Major project

Project report on

"A Comparative Study of various ML Techniques for Heart Disease Prediction"

Submitted By

Munish Kumar

(2K19/SWE/19)

MASTER OF TECHNOLOY

IN

SOFTWARE ENGINEERING

Under the supervision
of

DR. ABHILASHA SHARMA
Assistant Professor
Department of Software Engineering
DTU, Delhi

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
Bawana Road, Delhi-110042

JUNE, 2021

DECLARATION

I, Munish Kumar, 2K19/SWE/19 student of MTech (SWE), hereby declare that the project entitled **"A Comparative Study of Various ML Techniques for Heart Disease Prediction"** which is submitted by me to the department of Software Engineering, Delhi Technological University, Delhi in limited achievement of the requirement for the award of the degree of Master of Technology in Software Engineering, has not been previously formed the groundwork for any completion of the requirement in any degree or other similar title or recognition.

This report is a legitimate documentation of my study carried out during my degree under the guidance of Dr. Abhilasha Sharma.

Munish Kumar

Place: Delhi **Munish Kumar**

Date: 26th June, 2021 **(2K19/SWE/19)**

CERTIFICATE

I hereby certify that the project entitled **"A Comparative Study of Various ML Techniques for Heart Disease Prediction"** which is submitted by Munish Kumar (2K19/SWE/19) to the department of Software Engineering, Delhi Technological University, Delhi in partial achievement of the requirement for the award of the degree of Master of Technology in Software Engineering, is an evidence of the project study carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any degree or diploma to this university or elsewhere.

Place: Delhi

Date:26th June,2021

Dr. Abhilasha Sharma

Supervisor

Assistant Professor

Dept. of Software Engineering

ABSTRACT

Heart is one in every of the most important organs that has a lot of precedence in flesh. It provides the blood to any or all organs of the entire body by pumping it. Heart condition may be a prime root of death within the world. A large quantity of information is collected in medical business associated with heart condition. However, this knowledge isn't mined properly. Prediction of heart diseases in care field is critical work. Several researchers have already been operating within the field of heart condition prediction exploitation some machine learning algorithms. The results of analysis vary from dataset to dataset. Advanced machine learning models area unit accustomed discover data in information and for medical analysis, significantly in heart condition prediction.

In this research, the analysis of predictive frameworks is finished for cardiopathy utilizing a bigger range of input attributes. We've applied the various ML classification techniques for the detection of similar designs and elements within the Cleveland information from the UCI Machine Learning Repository utilizing python information manipulation applications. For the prediction of the presence of cardiopathy in a person, we've applied six ML classification techniques specifically SVM, KNN, NB, RF, LR and DT.

CONTENTS

List of Figures

List of Tables

List of Charts

List of abbreviations

Abbreviations	**Full Form**
AUC	Area Under ROC Curve
ROC	Receiver Operating Characteristics
KNN	K- Nearest Neighbour
DT	Decision Tree
RF	Random Forest
FN	False Negative
FP	False Positive
LR	Logistic Regression
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
CS	Clinical Support
QA	Quality Assistance
ML	Machine Learning
WEKA	Waikato Environment for Knowledge Analysis
UCI	Union Cycliste Internationale
BP	Blood Pressure
ECG	Electrocardiogram

ST	Segment
CP	Chest Pain
FBS	Fasting Blood Sugar

CHAPTER 1: INTRODUCTION

1.1 General

The providing the quality assistance at cheap prices is a major problem for health institutions (base hospital, health centers). Quality assistance means determining patients accurately and supervising the productive treatments. Disastrous consequences may occur due to poor clinical managements which are therefore intolerable. These clinical arrangements are mostly made based on doctor's instincts and experience rather than on the skills present in the data. Due to these practices causes undesirable partiality, mistakes and immoderate medical amounts. Quality of assistance, gets affected by this, provided to patients. The clinical arrangement help is integrated with computerized patient data. It could be helpful in lowering medical mistakes, increasing patient safety, reducing undesirable practice changes and upgrading patient outcomes.

Data mining is the computerized process. It is used for withdrawing useful data from huge places of data warehouses. Because of non-deterministic datasets from large size verification data mining is most helpful in an explorative analysis. It has big possibility of finding the designs in the datasets of the healthcare area. Further, the designs can be used for healthcare discovery. However, the present fresh clinical information should be gathered in order because it is widely distributed, voluminous and heterogeneous in nature. A medical information system can be formed with the unification of this collected information.

The data mining Appliances are helpful for prediction in healthcare, fraud detection, financial banking education etc. (as shown in Fig. 1). The data processing appliances are useful techniques for projecting the different diseases in the healthcare field.

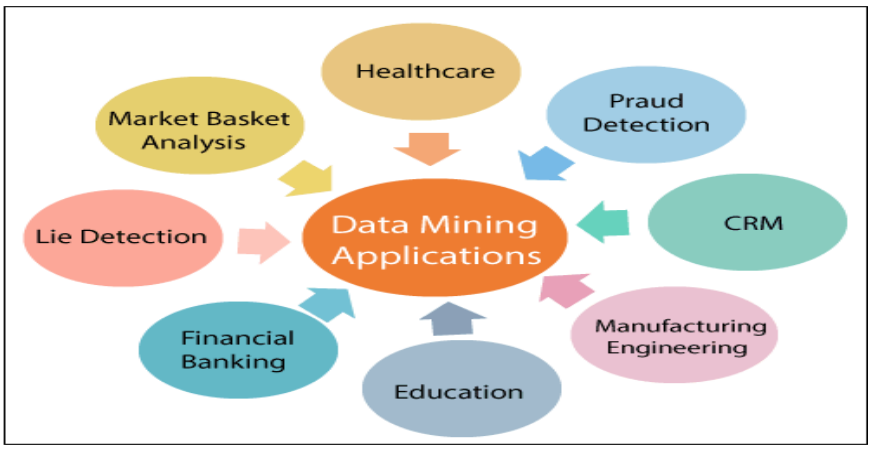


Fig. 1: Data mining Applications

Heart disease is one of the most murderous illness which can cause the decrease in life time of persons nowadays. Due to heart disease 17.5 million people lose their life each year. As heart is vital segment of humans, that is why we need the heart to function continuously to stay alive. Heart disease effects the working of heart. An analysis of possibility of heart disease in someone is essential for clinical advancement and medical treatment. A prediction framework can be derived by using various machine learning techniques. Medical management centers gather a large capacity of information in their data warehouse. The information is very compound and difficult to analyze. For the analysis of different data in medical centers machine learning algorithms play vital roles.

This research project mostly centered on designing a prediction framework for heart disease by using various data mining techniques. For the prediction of heart disease in the given dataset, we have applied six data mining techniques specifically decision tree, SVM, KNN, Naive Bayes, random forest and logistic regression.

1.2 Problem Statement

Information systems in many hospitals are designed to generation of simple statistics, inventory management and support patient billing. Largely health centers use management assistance frameworks which are restricted to small simple queries like the standard age of the patients, the number of operations required for a person to stay in health center for more than 10 days. However, the frameworks are unable to answer compound questions. On the basis of this problem following questions arises:

1. What is the current scenario of heart disease prediction?
2. What are the existing models available for machine learning?
3. Which datasets are available for heart disease prediction?
4. What is the accuracy achieved by a heart disease prediction system?
5. How can we reduce the time to upgrade the existing systems while maintaining the existing system?

1.3 Objective

The main purpose of this research project is to compare the various classification techniques for prediction of heart disease. Following are the objectives needed to be achieved for this project:

1. Gathering a custom dataset for training and validating the machine learning model.
2. The performance training and validation of machine learning models.
3. The performance comparison of different machine learning models.
4. Calculate the f1 score, precision, recall and ROC value of available machine learning models.
5. Comparison of the efficiency of the derived techniques with other models.

CHAPTER 2: RELATED WORK AND PROPOSED WORK

2.1 Related work

The prognosis of heart disease using various classification techniques is an ongoing process. In past plenty of research tasks has been finished on heart disease prognosis using several classification techniques such as KNN, DT, random forest, NB, neural networks, bagging and boosting etc. which is giving different accuracies on multiple datasets around the world. - Yan, Zheng et al. 2003;

Prognosis of heart disease using several regression frameworks is proposed. It demonstrates that to predict the heart disease chance Multiple Linear Regression is appropriate. The task is executed using training dataset of 13 different attributes which consists of 3000 instances as raised before. Then the data is split into two segments that is 30% of the records are utilized in testing and 70% applied for training. The outcomes specify that the accuracy of regression techniques is preferable than other techniques - K. Pola-Raju et al [1].

Megha Shahi et al, [12] using ML techniques proposed prognosis framework for heart disease. For automated identification of disease WEKA software used. This proposed framework gives standards of solutions in health departments. Several techniques like support vector machine, NB, Association rule, K-NN, ANN, and Decision Tree are applied in the framework. The outcome from the framework suggests SVM gives better results and more accuracy as compared with other machine learning algorithms.

Jaya-mi Patel et al. [2] suggested the heart disease prognosis using different machine learning algorithms. This research identifies different and unique designs with the help of the applications of machine learning methods. The J48 algorithms, provides the highest accuracy rate rather than Local Mean Time, determined by UCI record [3].

Sai-rabi H. Mujawar et al. [5] proposed analysis of NB and k-means for heart disease prognosis. The purpose of this project is to develop a framework utilizing the standard heart database which gives better detection. For gathering the data from the datasets using classification methods such as clustering can be used which is a specific data mining method.

2.2 Proposed work

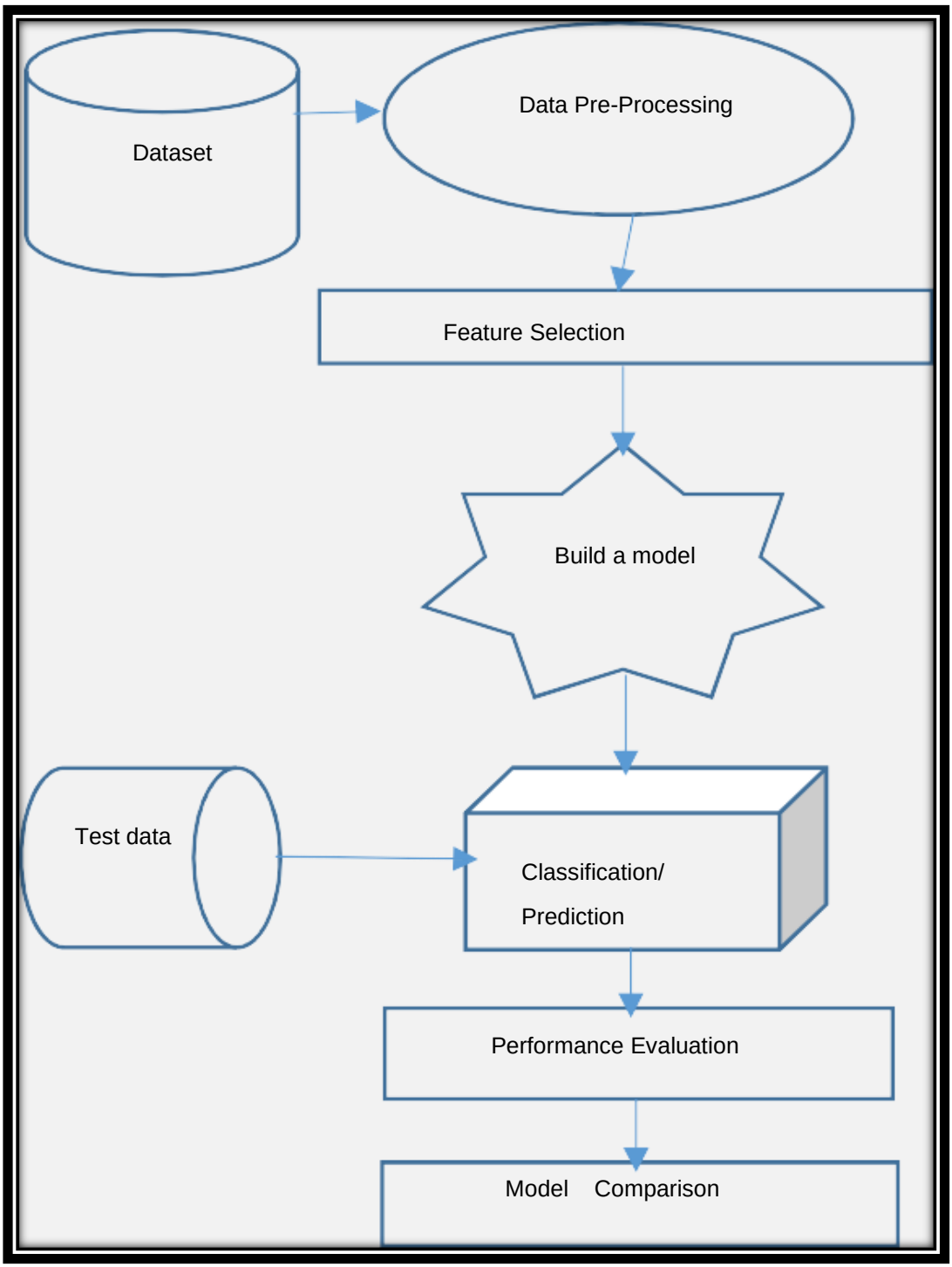


Fig 2: Proposed Model

CHAPTER 3: DATASET DISCRIPTION

The dataset is a type of multivariate data set intended to bring forth or entail a range of distinct arithmetic or statistical varying information, analysis of variable numeric information. The dataset is taken from Cleveland database consists of 14 attributes including age, gender, chest pain type, resting BP, serum-cholesterol, static ECG results, max-heart rate, angina, thalassemia, exercise-instigated chest, old-peak, ST exercise-instigated depression, ST slope, vessel count, and fasting blood glucose. The database comprises of 76 attributes, but most of research published till now have utilized only 14 of the attributes. The ML researchers use only the Cleveland database to date. The specific purpose of dataset is to anticipate whether the heart disease in the person is present or not and the other is the experimental function of assisting and finding several perspectives of the data-set which can assist better recognize the issue. The dataset was developed by:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Stein brunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Long Beach healthcare center
5. and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
97	52	1	0	108	233	1	1	147	0	0.1	2	3	3	1
70	54	1	2	120	258	0	0	147	0	0.4	1	0	3	1
82	60	0	2	102	318	0	1	160	0	0.0	2	1	2	1
162	41	1	1	120	157	0	1	182	0	0.0	2	0	2	1
189	41	1	0	110	172	0	0	158	0	0.0	2	0	3	0
285	46	1	0	140	311	0	1	120	1	1.8	1	2	3	0
86	68	1	2	118	277	0	1	151	0	1.0	2	1	3	1
141	43	1	0	115	303	0	1	181	0	1.2	1	0	2	1
95	53	1	0	142	226	0	0	111	1	0.0	2	0	3	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1

Fig. 3: Dataset Schema

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 303 entries, 97 to 197
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trestbps  303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalach   303 non-null    int64
 8   exang     303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64
 11  ca        303 non-null    int64
 12  thal      303 non-null    int64
 13  target    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 35.5 KB
```

Fig 3: - Description of attributes

```
Out[86]: age         0
         sex         0
         cp          0
         trestbps    0
         chol        0
         fbs         0
         restecg     0
         thalach     0
         exang       0
         oldpeak     0
         slope       0
         ca          0
         thal        0
         target      0
         dtype: int64
```

fig 4: check for null values

Table 1. Attributes of the heart disease dataset

Attribute	Representation	Information	Description
Age	Age	Integer	Age in years (0 to 99)
Sex	Sex	Integer	Gender instance (0 = Female, 1 = Male)
Chest-Pain-Type	Cp	Integer	Chest pain type (1: typical angina, 2: atypical angina, 3: non- anginal pain, 4: asymptomatic)
Rest-Blood-Pressure	Trestbps	Integer	Resting blood pressure in mm Hg
Serum-Cholesterol	Chol	Integer	Serum cholesterol in mg/dl
Fasting-Blood-Sugar	Fbs	Integer	Fasting blood sugar > 120 mg/dl (0 = False, 1= True)
Res-Electro-cardio-graphic	Restecg	Integer	Resting ECG results (0: definite left ventricular hypertrophy, 1: normal, 2: ST-T wave abnormality)
Max-Heart-Rate	Thalach	Integer	Maximum heart rate achieved
Exercise-Induced	Exang	Integer	Exercise induced angina (0: No, 1: Yes)
Old-peak	Oldpeak	Real	ST depression induced by exercise relative to rest
Slope	Slope	Integer	Slope of the peak exercise ST segment (2: up-sloping, 1: flat, 0: down-sloping)
Major-Vessels	Ca	Integer	Number of major vessels colored by fluoroscopy (values 0-3)
Thal	Thal	Integer	Defect types: value 0: null, 1: normal blood flow 2: fixed defect, 3: irreversible defect
Target	Target	Integer	Diagnosis of heart disease (0: disease, 1: no disease)

CHAPTER 4: DATA ANALYSIS

Feature Selection

4.1 Univariate Selection - Some features are picked by using statistical tests. These features may have the great relationship with the performance variable.

A certain quantity of features in a collection of various statistical tests are picked with the help of the SelectKBest class.

The chi-squared (chi2) statistical test for positive features where we choose the 13 best attributes in the given example.

	Specs	Score
7	thalach	188.320472
9	oldpeak	72.644253
11	ca	66.440765
2	cp	62.598098
8	exang	38.914377
4	chol	23.936394
0	age	23.286624
3	trestbps	14.823925
10	slope	9.804095
1	sex	7.576835
12	thal	5.791853
6	restecg	2.978271

Fig 6: Feature Score

4.2 Feature Importance — With the help of Model Characteristics property significance of each feature of given dataset.

For every function of the results a score is provided by feature value, the significance of the performance variable depends on the score of the specification.

Tree Based Classifiers provides the built-in class feature importance. The Extra Tree Classifier is used to take out the best attributes.

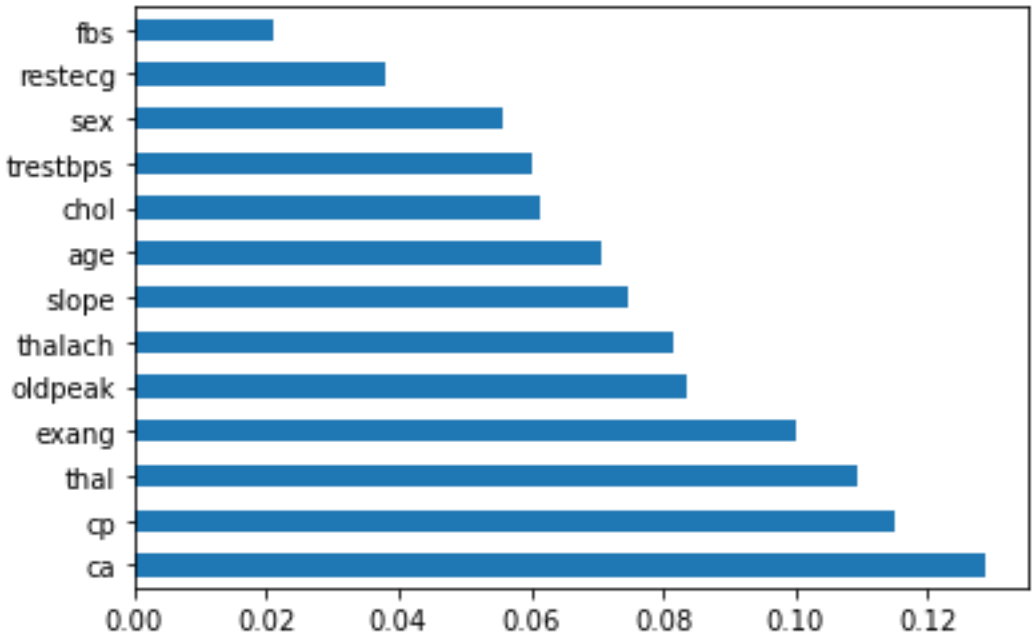


Fig. 7: Top features for the dataset

4.3 Correlation Matrix with Heatmap — Correlation matrix specifies the relation between the variables. It shows the correlation coefficient between them.

The correlation may be positive or negative. Classification of features whether these are most relevant or not, is made is by heat map and using the seaborn library the related attributes are plotted in the heat map.

Whether the characteristics are interrelated or the target variable is determined by the correlation. With the increase of a value, the value of the target variable decreases (positive correlation) and vice versa (negative correlation). From this heat map, it is clearly visible that the chest pain "cp" is strongly associated to the target attribute. In contrast to the connection between the other two attributes, it can be determined that chest pain hand out the most to the prognosis of the presence of heart disease. A heart attack is a medical crisis. Cardiac typically happens when blood motion to the heart is stopped due to blood clotting. Without the presence blood the tissues stop getting oxygen and die inflecting pain.

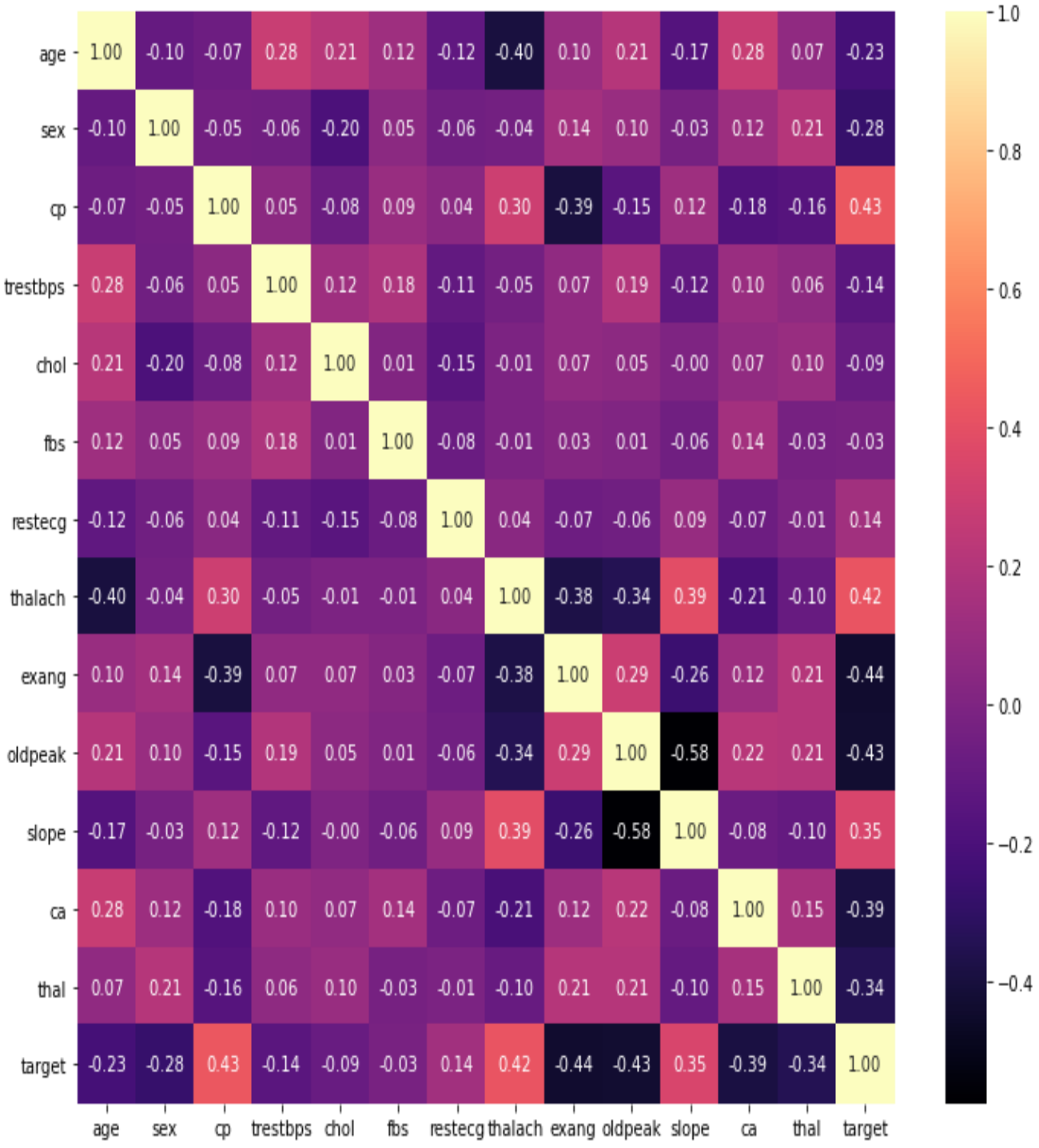


Fig 8: Correlation Matrix with heat Map

CHAPTER 5: DATA VISUALIZATION

5.1 Count plot

The Cleveland dataset specifies that men have higher chances getting heart problems than women. Women are affected less by heart arrest than men. About 70 to 90 percent of men have had cardiac arrest at some stage. When women experience heart attack, they usually have nausea or vomiting, which is frequently miscalculated for acid reflux or some kind of flu. They do not experience chest pressure while having heart attack.

Fig. 9: Distinguish Heart Disease according to gender

Chest pains are of 4 types: asymptomatic angina, atypical angina, windless pain, and special angina. Most of heart patients have symptomless chest pain. People of this category may have cp of type 1 which is called atypical angina such as hyperacidity, the infection, or tension in the muscles of the chest. In case of asymptomatic angina, the motion of blood to your heart gets blocked. The muscles of the heart get injured due to it. The possible causes for heart problems are close to that of asymptomatic attack.

Following are some elements which are the causing heart problems:

- Age
- Weight

- History of heart disease
- Cholesterol level
- Blood pressure level
- Smoking
- Previous heart attack
- Lack of exercise

The risk of another heart attack is increased due to an asymptomatic heart attack, which can be fatal. The issues such as heart failure have more chances of occurrence due to second heart attack. There no such test which can check whether someone has more chances of asymptomatic seizure. An echocardiogram can be the only procedure to check someone is experiencing asymptomatic seizure. This procedure can check the factors that are determining a heart attack.

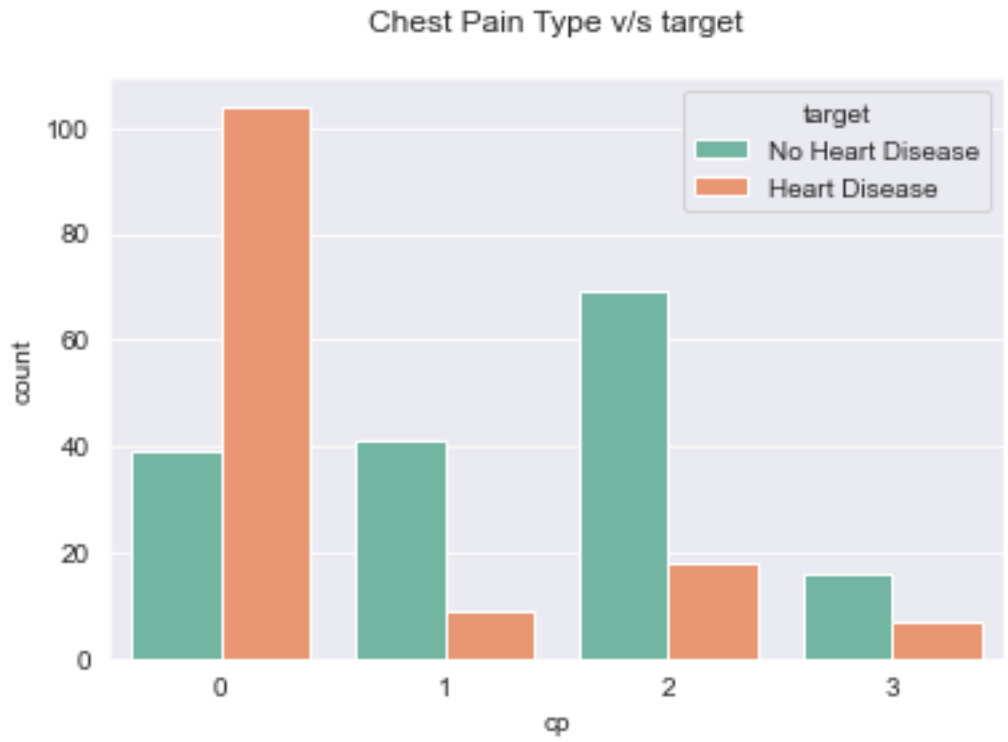


Fig 10: Chest Pain Type v/s heart disease

Beta-thalassemia heart condition is especially characterized by 2 differing kinds of phenotypes, one sort of dilation, altered left chamber dilation and ability and restricted constitution, with restricted left chamber sensation, respiratory organ cardiovascular disease and right chamber

failure. Heart issues, symptom failure, and abnormal heart rhythms will be related to severe hypochromic anemia.

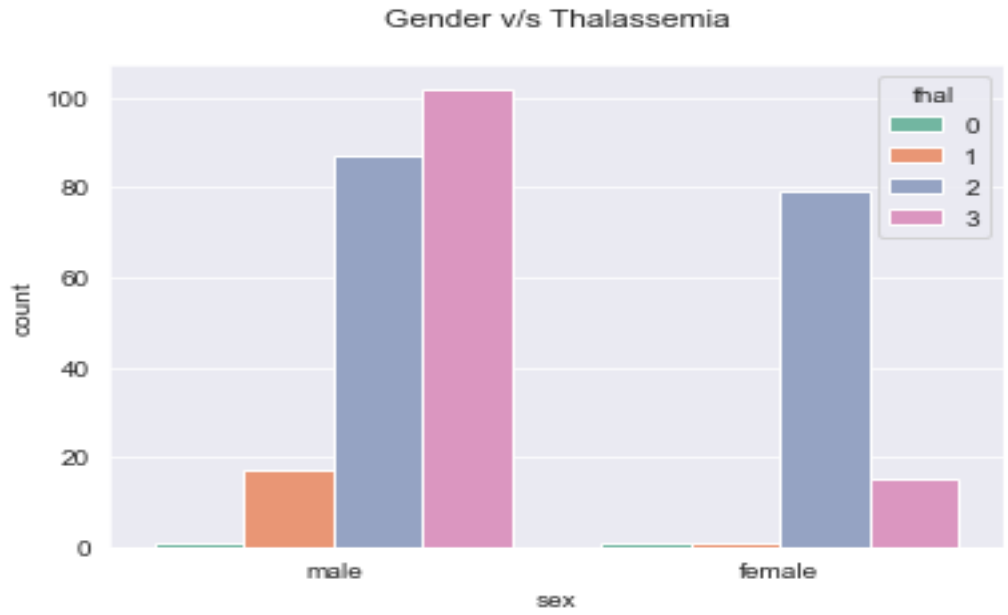


Fig. 11: Presence of Thalassemia in different gender

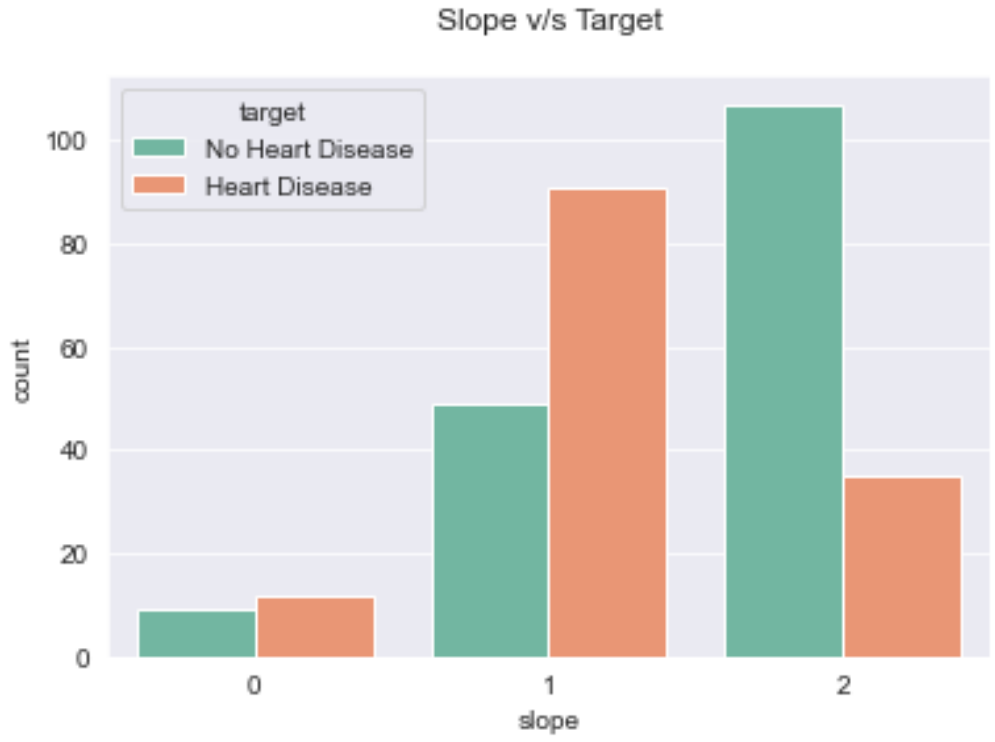


Fig. 12: Type of ST slope Present in Heart Disease

5.2 Histogram

Fig. 13: Age of Heart Disease Patients

People aged 60 and over and in adults between 41 and 60 years old the chances of heart disease are very high. But this is rare in 19–40-year-olds and very rare in 0–18-year-old.

5.3 Box plot/ Violin plot

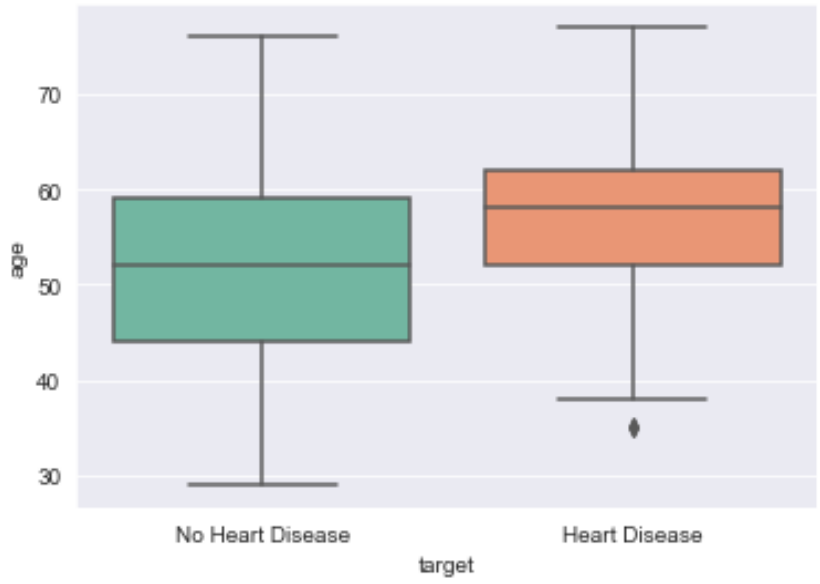


Fig 14: box plot of age

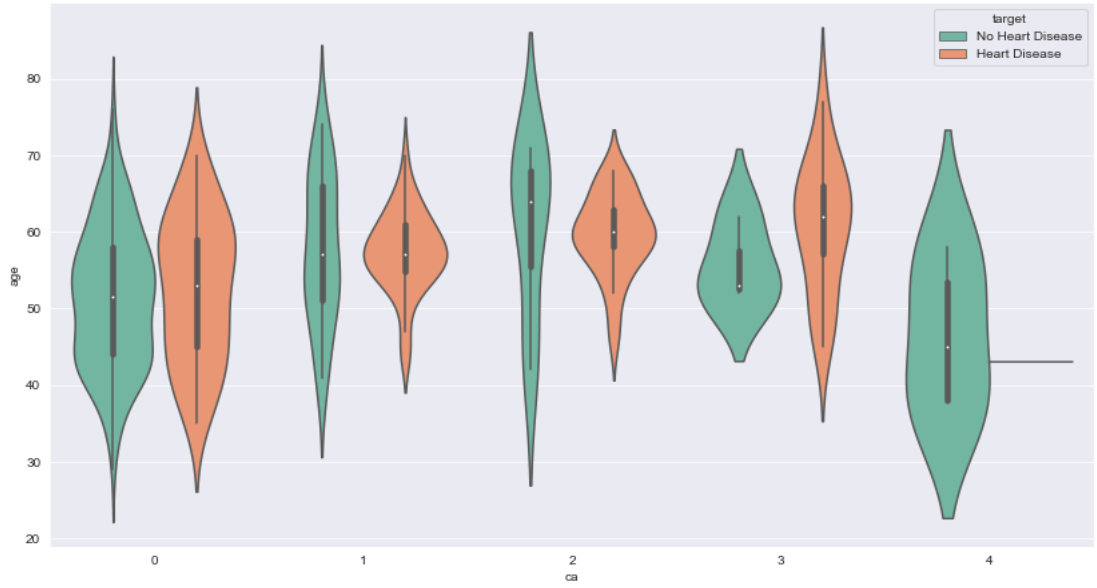


Fig 15: violin plot of age

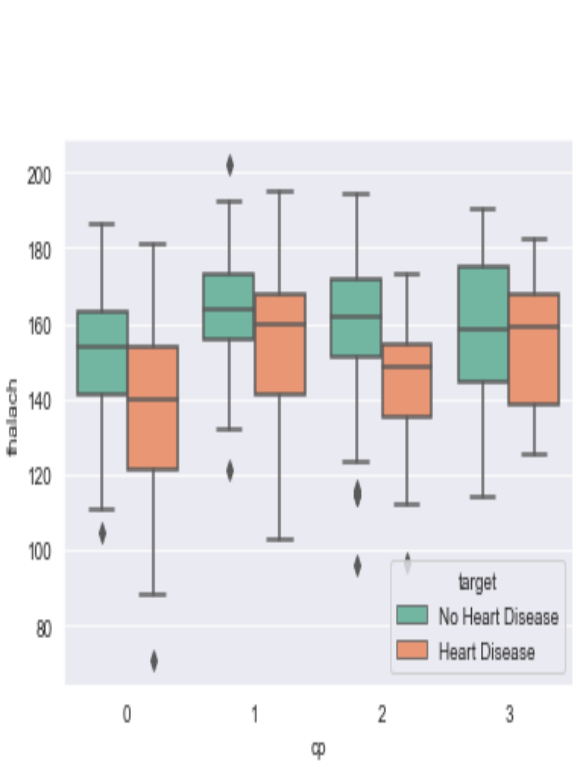


Fig 16: Box plot for thalach vs cp

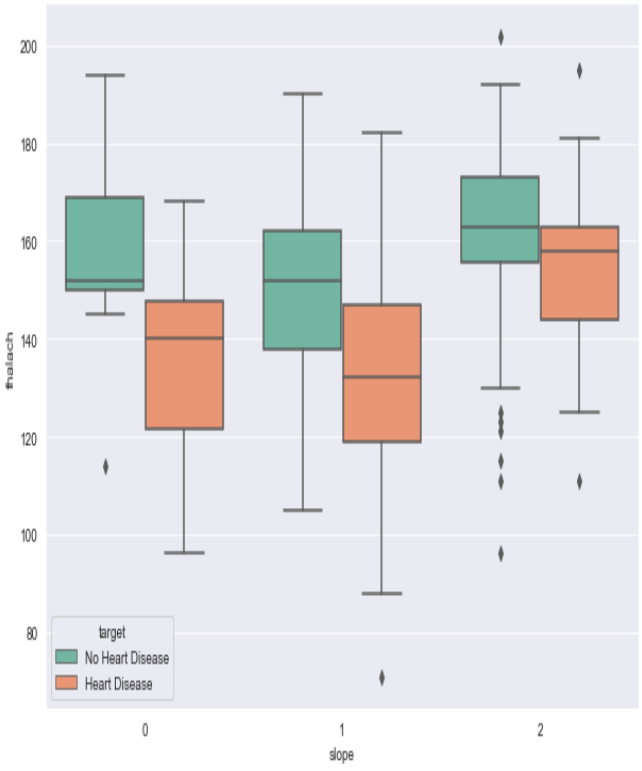


Fig 17: Box plot for thalach vs slope

5.4 Pair Plot

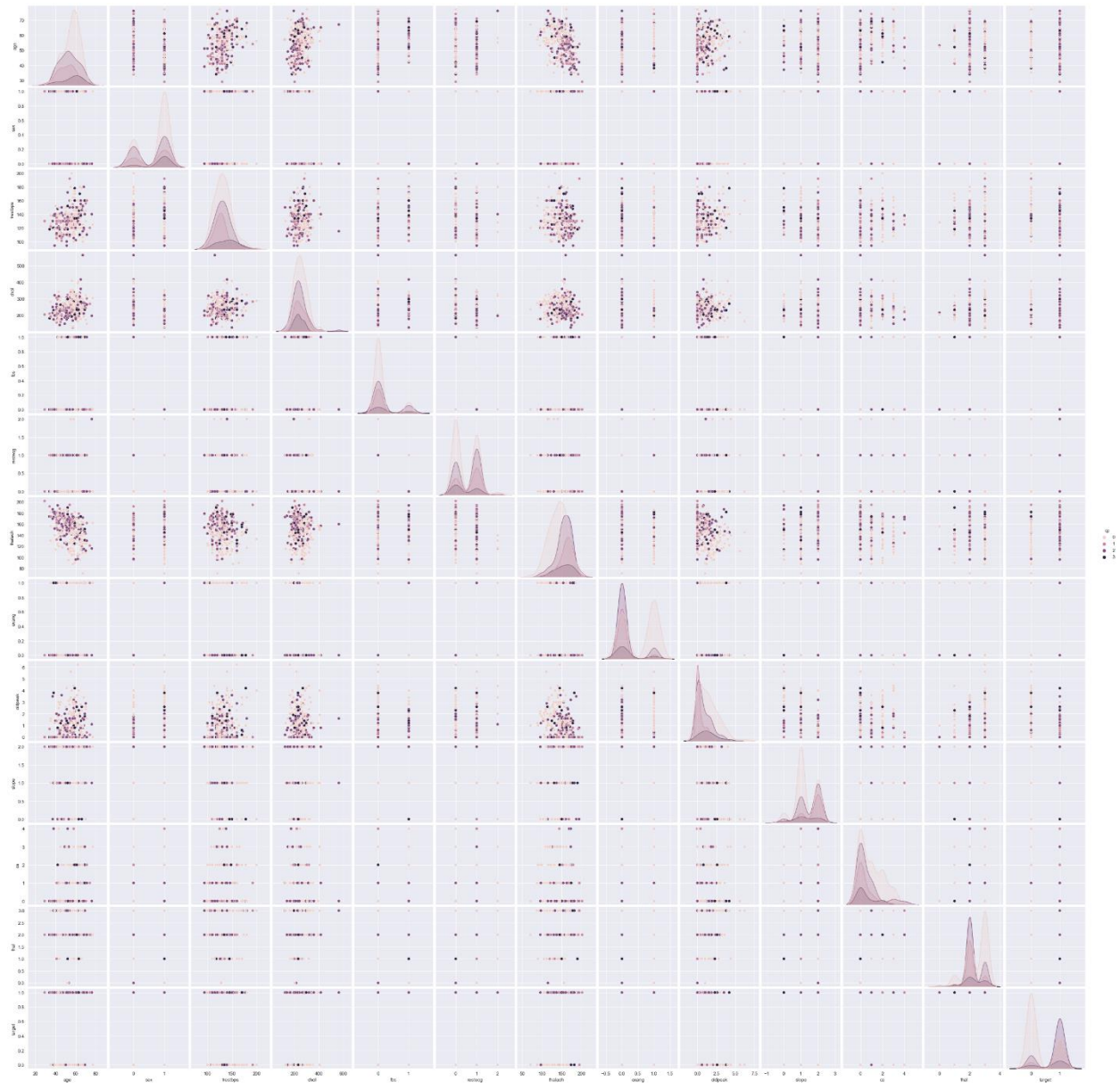


Fig 18: Pair plot for Chest pain

CHAPTER 6: CLASSIFICATION TECHNIQUES AND METRICS

6.1 Random Forest

The Random Forest consists of the many call trees. it's a collection classification. category results are delineated by individual trees. This comes from random forest results projected by Tin Kam Ho of Bell Labs in 1995 [9]. This technique is combined with random feature choice to construct a choice tree with controlled variation. Trees were made victimization the algorithmic rule as mentioned.

Let N be the amount of learning categories and M be number of attributes within the classification.

- The input variable m is employed to see the tree node.
- Select n learning times by work all N on the market learning cases by predicting the category, estimating the tree error.
- Select the variable m haphazardly for every tree node and calculate the most effective distribution.
- Finally, the tree has reached full growth and isn't cropped. The tree is pushed right down to predict a brand-new sample. once the terminal node ends, the tag is assigned a coaching sample. This procedure was performed repeatedly on all trees and was reported as a random forest forecast.

Random Forest (RF) is one such procedural case. RF as a multiple classifier formed from a choice tree in which each ht tree has been created from an information preparation group and the vector t numbers are arbitrarily distributed independently and independently of the vector. Each resulted tree is generated from a portion of random data sets. This uses a random vector generated from a number of variable probability distributions, in which the flow of probabilities is shifted to a central sample that is difficult to sort. Random vectors can be incorporated into the process of creating trees from a variety of views. The leaf centers of every tree area unit named with a back unfold rating on behalf of the data category. every internal hub contains the check that best matches the information house to be compiled. Another, hidden events area unit instructed by causing them to every tree and collection the incoming sends

- The random forest procedure has some desirable qualities, e.g.
- Not difficult to use, basic and easy to parallel.
- Doesn't need the model or parameter to pick out aside from the quantity of flags to be every which way hand-picked at every node.
- It works expeditiously on in-depth databases; it's sturdy enough against anomalies and agitation.
- It will manage an outsized variety of knowledge variables while not deleting the variables; it provides Associate in Nursing assessment of the vital variables within the classification.
- It has Associate in Nursing economical system for evaluating missing info and maintaining accuracy once most information is lost, it's ways for correcting errors in an exceedingly population of unequal information categories.

6.2 Decision Tree

A decision tree may be a decision support tool which utilizes a tree like a graph or a framework of outcomes and its potential results, together with the outcomes of fortified events, the resources, prices and services. this can be how to show the algorithmic rule.

Decision trees area unit usually employed in operational analysis, notably within the decision-making analysis, to assist establish a method presumably achieves the target, however it's conjointly achieved a preferred tool for data processing.

The decision tree may be a structure within the sort of a flow chart in every internal node represents the "test" on the attribute, of every branch represents the check results and every sheet node category prescript account. the trail from root to leaf describe the classification rules.

The basic algorithmic program for inducement decision tree is divide and conquer that build decision trees on a top-down basis strategies of division and perennial conquest. algorithmic program beginning with the complete row within the set of exercises, select the most effective attribute that offers the maximum amount info as potential for classification, and generate a take a look at node for this attribute. Then, inducement from prime to bottom of the results tree divides the present teams of tuples supported the values of their current take a look at attributes. The creation of the classifier comes to an end if all the tuples of a set talk over with constant class, or if it's unattainable to hold out an extra separation into alternative subsets, it's that's, if alternative

attribute tests solely provide info for classification below a antecedently such threshold. call tree algorithms generally use entropy-based measures referred to as "information acquisition" as a heuristic to pick out the attributes which will best separate the coaching information into distinct categories. This rule calculates the addition of knowledge for every attribute, and with every rotation, the one with the best info gain are elect because the check attribute for a given coaching information set. A felicitous split purpose can facilitate divide the info to the most effective limit. However, the most criterion within the greedy call tree approach is to make shorter trees. the most effective split purpose will be quickly assessed by considering every distinctive price for that characteristic within the given information as a potential split purpose and hard the gain from the relevant info.

6.3 Naïve Bayes

This classification supported Baye's theorem with assumption of freedom between characteristics. The theorem classifiers use applied arithmetic analysis to foresee future variables and are appropriate for giant knowledge sets. Rhythm algo two shows naive theorem classification to forecast future stage of condition of the cluster.

Baye's theorem gives how to calculate the probability P(c | x) from P (c), P (x) and

P (x | c).

$$P\left(\frac{c}{x}\right) = P\left(\frac{x}{c}\right) P(c) \Big/ P(x) \qquad (1)$$

where, P(c|x) - is the posterior probability of class (c, target) given predictor (x, Ai), where Ai = {A1, A2, …., A14}

P(c) - is the prior probability of class p(yes/no)

P(x|c) - is the likelihood which is the probability of predictor given class.

P(x) - is the prior probability of predictor.

6.4 KNN (K- Nearest Neighbor)

Neighbor-K is that the simplest supervised learning algorithmic program used for classification and regression issues. this is often a supervised learning algorithmic program, which implies that

the info should contain input and output parameters supported the model to be trained. The K-NN algorithmic program for a given worth of k can realize the closest k information. Then category is going to be appointed to a knowledge purpose supported the category of the most important cluster of information points having constant class. To sight the closest neighbor K, it uses geometer similarity or distance metrics. Here is that the formula for Euclidian distance:

$$d(x,y) = \sum_{j=1}^{k} \sqrt{(x_j - y_j)^2} \quad (4)$$

After this, data points are assigned to the class which has the highest probability. The probability can be represented as

$$P(y = j|X = x) = \frac{1}{k} \sum_{y \in A} I(y^i = j) \quad (5)$$

For regression issues the methodology is that the same however rather than neighboring categories it uses target values. one in every of the most important issues with KNN is selecting the right k. If k is smaller than the result limit, it'll be additional irregular and on the opposite hand, the next worth of k can end in a power tool result limit.

6.5 Logistic Regression

Logistic regression is employed in classification issues. The name logistic regression comes from the work it uses that is that the supply performs or the sigmoid function. The sigmoid perform takes associate degree input worth and maps it to a price between zero and one. The sigmoid perform is diagrammatic as follows

$$\text{sigmoid (value)} = \frac{1}{1+e^{-value}} \quad (6)$$

In this technique, information is classified consistent with whether or not the anticipated chance exceeds a average value or not. In this technique, call boundaries will be linear and nonlinear in nature, reckoning on the distribution of information points.

6.6 Support Vector Machine

The Support vector machine (SVM) is additionally a supervised machine learning formula. In classification and regression issues, this formula is often used. in a very support vector machine, information points area unit collected and portrayed in house. In SVM, the p-1 plane is employed to separate teams of knowledge that area unit thought-about teams of p vectors. This craft is understood as associate degree plane. the information purpose shouldn't fall terribly on the brink of the plane. it'll follow the most effective planes among all in line with the gap between the categories. A hyperplane of most margin is often outlined as a hyperplane that makes a most gap between categories. If there are a unit n information points, it is often expressed as

$$(\overline{x_1} y_1) \ldots \ldots (\overline{x_n} y_n) \quad (7)$$

6.7 Metrics for performance

The efficiency of the frameworks, is examined with the help of following metrics, that are calculated using confusion matrix (as shown in Fig. 25) generated on the dataset:

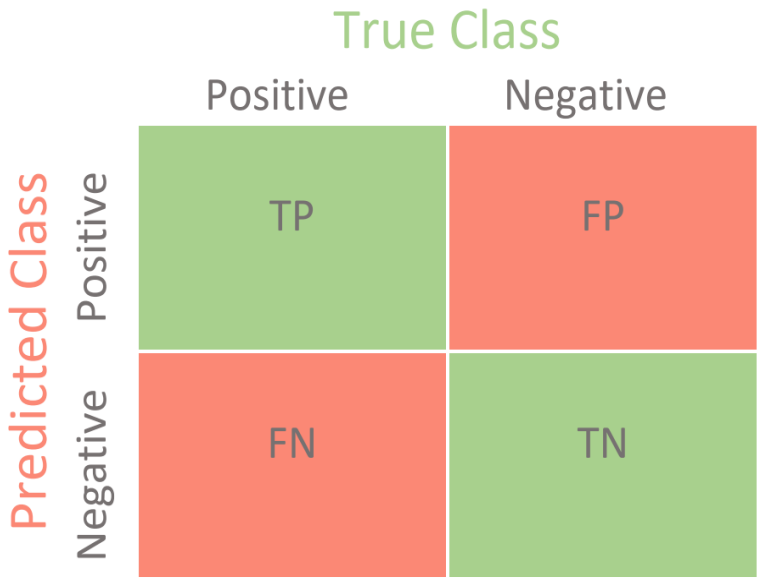


Fig. 19: Confusion Matrix generated on dataset

1. **Accuracy**

It is the ratio of rightfully analyzed results to the total outcomes of the experiment. Following formula is used to calculate the Accuracy:

$$Accuracy = \frac{true\ positive + true\ negative}{true\ positive + false\ Positive + false\ negative + true\ negative} \quad (9)$$

Accuracy is the rightness of analysis made by the framework.

2. **AUC (Area Under ROC Curve)**

AUC represents the model's prediction fit (Fig. 26). It is the extent to which the upper model is able to distinguish between positive and negative events.

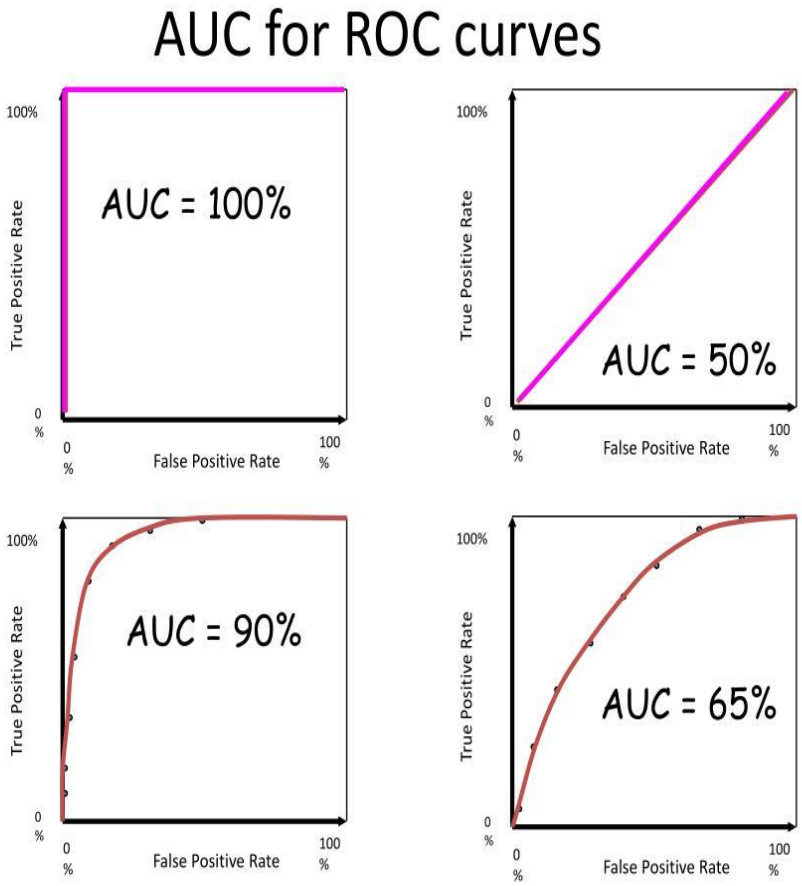


Fig. 20: AUC for ROC Curve

A value of 1.0 in the AUC means the model prediction is 100% accurate and 0.5 means the model prediction is worthless for an unknown event prediction.

3. **Precision**

Precision is the ratio of the correct positive result to the total positive result for a class. The precision is calculated using the following formula:

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

The precision indicates how many positive results are correct for the class.

4. **Recall**

Sensitivity or recall is described as the ratio of correctly classified positive results to the total of actual positive occurrences. Withdrawals are calculated according to the following formula:

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

The callback indicates a misclassified positive prediction.

5. **F1 Score**

F1 scores, also called balanced F measures. The F1 score is simplified as the weighted mean of precision and sensitivity, where the F1 score lies between 0 to1. The corresponding benefaction of precision and sensitivity to the F1 score is the similar. The F1 score is calculated as follows:

$$F1\ Score = 2 * \left(\frac{Precision*Recall}{Precision+Recall}\right) \quad (12)$$

In the case of multi-classes and multiple labels, this is the mean F1 score of each class with weights according to the mean parameters.

The following figure (Figure 27) represents the confusion matrix as well as the metric calculations:

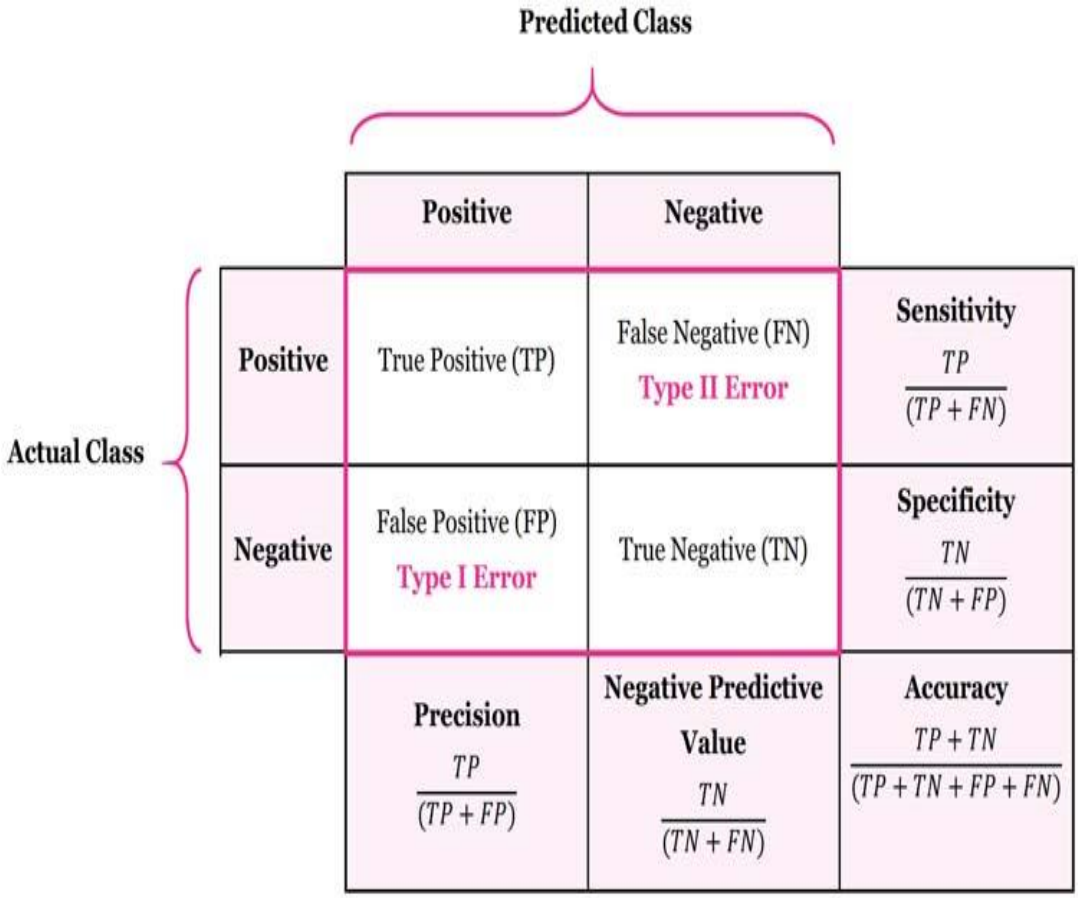


Fig. 21: Confusion Matrix and different classes

CHAPTER 7: RESULT AND ANALYSIS

Performance of different machine learning classification techniques is in the following table.

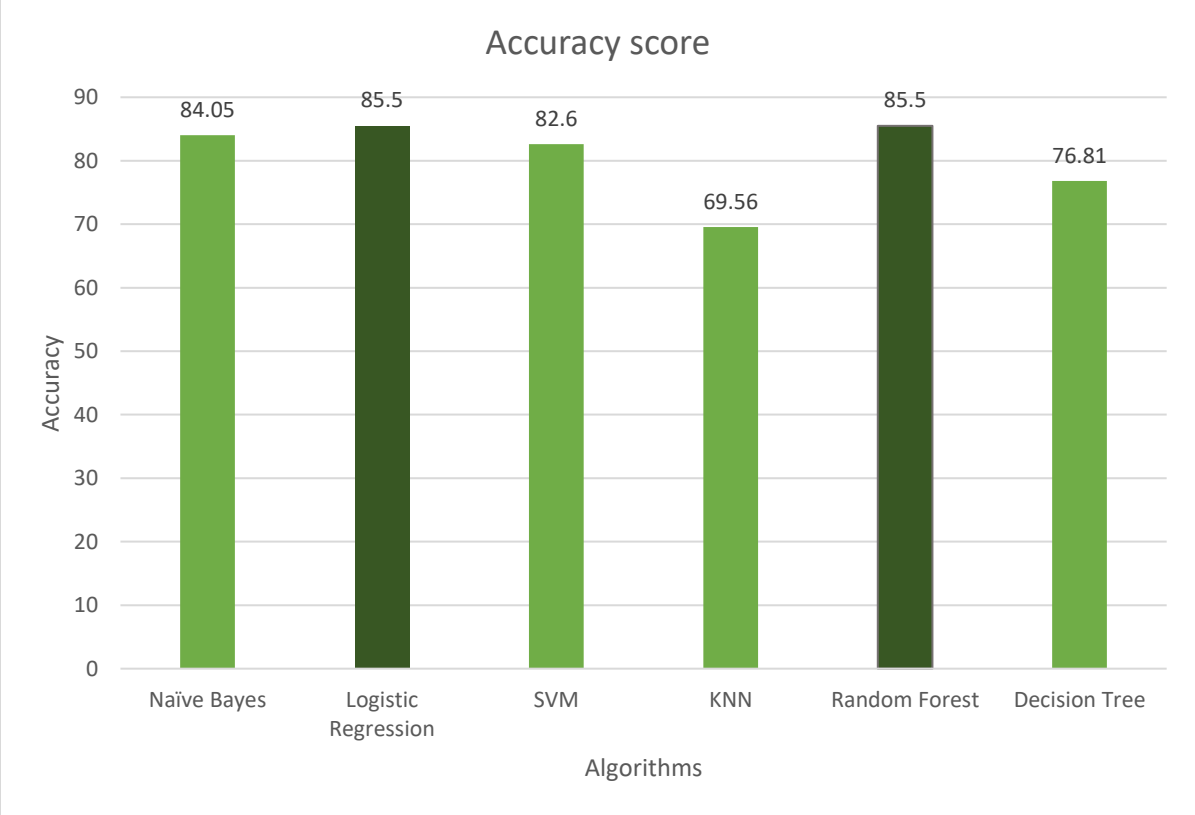
Table II: Performance Metrics for different models

Model	Accuracy	AUC-ROC Score	Precision	Recall	F1 Score
Naïve Bayes	84.05	84.52	93	83	88
Logistic Regression	85.50	84.22	91	88	89
Support vector machine	82.60	80.80	89	85	87
K - nearest neighbor	69.56	67.41	81	73	77
Random Forest	85.50	86.90	95	83	89
Decision Tree	76.81	75.29	86	79	83

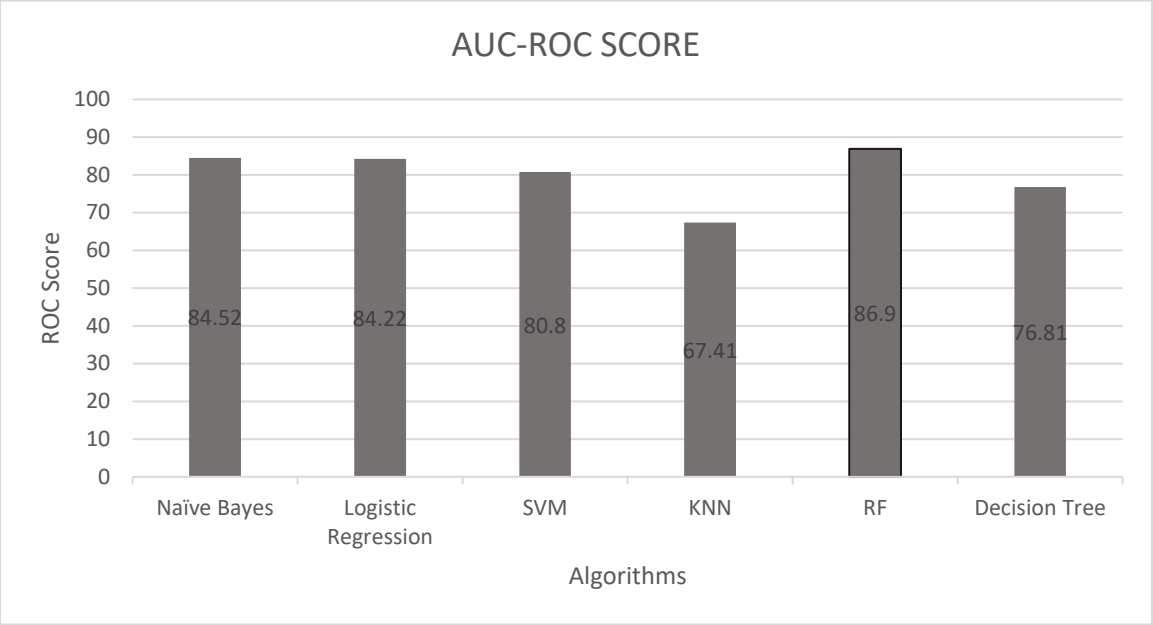
The detection of heart disease at premature stage can save the person's life. Multiple frameworks are generated with the help of ML classification techniques. The dataset is split 70% into training and 30% into testing. Six separate ML classification techniques are used which are K-nearest neighbor, NB, SVM, LR, DT and RF to forecast if the person has heart disease or not. Accuracy, Precision, Recall, F1 Score and AUC (Area under ROC curve) Score are the metrics used to compare the above techniques. ROC-AUC curve will brief about the efficiency of the techniques with the help of distinction of classes. By applying six classification techniques on the Cleveland dataset, we get the outcomes as specified in the above table. The Table specifies various metric

scores for all the techniques. Out of all the classification models under evaluation, Random Forest (RF) and Logistic Regression achieved the highest accuracy (85.50%). Random forest got the highest ROC score (86.92) followed by Naïve Bayes (84.52) and Logistic Regression (84.2). The highest F1 score (89) is achieved by Random Forest and Logistic Regression. Overall, Random Forest give the best performance.

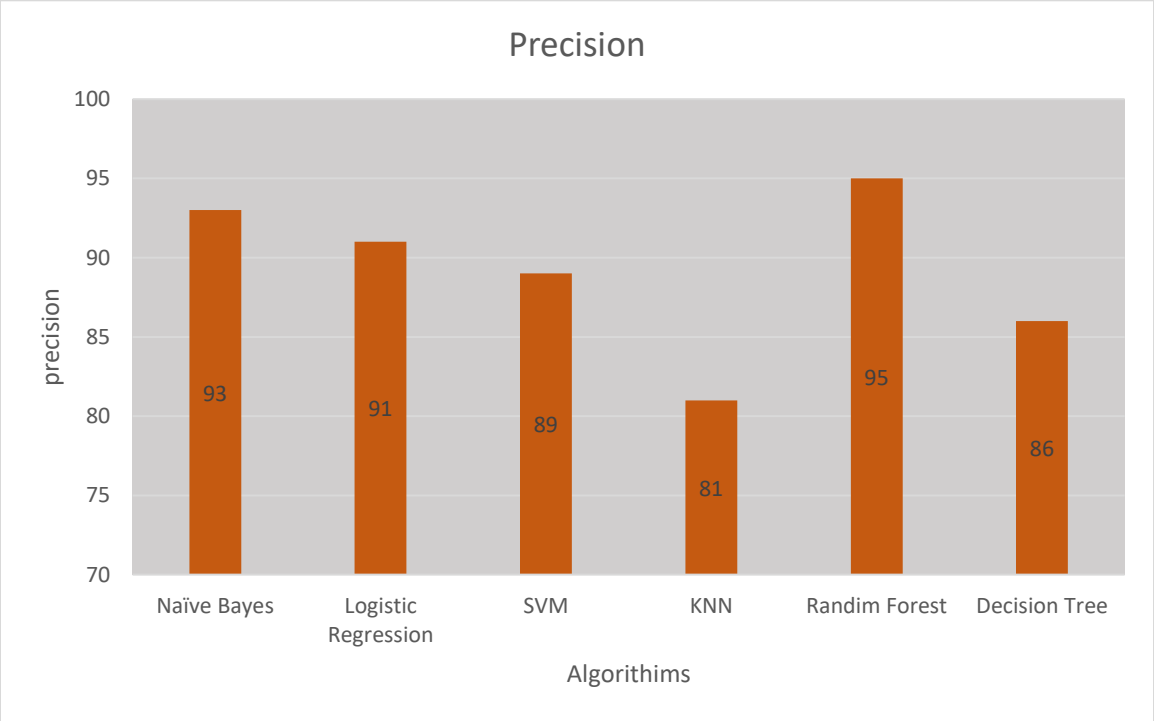
Accuracy plot for the different classification techniques is shown in following figure.



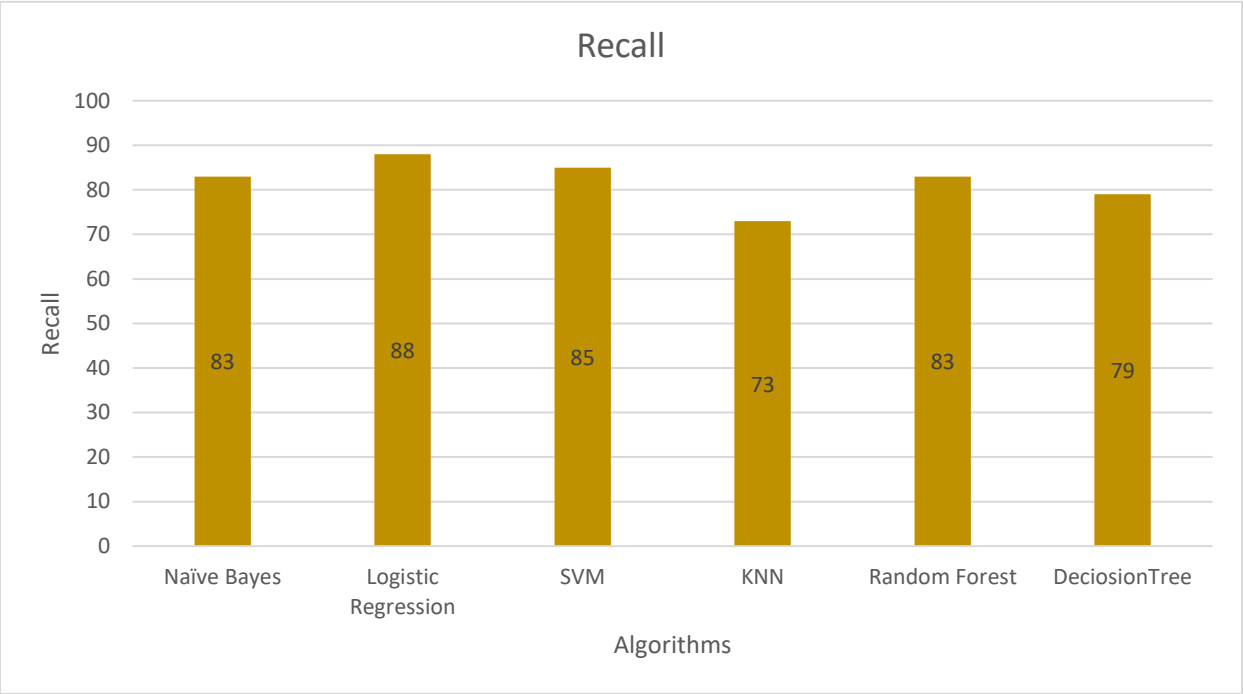
AUC plot for the different classification techniques is shown in following figure.



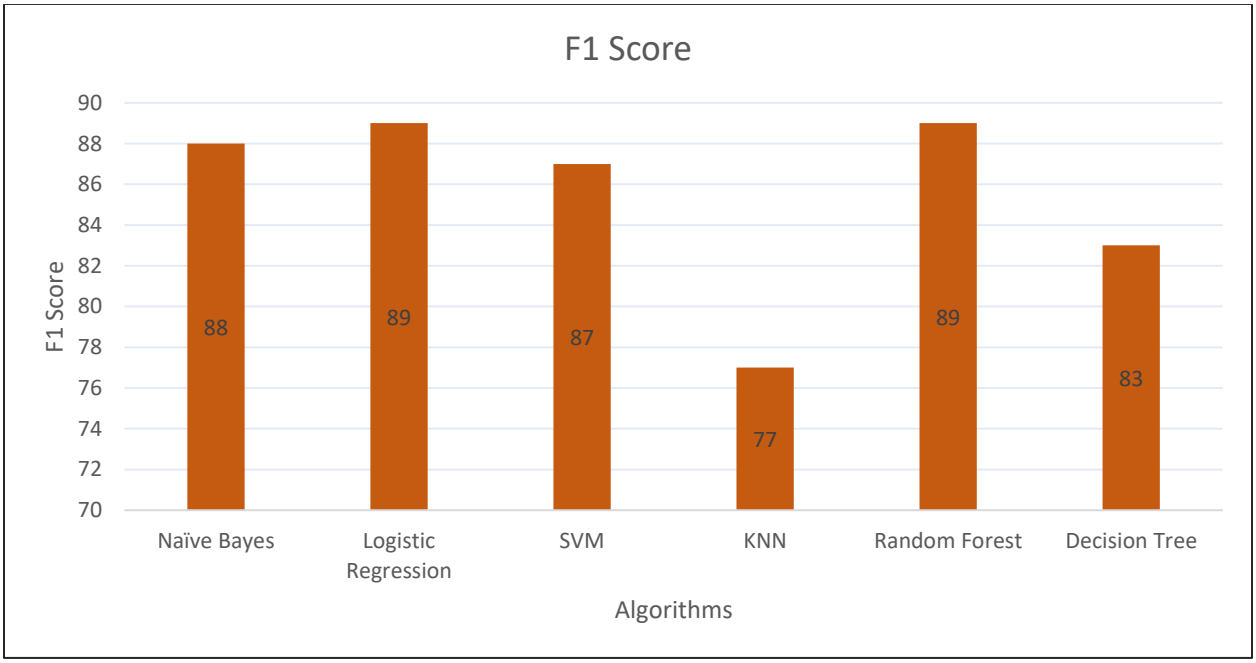
Precision plot for the different classification techniques is shown in following figure.



Recall plot for the different classification techniques is shown in following figure.



Fi score plot for the different classification techniques is shown in following figure.



The following analysis is made on the machine learning techniques.

- Random Forest outperform all the other classification techniques in accuracy, precision, f1 Score, and AUC score. It has maintained the high score in case of recall also.
- Logistic Regression gives the 2nd highest result in case of all the performance metrics.
- In terms of recall, Logistic Regression has the highest value. Accuracy and f1 score value of logistic Regression is same as of random forest, thus it is very much capable in determining if a person is diagnosed with heart disease or not.
- Random Forest is established to be more capable of heart disease projection as compared to logistic regression when all the performance factors are taken into consideration as it is performing highest in 4 out 5 performance criteria and 3rd highest in one of the remaining criteria.
- In term of performance, random forest is followed by logistic regression as it has same accuracy and f1 score value.
- K- nearest neighbor is the lowest performer in all the classification techniques. It has lowest score in all the performance metrics such as accuracy (69.56), AUC (67.41), precision (81), recall (73) and f1 score (77).

CHAPTER 8: CONCLUSION

The principle aim of this project was to explore the various classification techniques and compare them in order to find which method gives better heart disease prediction performance. A categorized research is completed on the basis of six distinct ML classification techniques. Those techniques are DT, KNN, Naïve Bayes, SVM, RF and LR. These classification techniques are differentiated on the basis of five metrics such as accuracy, AUC, precision, recall and f1 score. The conclusion made are as follows:

- It has been found that Random Forest has the most potential to solve the heart disease prediction problem as compared to any of the machine learning model available.
- Random Forest gives the best performance for all the given metrics. It gives highest score for 4 out of 5 metrics.
- It is followed by Logistic Regression (LR) which gives same result as of RF in case of 2 metrics. Overall LR gives best result in 3 out of 5 metrics.

CHAPTER 9: REFERENCES

1) K. Polaraju, D. Durga Prasad, ‘‘Prediction of heart disease using multiple linear regression model”, Int. J. Eng. Dev. Res. Dev., ISSN:2321-9939, 2017.
2) Jaya mi Patel, Prof. Tejal Upadhyay, Dr. Samir Patel.
3) J. Patel Prof. Tejal Upadhyay, Dr. Samir Patel, ‘heart disease prediction using machine learning and data mining technique’, Int. J. Computer Science Communication, 2016, pp. 129-137.
4) Prof. (Dr.) Kanak Saxena, Purushottam, Richa Sharma, ‘‘Efficient heart disease prediction system”, 2016, pp.962-969.
5) Sai rabi H. Mujawar, P.R. Devale, Prediction of heart disease using modified K-means and by using naïve bayes, Int. J. Innov. Res. Computer Communication Engineering. 3 (2015) 10265–10273.
6) Mr. Chala Beyene, Prof. Pooja Kamat, ‘‘Survey on prediction and analysis the occurrence of heart disease using data mining techniques”, Int. J. Pure Appl. Math., 2018.
7) Sellappan Palaniappan and Rafiah Awang, “Intelligent heart disease prediction system using data mining techniques”, International Journal of Computer Science and Network Security, vol.8, no.8, pp. 343-350,2008.
8) J. Laaksonen and E. Oja, “Classification with learning k-nearest neighbors”, In: Proc. of Neural Networks, 1996., IEEE International Conference, Vol.3, pp.1480–1483, 1996
9) R. Jing and Y. Zhang, “A View of Support Vector Machines Algorithm on Classification Problems”, In: Proc. of 2010 International Conference on Multimedia Communications, pp. 13–16, 2010.
10) G. Biau, “Analysis of a random forests model,” J. Mach. Learn. Res., Vol.13, pp.1063–1095, 2012.
11) G. Louppe, “Understanding random forests: From theory to practice”, arXiv Prepr. arXiv1407.7502, 2014.
12) Megha Shahi, R. Kaur Gurm, “Heart Disease Prediction System using Data Mining Techniques”, Orient J. Computer Science Technology, vol.6 2017, pp.457-466.
13) Subbalakshmi, K. Ramesh and N. Chinna Rao, “Decision support in heart disease prediction system using Naïve Bayes”, ISSN: 0976-5166, vol. 2, no. 2. pp.170-176, 2011.

14) C. S. Dangare and S. S. Apte, “Improved study of heart disease prediction system using data mining classification techniques,” International Journal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012.
15) C. M. Bishop, “Pattern recognition and machine learning”, Inf. Sci. Stat., 2006.
16) G. Louppe, “Understanding random forests: From theory to practice”, arXiv Prepr. arXiv1407.7502, 2014.
17) G. Biau, “Analysis of a random forests model,” J. Mach. Learn. Res., Vol.13, pp.1063–1095, 2012.
18) R. Jing and Y. Zhang, “A View of Support Vector Machines Algorithm on Classification Problems”, In: Proc. of 2010 International Conference on Multimedia Communications, pp. 13–16, 2010.
19) J. Laaksonen and E. Oja, “Classification with learning k-nearest neighbors”, In: Proc. of Neural Networks, 1996., IEEE International Conference, Vol.3, pp.1480–1483, 1996
20) S. Thirumuruganathan, “A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm”, [Online]. https//saravananthirumuruganathan.wordpress.c om/2010/05/17/a-detailed- Introduction, 2010.
21) L. Van Cauwenberge, “Top 10 Machine Learning Algorithms”, Data Sci. Cent., 2015.
22) B. Tarle and S. Jena, ‘‘An artificial neural network-based pattern classification algorithm for diagnosis of heart disease,’’ in Proc. Int. Conf. Comput., Commun., Control Automat. (ICCUBEA), Aug. 2017, pp. 1–4
23) J. Laaksonen and E. Oja, “Classification with learning k-nearest neighbors”, In: Proc. of Neural Networks, 1996., IEEE International Conference, Vol.3, pp.1480–1483, 1996
24) C. Sowmiya and P. Sumitra, ‘‘Analytical study of heart disease diagnosis using classification techniques,’’ in Proc. IEEE Int. Conf. Intell. Techn. Control, Optim. Signal Process. (INCOS), Mar. 2017, pp. 1–5.
25) J. S. Sonawane and D. R. Patil, ‘‘Prediction of heart disease using multilayer perceptron neural network,’’ in Proc. Int. Conf. Inf. Commun. Embedded Syst., Feb. 2014, pp. 1–6.
26) Y. E. Shao, C.-D. Hou, and C.-C. Chiu, ‘‘Hybrid intelligent modeling schemes for heart disease classification,’’ Appl. Soft Comput. J., vol. 14, pp. 47–52, Jan. 2014. doi: 10.1016/j.asoc.2013.09.020.

27) S. M. S. Shah, S. Batool, I. Khan, M. U. Ashraf, S. H. Abbas, and S. A. Hussain, ‘‘Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis,’’ Phys. A, Stat. Mech. Appl., vol. 482, pp. 796–807, 2017. doi: 10.1016/j.physa.2017.04.113.

28) M. S. Amin, Y. K. Chiam, K. D. Varathan, ‘‘Identification of significant features and data mining techniques in predicting heart disease,’’ Telematics Inform., vol. 36, pp. 82–93, Mar. 2019. [Online]. Available: https:// linkinghub.elsevier.com/retrieve/pii/S0736585318308876.

29) Purushottam, Prof. (Dr.) Kanak Saxena, Richa Sharma, ‘‘Efficient heart disease prediction system”, 2016, pp.962-969.

30) Mr. ChalaBeyene, Prof. Pooja Kamat, ‘‘Survey on prediction and analysis the occurrence of heart disease using data mining techniques”, Int. J. Pure Appl. Math., 2018.