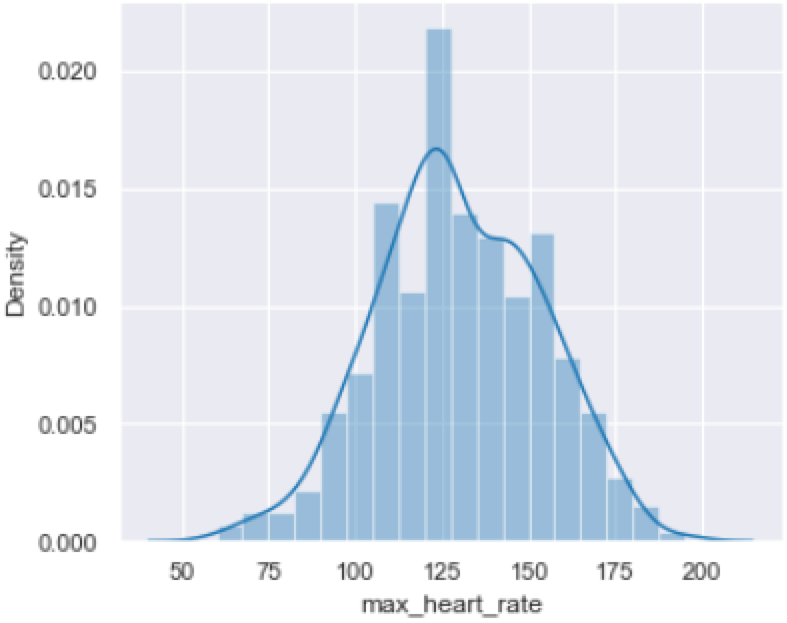
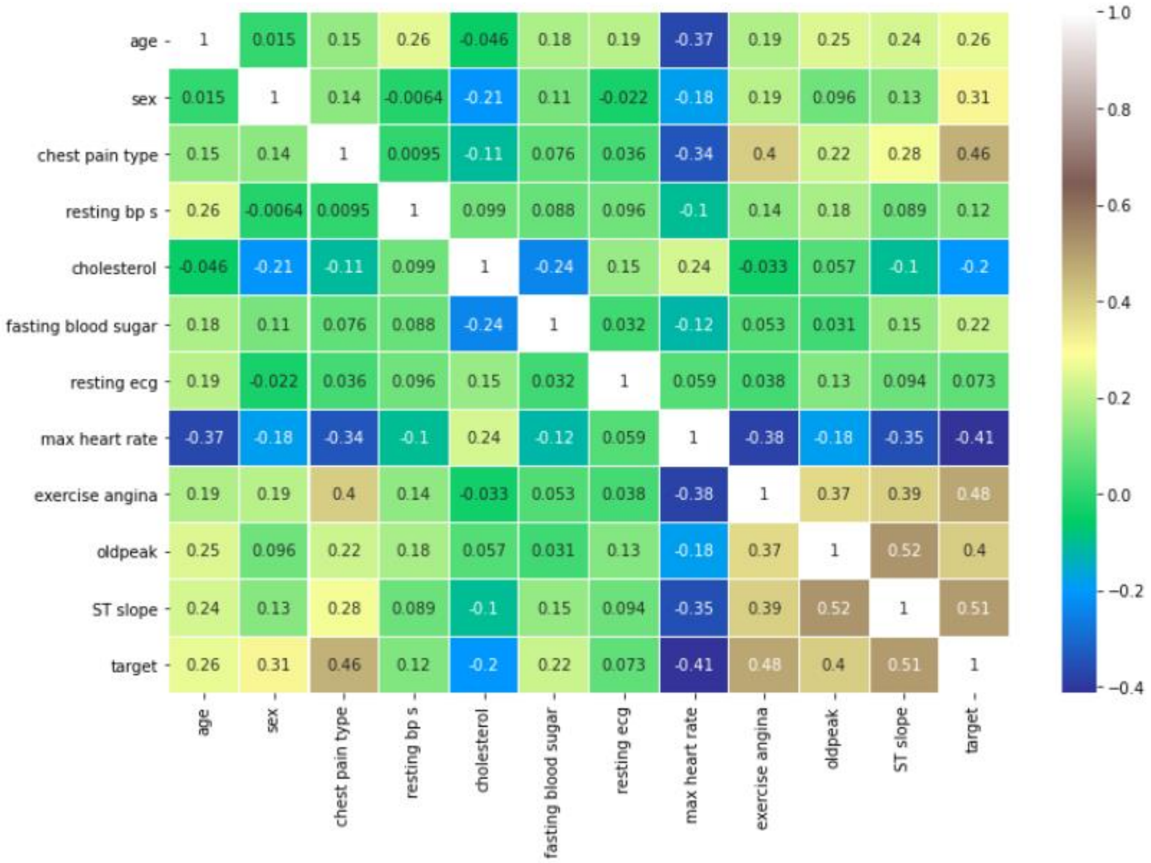


Figure 4.11: Max heart rate of patient with heart disease

The univariate distribution of the features are plotted by using dist plot graph. Distplot is another data visualizing tool of seaborn library which show a histogram with a line on it and we have used it to show the density of the values of features according to target feature that indicates the result of having heart disease or not.

After checking that the data is balanced we found out the correlation between the data that is helpful to understand the relation between any two features value. The heat map which provided by seaborn library is plotted the correlation between the features value that is used.



	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
age	1	0.015	0.15	0.26	-0.046	0.18	0.19	-0.37	0.19	0.25	0.24	0.26
sex	0.015	1	0.14	-0.0064	-0.21	0.11	-0.022	-0.18	0.19	0.096	0.13	0.31
chest pain type	0.15	0.14	1	0.0095	-0.11	0.076	0.036	-0.34	0.4	0.22	0.28	0.46
resting bp s	0.26	-0.0064	0.0095	1	0.099	0.088	0.096	-0.1	0.14	0.18	0.089	0.12
cholesterol	-0.046	-0.21	-0.11	0.099	1	-0.24	0.15	0.24	-0.033	0.057	-0.1	-0.2
fasting blood sugar	0.18	0.11	0.076	0.088	-0.24	1	0.032	-0.12	0.053	0.031	0.15	0.22
resting ecg	0.19	-0.022	0.036	0.096	0.15	0.032	1	0.059	0.038	0.13	0.094	0.073
max heart rate	-0.37	-0.18	-0.34	-0.1	0.24	-0.12	0.059	1	-0.38	-0.18	-0.35	-0.41
exercise angina	0.19	0.19	0.4	0.14	-0.033	0.053	0.038	-0.38	1	0.37	0.39	0.48
oldpeak	0.25	0.096	0.22	0.18	0.057	0.031	0.13	-0.18	0.37	1	0.52	0.4
ST slope	0.24	0.13	0.28	0.089	-0.1	0.15	0.094	-0.35	0.39	0.52	1	0.51
target	0.26	0.31	0.46	0.12	-0.2	0.22	0.073	-0.41	0.48	0.4	0.51	1

Figure 4.12: Correlation between features

The heat map clearly shows that the features like chest pain type, exercise angina, oldpeak and ST slope have positive correlation with the target feature.

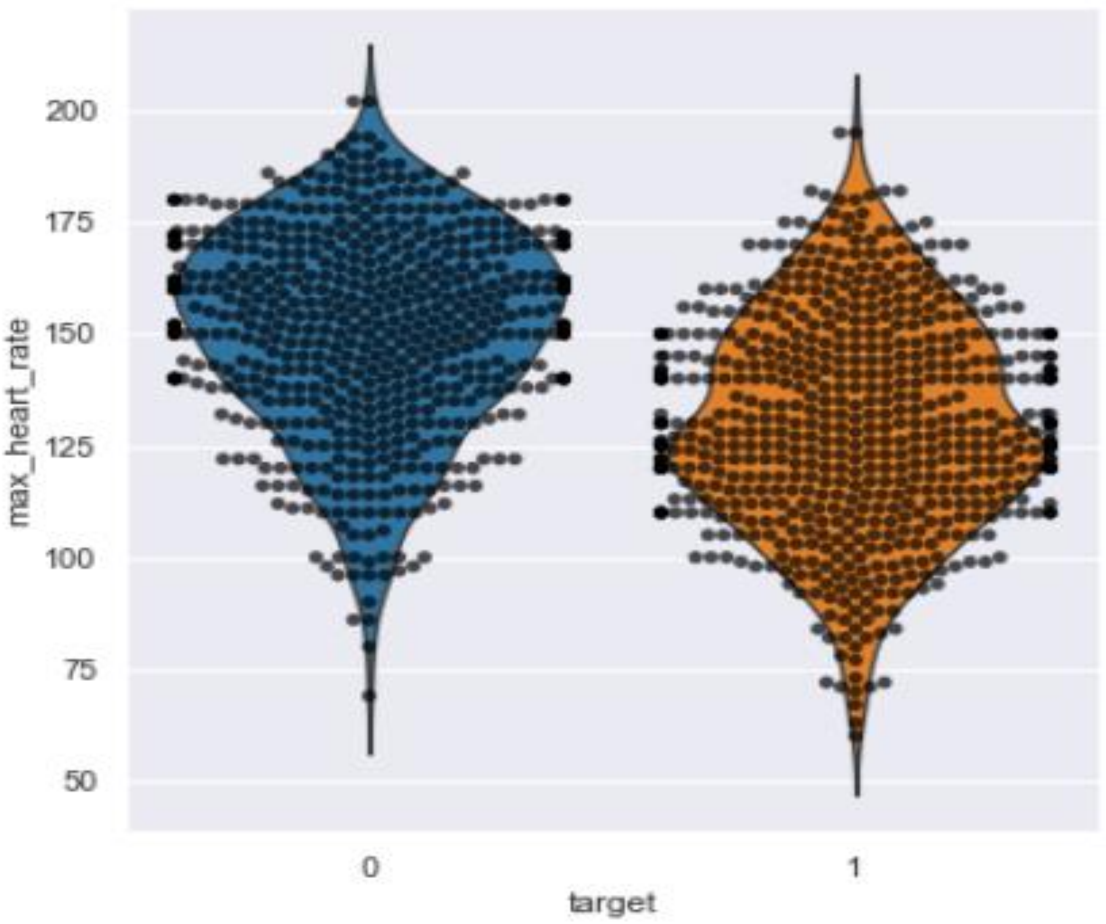
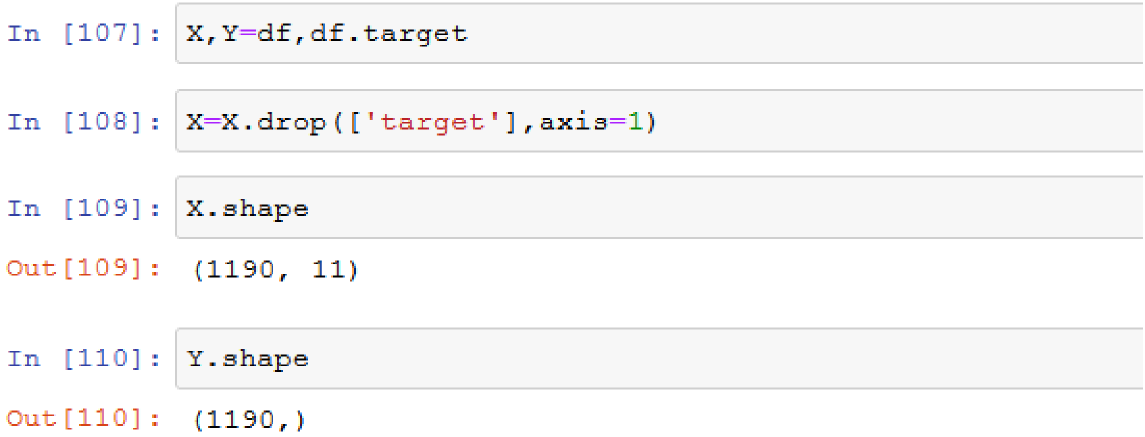


Figure 4.13: Violin plot of relation between max heart rate and target feature

Here, we also used some various plotting graph by using seaborn library such as violin plot that are effective for gathering proper idea about the features relation along with target values.

4.2.2 Data Splitting

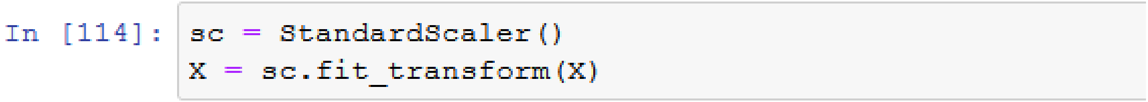
The data set are mainly used for two purposes. The big portion of data set is used for training the model and other is used for testing. So, for this purpose we have to split the data set into training and testing part. We are using the train_test_split module of sklearn for splitting the data set into training and testing part. As target variable indicates the result of having heart disease or not, so first of all we have to separate this target variable from the features of the data set. After separation, we have put them into X & Y variables.



```
In [107]: X,Y=df,df.target

In [108]: X=X.drop(['target'],axis=1)

In [109]: X.shape

Out[109]: (1190, 11)

In [110]: Y.shape

Out[110]: (1190,)
```

Figure 4.14: Shape of train & test dataset

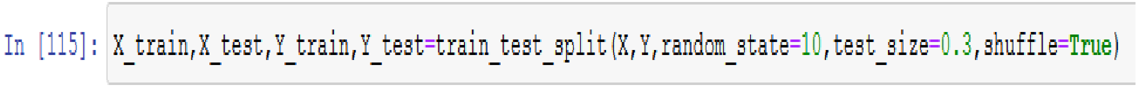
Here, X contains the features of data set without target variable and Y contains the only target variable.



```
In [114]: sc = StandardScaler()
          X = sc.fit_transform(X)
```

Figure 4.15: Standard Scaling

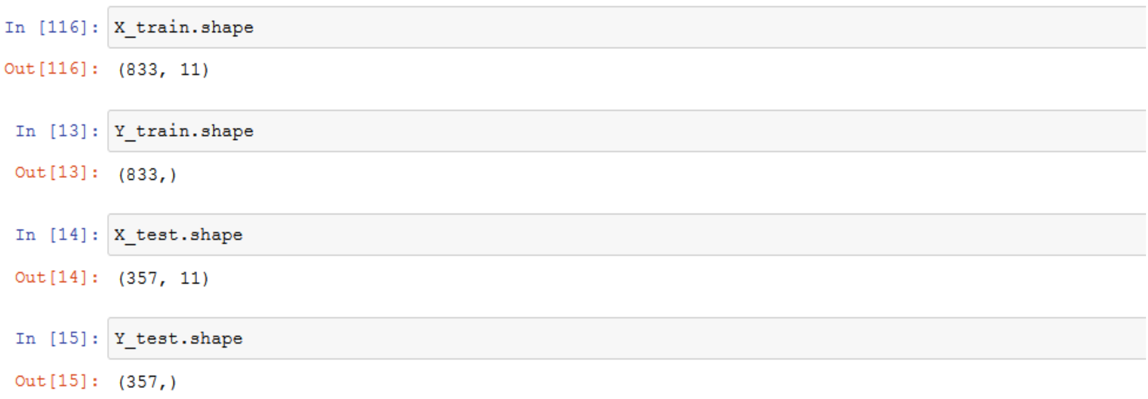
To gain better performance from machine learning algorithm the features need to be on a relatively similar scale. Standard Scaler is one of the most effective scikit-learn method which can preprocess the data to make it similar scale.



```
In [115]: X_train,X_test,Y_train,Y_test=train_test_split(X,Y,random_state=10,test_size=0.3,shuffle=True)
```

Figure 4.16: Splitting the dataset

Then we have split the data set for dividing it into training and testing part. Then we were splitting the data set in 70:30 ratio that means we have used 70% data for training our proposed methodology and 30% data used for testing. Random state variable has also passed as parameter of train_test_split module.



```
In [116]: X_train.shape
Out[116]: (833, 11)

In [13]: Y_train.shape
Out[13]: (833,)

In [14]: X_test.shape
Out[14]: (357, 11)

In [15]: Y_test.shape
Out[15]: (357,)
```

Figure 4.17: Dataset after splitting

After splitting X & Y into training and testing part, we have put them in X_train, X_test, Y_train and Y_test variables. Here, X_train, Y_train contain training data and X_test, Y_test contain test data.

4.3 Experimental Results & Analysis

In our project we used our dataset and applied three machine learning algorithm Decision Tree, K-Nearest Neighbor and Support Vector Machine. The results have been obtained from the algorithms that we applied.

The accuracy of the algorithms can be calculated from the confusion matrix using the formula:

$$\textbf{Accuracy} = \{(TP + TN) / TP + FP + TN + FN)\} * 100$$

First of all we applied Decision Tree classifier algorithm and we didn’t pass any parameter for the classifier function. So, the algorithm automatically followed the Gini Index method for building the decision tree.

Table 4.1: The node considered as the root node and the gini for the root node

Root Node	ST slope
Root Creation Gini	0.499

After applying this algorithm we have obtained the confusion matrix as follows:

[[148 11]

[24 174]]

From the confusion matrix, we found the accuracy score of this algorithm and which is 90.196 %

Then the K-NN classifier algorithm has been applied on the dataset. In this algorithm we set 8 as the value of K because 8 was the expected value of K which gave the highest accuracy of this algorithm.

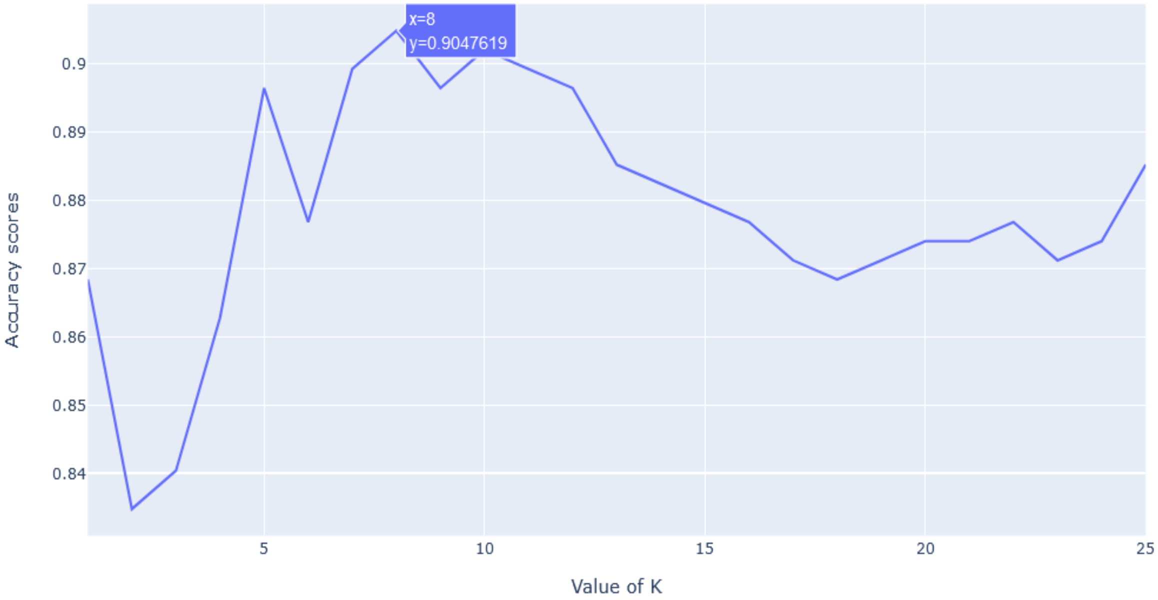


Figure 4.18: Accuracy score for different K values

The confusion matrix that was obtained from the algorithm is as follows:

$$\begin{matrix} [[144 & 15] \\ [22 & 176]] \end{matrix}$$

The accuracy score was found from the confusion matrix of this algorithm is 90.476 %.

After that, we applied another classification algorithm on our dataset which was Support Vector Machine. Here, we set Radial Basis Function (rbf) as kernel type and auto as gamma value for the parameters of this classifier function.

The confusion matrix which was obtained after applying the algorithm as follows:

$$\begin{matrix} [[141 & 18] \\ [16 & 182]] \end{matrix}$$

90.283 % accuracy was obtained from the confusion matrix of this algorithm.

The classification reports show the quality of prediction from these classification algorithm.

Table 4.2: Classification report of Decision Tree algorithm

	precision	recall	f1-score	support
0	0.86	0.93	0.89	159
1	0.94	0.88	0.91	198
accuracy			0.90	357
macro avg.	0.90	0.90	0.90	357
weighted avg.	0.90	0.90	0.90	357

Table 4.3: Classification report of K-NN algorithm

	precision	recall	f1-score	support
0	0.87	0.91	0.89	159
1	0.92	0.89	0.90	198
accuracy			0.90	357
macro avg.	0.89	0.90	0.90	357
weighted avg.	0.90	0.90	0.90	357

Table 4.4: Classification report of SVM algorithm

	precision	recall	f1-score	support
0	0.90	0.89	0.89	159
1	0.91	0.92	0.91	198
accuracy			0.90	357
macro avg.	0.90	0.90	0.90	357
weighted avg.	0.90	0.90	0.90	357

Now, we are going to create a summary table that contains different accuracy which we were obtained from the algorithms.

Table 4.5: Accuracy scores of three algorithm

Algorithms	Decision Tree	K-NN	SVM
Accuracy score	90.196 %	90.476 %	90.283 %

The bar plot represents the accuracy score of three classifier algorithms graphically.

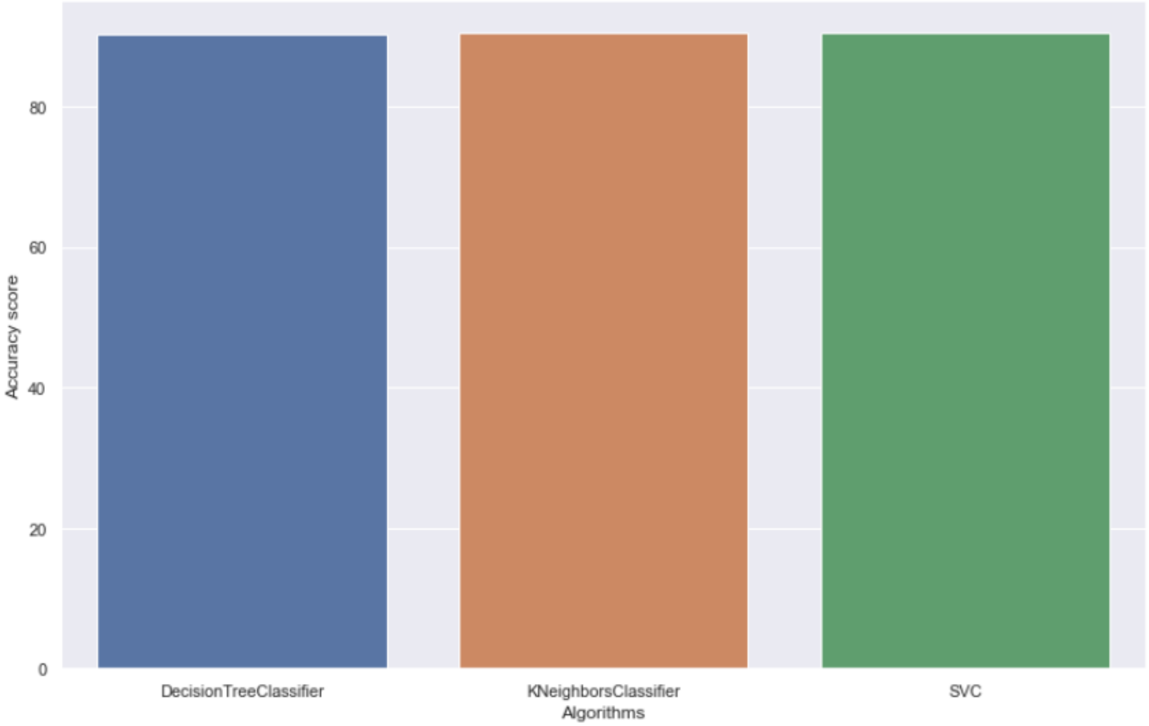


Figure 4.19: Accuracy score of three algorithms

Every algorithm has its own capacity to perform the best in its own favorable situation. But overall from Table 4.5, it is concluded that K-NN has got the best accuracy score among the three algorithms.

4.3.1 Feature Importance

A very basic question comes to our mind after applying the algorithms is what features have the maximum impact on predictions? The answer of the question can be found from the feature importance. Some features in the dataset don’t effect the prediction that much.

On the other hand few features play a big role in prediction. It can be changed the accuracy level of the model. So, the work only with important features may increase the accuracy of the model.

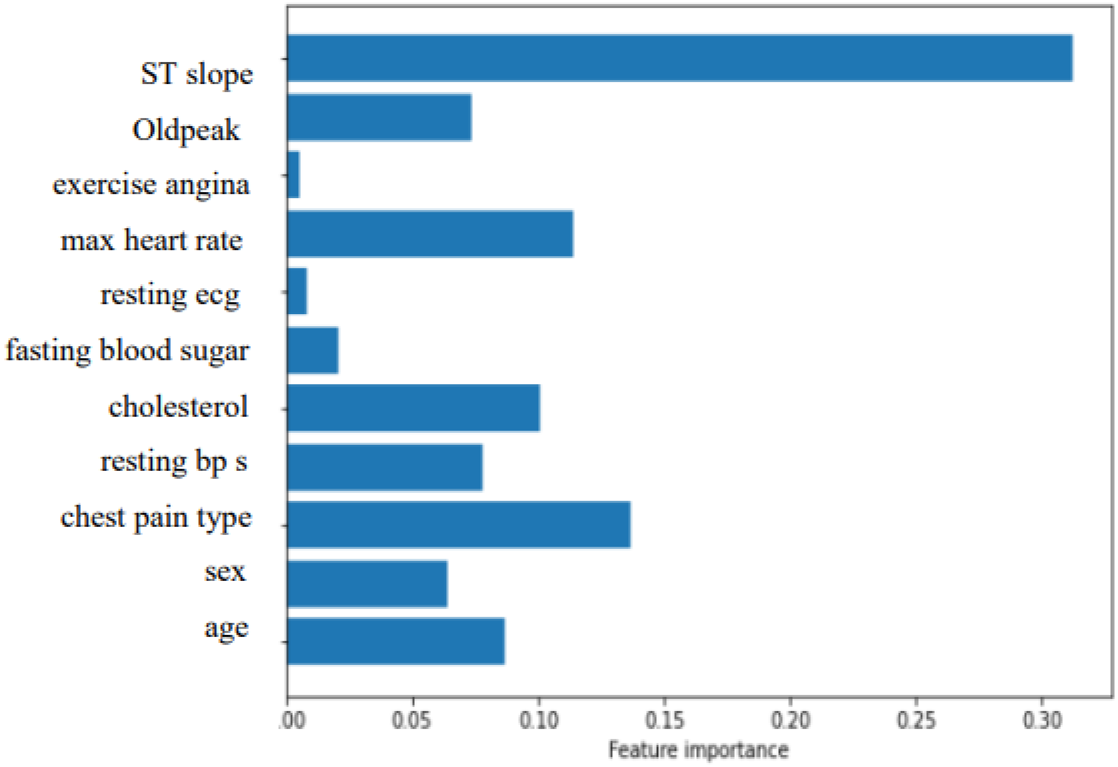


Figure 4.20: Decision Tree importance checker

4.4 Discussion

From these result we can see that the algorithms that we are used are very effective for this predicting purpose. The accuracy scores which were obtained from these algorithm were almost equal. The algorithms that we used are more accurate and helps us to save the money. We used Decision Tree which is an eager learner and K-NN which is lazy learner. Then we checked the accuracy of the two types of algorithm and found almost same accuracy. K-NN is a little more accurate than Decision Tree. We set 8 as value of K in K-NN algorithm for gaining highest accuracy. Complex algorithm like SVM also generate good accuracy in our project as like as basic ones. From this discussion we summarize that the algorithms that we used are very useful for this purpose.

CHAPTER 5

Impact on Society & Environment

5.1 Impact on Society

Bangladesh is a low-income country and most of the people in our society don’t have enough money for treatment of disease. Like other chronic disease, cardiovascular disease is one of the leading cause of unwanted death. There are a large number of people around us are suffering from heart related problem but can’t detect disease for high cost of medical service. The system we have made that predict heart disease with minimum cost and give almost accurate result. So, the people who couldn’t detect their problem of heart for excess cost of health service can easily detect their disease without more expense. People can get quality services at affordable costs with the help of these system. As our system was created with the help of appropriate computer-based data and decision support technique, so minimal cost of clinical test can easily achieved. For this the affected people in our society can save their valuable lives and the mortality rate will decrease day by day.

5.2 Impact on Environment

The data that we have used in this project was healthcare related. Patient data were included in it. The system have the ability to analyze the data and used data mining techniques to take clinical decisions. These clinical decision that support with computer-based patient records could improve the medical system by reducing medical errors, improving patient safety and outcome, decreasing unwanted practice variation. Data mining is very effective data modeling and analysis tool that have generated a knowledge –rich environment. It can enhance the quality of clinical decisions. This computer-based environment will take the place of the practice that take clinical decision based on doctors’ intuition and experience.

Besides, people can know the risk factors behind the heart disease through this system such as arsenic contamination in water, food-staff, air pollution. So, we can increase our consciousness about the factors of environment that are responsible for heart related disease.

5.3 Ethical Aspects

- Not to harm any patient.
- Privacy of personal data,
- Not to show genetic discrimination.
- Fairness in research design.
- Take the responsibility of research result.

Chapter 6

Summary, Conclusion & Implication for Future Research

6.1 Summary of the Study

Our research work is classical machine learning based. Our goal was predicting heart disease by using some machine learning algorithms. For this purpose first of all we collected a dataset from online. After collecting the dataset we analyzed it very well. We also visualized the dataset that represents the data graphically and it is useful for better understanding the data. Then, we preprocessed the data for applying various classifier algorithm on it. Three classifier algorithms have been applied on the dataset. The algorithms that were applied were Decision tree, K-NN and SVM. After applying these algorithms we have obtained different accuracy scores of the algorithms. After this we analyzed the results that we have gained.

6.2 Conclusions

This project has been developed to predict cardiovascular disease analyzing the patient medical history from the dataset. Medical history of patient such as max heart rate, cholesterol, chest pain type, resting bp s etc. which are lead to fatal heart disease. For prediction purpose we choose three machine learning algorithms i.e. Decision tree, K-NN and SVM. After applying these algorithms, it can be said that Machine learning is extremely effective in predicting heart related disease. More and more development should be done in this field of machine learning. The algorithms that we have used show excellent performance with the feature of our dataset. The conclusion can be finally drawn that the project that we are created using machine learning can be saved valuable lives by predicting heart disease.

6.3 Implication for Further Study

Efficient data mining on our dataset played the vital role to predict heart diseases. We have applied three classifier algorithms and got satisfactory results. In future we will try to evaluate the efficiency of these algorithms and use some others algorithm for finding better accuracy. We will also try to collect the data from hospitals of Bangladesh and use it as our dataset. It will be helpful for our people.

Anyone can used these data mining technique to build a software to predict the heart disease easily. This software would be useful for saving the valuable lives of heart disease patients.

References

[1] WHO, “Cardiovascular diseases (CVDs),” Published by WHO, 2013. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/. [Accessed: 30-March-2021].

[2] Devansh Shah, Samir Patel, Santosh Kumar Bharti. "Heart Disease Prediction using Machine Learning Techniques", SN Computer Science, 2020

[3] Apurv Garg, Bhartendu Sharma, Rijwan Khan. "Heart disease prediction using machine learning techniques", IOP Conference Series: Materials Science and Engineering, 2021

[4] G. Karthiga, C. Preethi, and R. D. H. Devi, “Heart Disease Analysis System Using Data,” vol. 3, no. 3, pp. 3101–3105, 2014.

[5] GeeksforGeeks, available at << https://www.geeksforgeeks.org/>>, last accessed on 12-04-2021 at 11:00 PM.

[6] Research and Reviews - International Journals, available at <<https://rroij.com>>, last accessed on 17-04-2021 at 10:30 PM.

[7] S. Gupta, D. Kumar, and A. Sharma, “DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR,” vol. 2, no. 2, pp. 188–195, 2011.

[8] J. Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction,” vol. 17, no. 8, pp. 43–48, 2011.

[9] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," 2008 IEEE/ACS International Conference on Computer Systems and Applications, Doha, 2008, pp. 108-115.

[10] Sibo Prasad Patro, Neelamadhab Padhy, Dukuru Chiranjevi. "Ambient assisted living predictive model for cardiovascular disease prediction using supervised learning", Evolutionary Intelligence, 2020

[11] Adi Purnomo, Mula Agung Barata, Moch Arief Soeleman, Farrikh Alzami. "Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease", Journal of Physics: Conference Series, 2020

[12] Dove Medical Press - Open Access Publisher of Medical Journals, available at <<https://dovepress.com>>, last accessed on 17-04-2021 at 10:30 PM.

[13] M. Akhil jabbar, B.L. Deekshatulu, Priti Chandra. "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", Procedia Technology, 2013

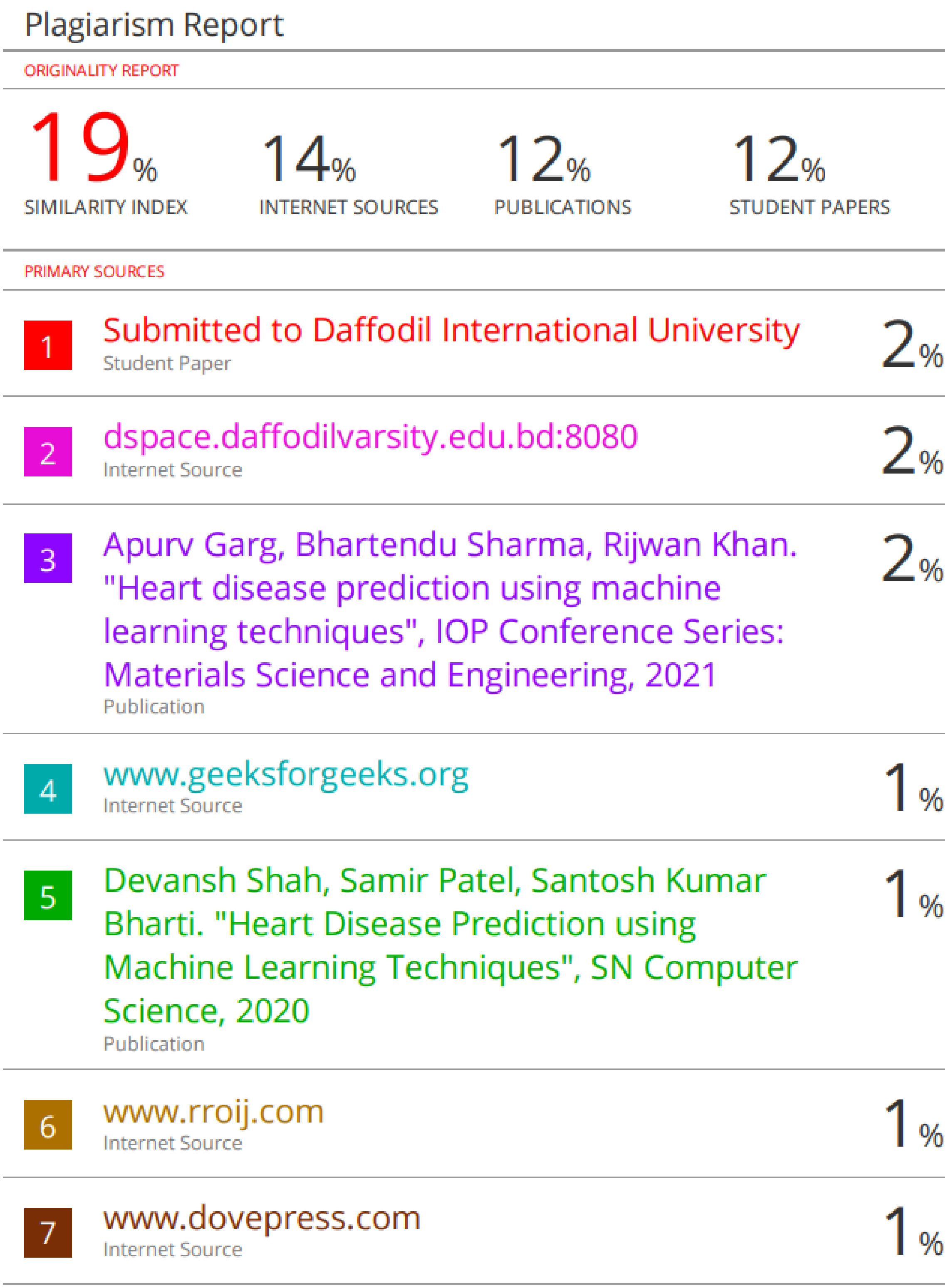
[14] K.Shekar, N.Deepika and D.Sujatha,“Association rule for classification of heart-attack patients”, International Journal of Advanced Engineering Sciences and Technologies, vol.11, no. 2, pp.253-257, 2011.

[15] M. Anbarasi, E. Anupriya and N.Iyengar, “Enhanced prediction of heart disease with feature subset selection using Genetic algorithm”, International Journal of Engineering Science and Technology vol.2, pp.5370- 5376, 2010.

[16] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol. 8, No. I (January - June 2007).

[17] Kiyong Noh, Heon Gyu Lee, Ho-Sun Shon, Bum Ju Lee, and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Springer 2006,Vol:345, page no. 721- 727.

Plagiarism Report:



Plagiarism Report

ORIGINALITY REPORT

19%	14%	12%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Apurv Garg, Bhartendu Sharma, Rijwan Khan. "Heart disease prediction using machine learning techniques", IOP Conference Series: Materials Science and Engineering, 2021 Publication	2%
4	www.geeksforgeeks.org Internet Source	1%
5	Devansh Shah, Samir Patel, Santosh Kumar Bharti. "Heart Disease Prediction using Machine Learning Techniques", SN Computer Science, 2020 Publication	1%
6	www.rroij.com Internet Source	1%
7	www.dovepress.com Internet Source	1%