Project Report

On

HEART ATTACK ANALYSIS AND PREDICTION

Submitted

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

APPLIED STATISTICS AND DATA ANALYTICS

by

PARVATHI MURALEEDHARAN

(Register No. SM21AS017)

(2021-2023)

Under the Supervision of

ARUNIMA P S

DEPARTMENT OF MATHEMATICS AND STATISTICS

ST. TERESA'S COLLEGE (AUTONOMOUS)

ERNAKULAM, KOCHI - 682011

APRIL 2023

ST. TERESA'S COLLEGE (AUTONOMOUS), ERNAKULAM

CERTIFICATE

This is to certify that the dissertation entitled, **HEART ATTACK ANALYSIS AND PREDICTION** is a bonafide record of the work done by Ms. **PARVATHI MURALEEDHARAN** under my guidance as partial fulfillment of the award of the degree of **Master of Science in Applied Statistics and Data Analysis** at St. Teresa's College (Autonomous), Ernakulam affiliated to Mahatma Gandhi University, Kottayam. No part of this work has been submitted for any other degree elsewhere.

Date: 18|5|2023
Place: Ernakulam

ST. TERESA'S COLLEGE
DEPT. OF MATHEMATICS
ERNAKULAM

ARUNIMA P S
Assistant Professor,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

Betty Joseph

SHREEMATI BETTY JOSEPH
Associate Professor and HoD,
Department of Mathematics and Statistics,
St. Teresa's College(Autonomous),
Ernakulam.

External Examiners

1: Dr, Joseph Joshin Rebello
18/5/23

2: SARI THOMAS

DECLARATION

I hereby declare that the work presented in this project is based on the original work done by me under the guidance of **ARUNIMA P S**, Assistant Professor, Department of Mathematics and Statistics, St. Teresa's College(Autonomous), Ernakulam and has not been included in any other project submitted previously for the award of any degree.

Ernakulam.

Date: 18|5|2023

PARVATHI MURALEEDHARAN

SM21AS017

ACKNOWLEDGEMENTS

I would like to convey my heartfelt gratitude to all individuals who encouraged me to carry this work and for their tremendous support and assistance in the completion of my project. The completion of the project would not have been possible without their help and insights.

I am very grateful to my project guide (Arunima P S) for the immense help during the period of work.

In addition, very energetic and competitive atmosphere of the Department had much to do with this work. I acknowledge with thanks to faculty, teaching and non-teaching staff of the department and Colleagues.

I also very thankful to HoD for their valuable suggestions, critical examination of work during the progress.

Ernakulam. PARVATHI MURALEEDHARAN

Date: 18|5|2023 **SM21AS017**

ABSTRACT

Heart disease, which is also called a Cardio Vascular ailment, is one of the major health concerns of many people and is also one of the leading causes of millions of people's death, only second to cancer. In 2020 19.1 million people died due to heart disease worldwide. It is estimated by World Heart Federation that by 2030 there will be more than 23 million Cardio Vascular diseases related deaths per year.

Timely detection of the chance of occurrence of heart attack among patients plays an important role in curing the disease and saving lives. That means predicting the chance of a heart attack has great relevance. Therefore the main aim of this study is to analyze the data under study and optimize an accurate prediction of heart disease.

Methods: This project is an attempt to predict heart attacks using a machine learning model. Different models are constructed and one with better accuracy is chosen. This is done by using different ML approaches such as Logistic Regression, Random Forest, and LGBM classifier and these models focus on predicting based on different parameters. The prediction is done by training the data to carry out optimized predictions. In addition to this EDA is also carried out.

Result: Among the three models used Random forest have better accuracy (98.54%) and produces a better area under the curve (0.985337) compared to the other two.

Conclusion: Further developing this project can be of great help clinically for analyzing various factors and causes of heart attack and thus helps interpret the patient scenario. Thus it helps in detecting and curing heart disease.

ST.TERESA'S COLLEGE (AUTONOMOUS) ERNAKULAM

Certificate of Plagiarism Check for Thesis

Author Name	Parvathi Muraleedharan
Course of Study	MSc Applied Statistics & Data Analytics
Name of Guide	Ms. Arunima P S
Department	Mathematics & Statistics
Acceptable Maximum Limit	20%
Submitted By	library@teresas.ac.in
Paper Title	HEART ATTACK ANALYSIS AND PREDICTION
Similarity	0%
Paper ID	726087
Submission Date	2023-04-24 12:30:09

Signature of Student

Signature of Guide

Checked By
College Librarian

* This report has been generated by DrillBit Anti-Plagiarism Software

Contents

Chapter 1

Introduction

A coronary heart attack is similar to acute myocardial infarction (AMI). It is one of the most serious diseases among heart diseases. A heart attack occurs when one of the coronary arteries becomes blocked and the blood supply to parts of your heart is interrupted. If the body fails to restore the blood supply fast enough, this causes heart muscles to be damaged.

A heart attack most commonly results from atherosclerosis or fat build-ups in the arteries which carries blood to the heart muscles. These plaque build-ups narrow the area inside the arteries, thus restricting the flow of blood.

If any of these plaques in a heart artery breaks open, then it leads to the formation of a blood clot and these clots thus formed thus block the blood flow. When it succeeds in completely stopping the flow of blood to different parts of the heart muscles a heart attack occurs. As a result, gradually that particular section of the heart which is supplied by the artery begins to die. As long as the artery remains blocked the damages continue to increase. Which may lead to that muscle's death or complete damage. Resulting in permanent heart damage, once that muscle dies.

The depth of the damage to the heart muscles is decided by the area of the blocked artery and along with the time taken to treat the injury. Therefore to reduce the heart damage and the severity of the condition of the patient, the artery which is blocked should be opened as soon as possible.

1.1 Major causes

Smoking, lack of physical activities, and high consumption of fat and salt are the major risk factors that lead to a heart attack. Increased Blood pressure and cholesterol are the leading causes of heart diseases, while tobacco, obesity, physical inactivity, diabetes, and metabolic syndrome are also important contributors.

1.2 Treatments

Treatments provided for heart attacks depend upon the severity of the patient's condition and the time between detecting and treating. It can range from lifestyle changes including certain cardiac rehabilitation to medications, surgeries like stenting, and bypass surgery. However early detection greatly helps in reducing the severity thus lightening the treatments.

But as the saying goes prevention is better than cure, a healthy lifestyle avoiding the above risk factors greatly helps in avoiding heart diseases.

1.3 Why this project?

With modern technologies, early detection of a patient's risk of heart attack and taking timely prompt treatment can save him and make him completely free from the risk of a heart attack.

This leads to the need for a method to predict heart attacks in a short time. So, this project is an attempt to meet this need.

This project applies some machine learning techniques to the data under study which is historical data. That is some machine learning algorithms are used to construct certain models that analyze the data and predict heart attacks. This is done by training the models with some input dataset and use of certain statistical methods to analyze the output. So, we are carrying out a supervised learning approach.

Thus, this attempt can be further developed, and making it more accurate can be of great use clinically as well as to common people.

1.4 OBJECTIVES

- **To conduct univariate, bivariate, and multivariate analyses of the data under consideration.**
- **To compare the accuracy of the models**
- **To predict whether the person has a high chance of heart attack from the given information.**

Chapter 2

Literature Review

1.Student's Placement Prediction Model Using Logistic Regression (INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH and TECHNOLOGY (IJERT))

They used Machine learning technique to design and implement a logistic classifier that predicts the probability of the student getting placed along with the Gradient Descent algorithm.

2.Software project failures prediction using logistic regression modeling(Mohammad Ahmad Ibraigheeth and Syed Abdullah Fadzli Faculty of Informatics and Computing, University Sultan Zainal Abidin 21300 Kuala Terengganu, Malaysia)

Several statistical tests were applied to the LR model to evaluate its performance. Confusion matrices that include information about actual and predicted model outputs are shown. For project failure prediction problems, the confusion matrix is used to evaluate the model performance.

Breast carcinoma in men(Sharon H. Giordano M.D., MPH, Deborah S. Cohen M.S., Aman U. Buzdar M.D., George Perkins M.D., Gabriel N. Hortobagyi M.D.)

Age-adjusted incidence rates were calculated. Characteristics of the patients and presenting tumors were compared between men and women. Univariate and multivariate analyses were performed to determine the effect of each variable on overall survival. Survival rates by disease stage were compared for men and women.

Although it remains a rare disease, the incidence of male breast carcinoma is increasing. Breast carcinoma in men has some epidemiologic and biological differences from breast carcinoma in women. Cancer 2004. © 2004 American Cancer Society.

Multivariate analysis of male and female professional career choice correlates. (Wertheim, E. G., Widom, C. S., Wortzel, L. H. (1978). Multivariate analysis of male and female professional career choice correlates. Journal of Applied Psychology, 63(2), 234–242)

Personality, aptitude, achievement, and social-demographic characteristics of graduate students in 4 professional degree programs were investigated in a comparative multivariate analysis of the correlates of professional career choice. 173 male and 175 female 1st-yr graduate students in 2 traditionally male fields (law and management) and 2 traditionally female fields (education and social work) completed an extensive questionnaire. Results confirmed the central hypothesis of the study: Differences across careers for each variable were greater than differences between the sexes within careers. In contrast to previous findings, no significant sex differences were found in assertiveness, locus of control, or Machiavellianism.

A regression analysis of long-term trends in cancer mortality in Japan (1958–2004) (Dongmei Qiu,Kota Katanoda,Tomomi Marugame, Tomotaka Sobue)

A regression model was used to analyze the long-term trends of mortality related to overall cancer. Since 1996, a decline has been seen in overall cancer for both sexes in Japan. Most of the common sites, including cancers of the stomach, colon, liver, gallbladder, and lung and leukemia in both sexes, cancer of the esophagus in men and rectal and ovarian cancers in women showed a decreasing trend, and cancers of the rectum, pancreas, prostate and urinary bladder and malignant lymphoma in men and cancers of the esophagus and uterus in women are level off during the most recent period. However, an increasing trend was confirmed for cancers of the pancreas, breast, and urinary bladder, and malignant lymphoma in women.

Predicting prokaryotic incubation times from genomic features (Maeva Fincher)

This project focused on predicting microbial incubation times from genomic features. All classifiers trained yielded precision and recall values in the 0.4 - 0.5 range, indicating that no model was vastly superior. Under optimal growth conditions, it is possible that the organisms in classes 2 and 3 would grow faster and would fall in class 0.

Future Sales Prediction (Digbalay Bose, Department of Electrical and Computer Engineering University of Southern California, Los Angeles and Souvik Kundu, Department of Electrical and Computer Engineering University of Southern California, Los Angeles)

In this work future sales prediction models based on decision tree structures were made. The evaluation showed the best-performing model can be achieved through the ensembling of LGBM and random forest giving equal weight to each.

Heart Disease Prediction (Singha Taqdees, Department of Software Engineering Fatima Jinnah Women University, The Mall, Rawalpindi, Pakistan Kanwal Dawood, Department of Software Engineering Fatima Jinnah Women University, The Mall, Rawalpindi, Pakistan Nayab Akhtar, Department of Software Engineering Fatima Jinnah Women University, The Mall, Rawalpindi, Pakistan)

This study mainly focused on the use of data mining techniques in healthcare, especially in the detection of heart disease. Data mining techniques were implemented using the following algorithm, KNN, Neural Networks, Decision Tree, Naive Bayes, and Random Forest. The performance was measured based on Accuracy, TN, FP, FN, and TP rate and in some algorithms. We conducted five experiments with the same data set to predict heart disease. This experiment shows that Naive Bayes gives the highest accuracy which is 88% followed by ANN and KNN with an accuracy of 87%. Our findings indicate that data mining can be used and applied in the healthcare industry to predict and diagnose the disease at its early stages.

Heart disease prediction uusing machine learning algorithms (bach-

elor of technology in computer science engineering submitted by gunturu deepthi, cherukuri shivani, koruprolu nagavinith, kesuboyina hanudeep)

In this project, it is concluded that the accuracy of the xgboost is better compared to other algorithms. All the seven machine learning methods' accuracies are compared based on which one prediction model is generated. Hence, it aims to use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently. Comparing all seven the extreme gradient boosting classifier gives the highest accuracy of 81%.

Heart Disease Prediction using Exploratory Data Analysis (R. Indrakumari, T. Poongodi, Soumya Ranjan Jena)

Heart stroke and vascular disease are the major causes. Chest pain is the key to recognizing heart disease. In this work, heart diseases are predicted by considering major factors with four types of chest pain.

Here the datasets are clustered and based on the clusters the happening of chest pain is predicted. The role of exploratory data using tableau provided a visually appealing and accurate clustering experience.

Chapter 3

Methodology

3.1 Dataset for implementation

We have used a built-in dataset from Kaggle for predicting heart disease. This database contains 14 attributes.

3.2 Exploratory Data Analalysis

Exploratory Data Analysis is the process of performing an initial study on a dataset in order to discover patterns, to spot anomalies, and check assumptions with the help of summary statistics and graphical representations. EDA is all about understanding and analysing the data in hand, before getting dirty with it.

3.3 MACHINE LEARNING

Machine learning is a computer program used to optimize a performance criterion based on example data or past experience.

3.3.1 Supervised Learning

Supervised learning is a method of machine learning in which machines are provided and trained using well "labelled" data, and based on which, machines predict the output or where the unlabelled belongs.

3.3.2 Unsupervised learning

Unsupervised machine learning is a machine learning method in which models are not provided with or supervised using any labeled datasets, they are provided with unlabeled data and are able to act on that data without any supervision. regression Logistic regression sometimes called the logistic model or logit model, analyses the relationship between multiple independent variables and a categorical dependent variable, and estimates the probability of occurrence of an event by fitting data to a logistic curve. Logistic regression comes under a supervised machine-learning algorithm.

The binary logistic regression model is the following:

$$f(z) = \frac{1}{1+e^{-z}}$$

where z=$\beta_0 + \beta_1 x_1 + \ldots \beta_k x_k$

3.4 Random forest method

A random forest is an estimator that fits a number of decision tree classifiers on sub-samples of the dataset and uses the averaging method to improve predictive accuracy and control over-fitting. Random Forest comes under a supervised machine-learning algorithm. The random forest classifier can be utilized for solving regression and classification problems.

3.5 K Nearest Neighbour Model

The k-nearest neighbors, also known as KNN , is a non-parametric, supervised learning classifier, used for classification and regression. It works by using proximity to make classifications and predictions about the grouping of each individual data point.

Chapter 4

DATASET AND EXPLORATORY DATA ANALYSIS

4.1 Preliminaries

Table 4.1 Gives attributes information.

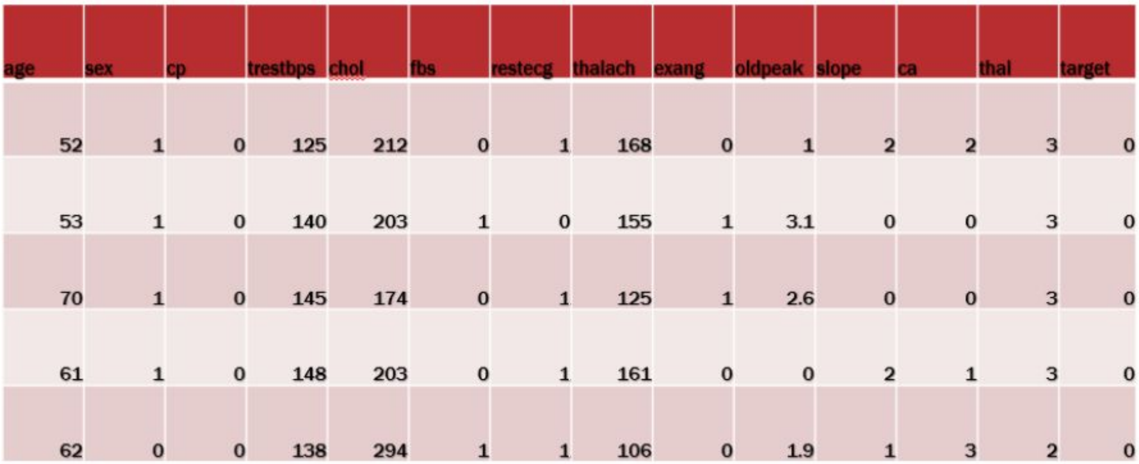
Table 4.1: Attributes

Attribute	Description	Datatype
age	Displays the age of an individual	int644
sex	Displays the gender of the individual using the following format: • 1 = male • 0 = female	int64
cp (Chest-Pain Type)	Displays the type of chest-pain experienced by the individual using the following format: • 0 = typical angina • 1 = atypical angina • 2= non — anginal pain • 3 = asymptotic	int64
trestbps (Resting Blood Pressure)	Displays the resting blood pressure value of an individual in mmHg (unit)	int64
chol (Serum Cholesterol)	Displays the serum cholesterol in mg/dl (unit)	int64

fbs (Fasting Blood Sugar)	compares an individual's fasting blood sugar value with 120mg/dl: • If fasting blood sugar ¿ 120mg/dl then: 1 (true) • else: 0 (false)	int64
restecg (Resting ECG)	displays resting electrocardiographic results: • 0 = normal • 1 = having ST-T wave abnormality • 2 = left ventricular hypertrophy	int64
thalach (Max Heart Rate Achieved)	displays the max heart rate achieved by an individual	int64
exang (Exercise induced angina)	• 1 = yes • 0 = no	int64
oldpeak (ST depression induced by exercise relative to rest)	Displays the value of an integer or float.	float64
slope (Peak exercise ST segment)	• 0 = upsloping • 1 = flat • 2 = down sloping	int64
ca (Number of major vessels (0–3) colored by fluoroscopy)	Displays the value as integer or float.	int64
thal (thalassemia)	Displays the thalassemia (is an inherited blood disorder that causes your body to have less hemoglobin than normal): • 0 = normal • 1 = fixed defect • 2 = reversible defect	int64
target (Diagnosis of heart disease)	Displays whether the individual is suffering from heart disease or not: • 0 = absence • 1 = present	int64

4.2 Data Sample

Figure 4.2 shows sample of first five units.

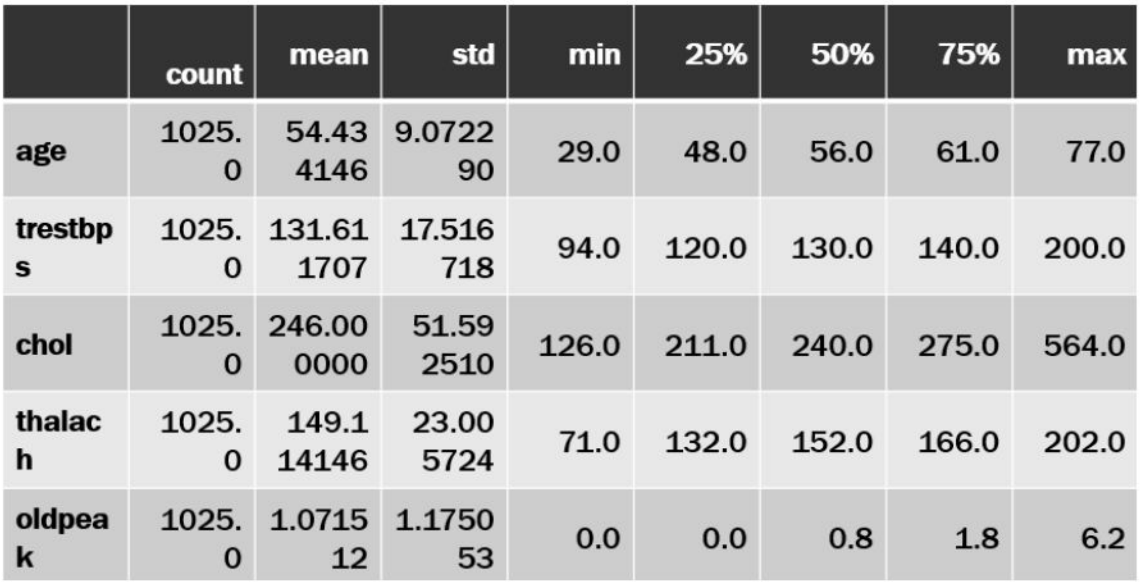


age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Figure 4.1: Data Sample

4.3 Descriptive Statistics

Figure 4.3 shows descriptive statistics.



	count	mean	std	min	25%	50%	75%	max
age	1025.0	54.434146	9.072290	29.0	48.0	56.0	61.0	77.0
trestbps	1025.0	131.611707	17.516718	94.0	120.0	130.0	140.0	200.0
chol	1025.0	246.000000	51.592510	126.0	211.0	240.0	275.0	564.0
thalach	1025.0	149.114146	23.005724	71.0	132.0	152.0	166.0	202.0
oldpeak	1025.0	1.071512	1.175053	0.0	0.0	0.8	1.8	6.2

Figure 4.2: Descriptive Statistics

4.4 Exploratory Data Analysis

4.4.1 Univariate Analysis

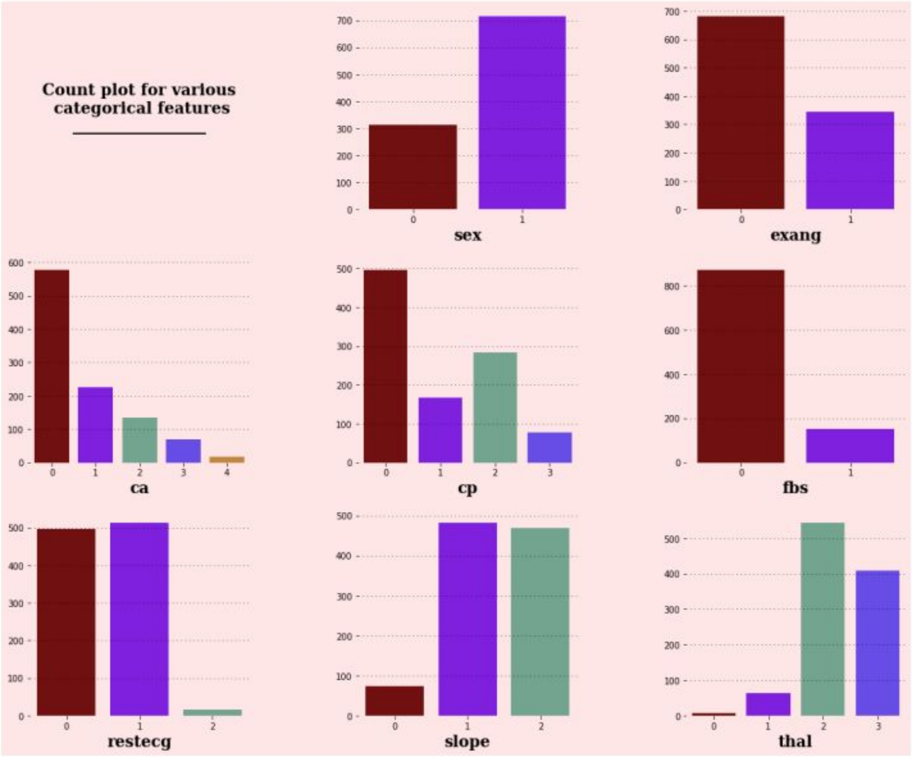


Figure 4.3: Count plot

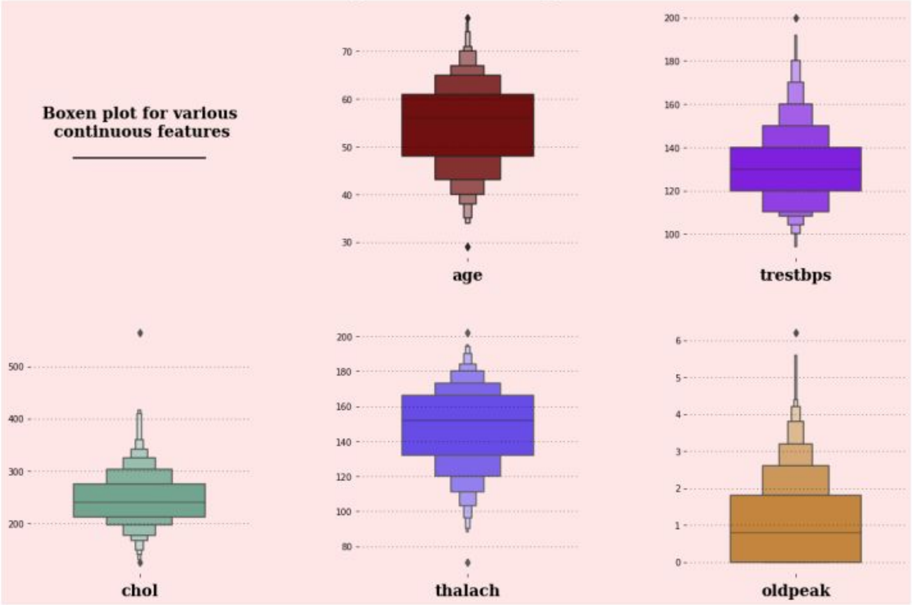


Figure 4.4: Boxen plot

- Age is normally distributed with little variance and after applying log results are same , so we will go ahead with original age values and Age have some outliers.
- In original form 'tretbps' is right skewed , 'chol' is right skewed and 'thalach' is left skewed.

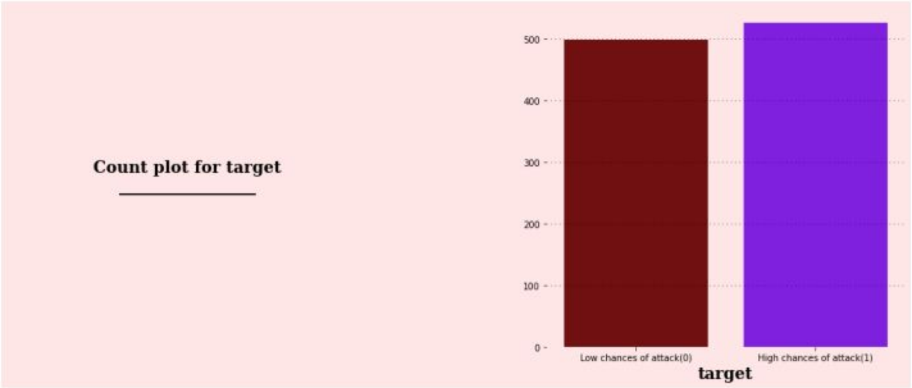
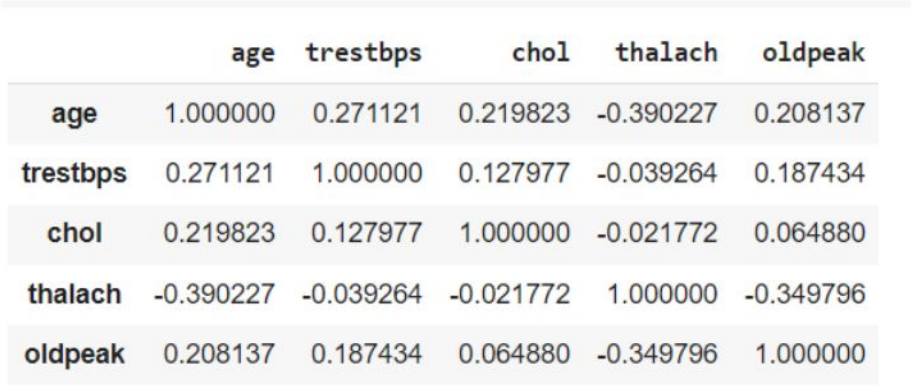


Figure 4.5: Target's count plot

- For the given data number of high chances of heart attack cases is slightly greater than cases with low chances of heart attack. That is the 526 positive cases and 500 negative cases.

4.4.2 Bivariate Analysis



	age	trestbps	chol	thalach	oldpeak
age	1.000000	0.271121	0.219823	-0.390227	0.208137
trestbps	0.271121	1.000000	0.127977	-0.039264	0.187434
chol	0.219823	0.127977	1.000000	-0.021772	0.064880
thalach	-0.390227	-0.039264	-0.021772	1.000000	-0.349796
oldpeak	0.208137	0.187434	0.064880	-0.349796	1.000000

Figure 4.6: Correlation Matrix

Variables are not much associated linearly as the correlation between the variables are weak.

- age and trestbps have a correlation of 0.271121, which is highest positive correlation among variables but it is a weak positive correlation.

- thalach and age have a correlation of -.390227 they are most negatively correlated among variables, but it is a weak negative correlation.

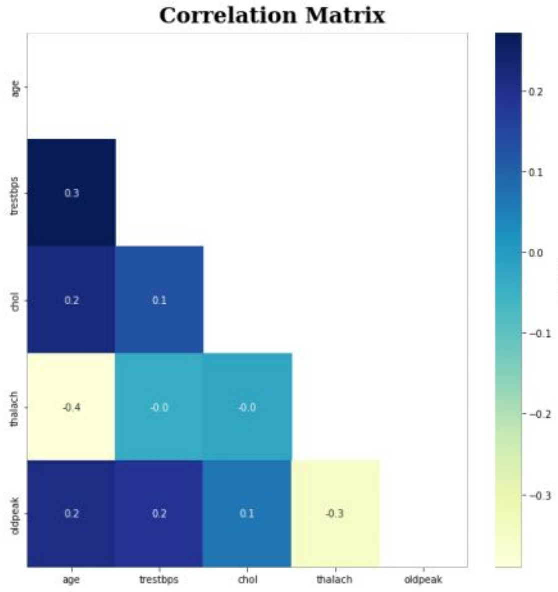


Figure 4.7: Correlation Matrix

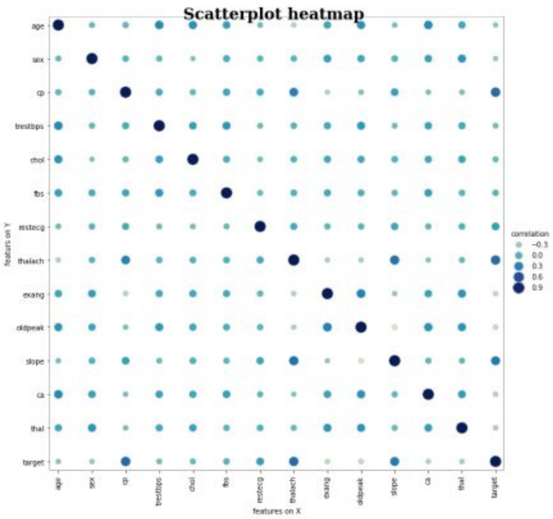


Figure 4.8: Scatter Plot

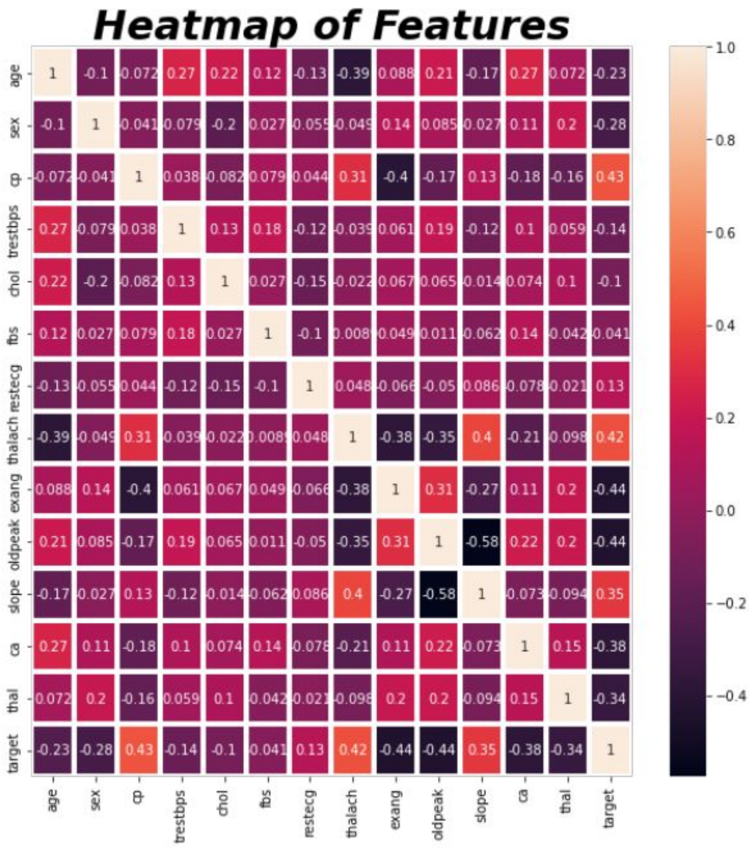


Figure 4.9: Heat Map

According to above heatmap there is no significant positive and negative correlation between two features so we don't need to take any actions. Highest positive and negative correlation are.

Age vs thalach = -0.39 correlated
cp vs target = 0.43 correlated
thalach vs target = 0.42 correlated
exng vs target = -0.44 correlated
slope vs oldpeak = -0.58 correlated

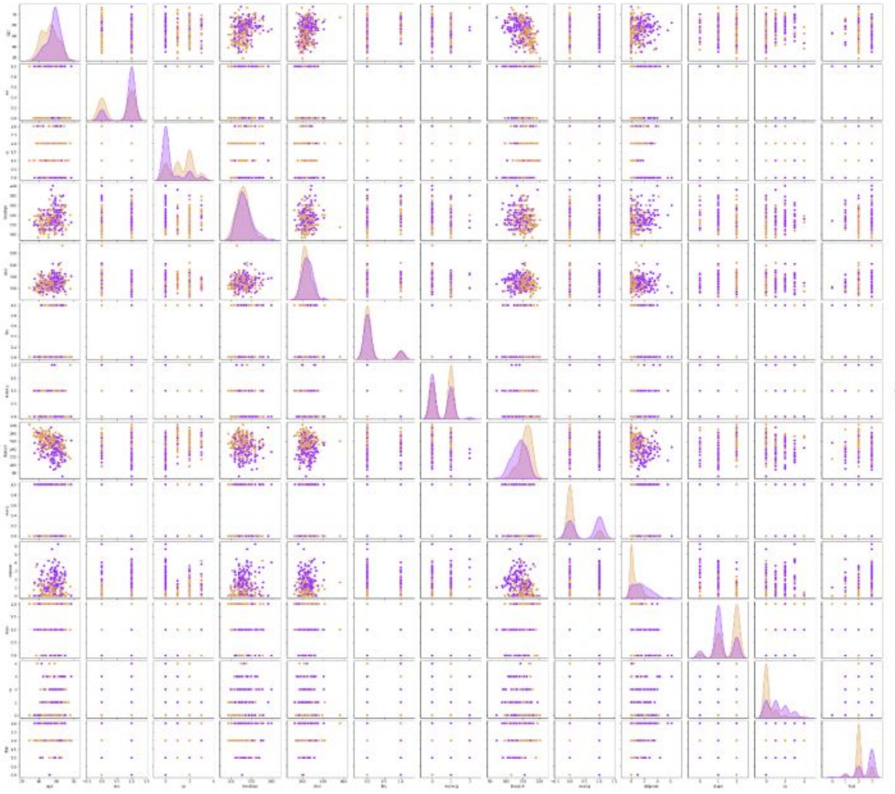


Figure 4.10: Pair Plot

4.5 Data Pre-processing

Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with the dataset's noises, duplicates, and missing values. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Pre-processing of data is required for improving the accuracy of the model.

The data under consideration doesn't have any null values, missing values, or duplicates.

4.5.1 Outlier detection and removal

- **No.of outliers in age = 0**
- **No.of outliers in sex = 0**
- **No.of outliers in cp = 0**
- **No.of outliers in trestbps = 30**
- **No.of outliers in chol = 16**
- **No.of outliers in fbs = 153**
- **No.of outliers in restecg = 0**
- **No.of outliers in thalach = 4**
- **No.of outliers in exang = 0**
- **No.of outliers in oldpeak = 7**
- **No.of outliers in slope = 0**
- **No.of outliers in ca = 87**
- **No.of outliers in thal = 7**
- **No.of outliers in target = 0**

These outliers are then removed.

The data under study is not that imbalanced as we have 500 negative cases and 526 positive cases, but we will use SMOTE oversampling so that our model will have equal opportunity to learn about positive and negative cases. But first, we will split data into training and testing sets before oversampling to avoid any data leakage. 20% of the data is kept for testing and the remaining is used for training the model to predict the output (target).

Then four models are built and trained using the training data which includes the Logistic Regression model, Random Forest classifier model and K Nearest Neighbour model. All these models are then used for predicting and the one with best accuracy is noted, which is then used for predicting.

Chapter 5

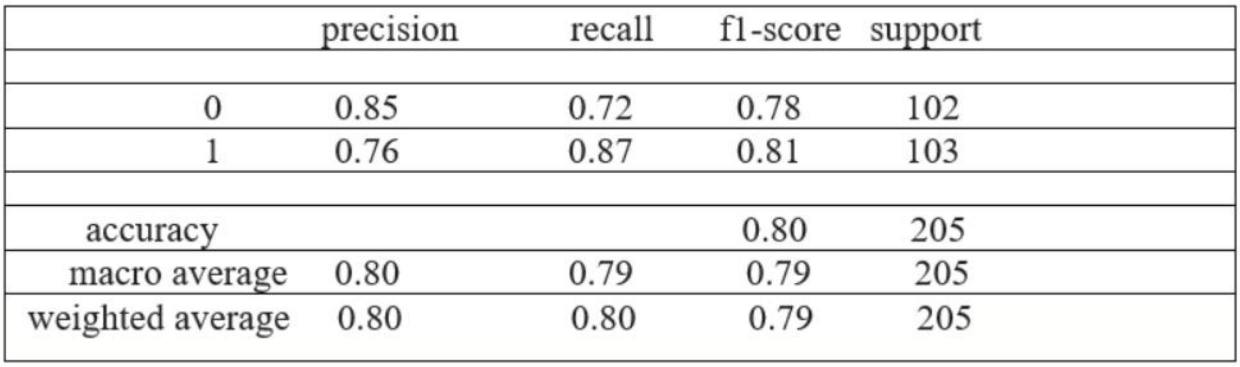
Results

5.1 Logistic Regression Model

Here the aim is to fit a logistic regression model for the given data.

This model gives an accuracy of 0.7951219512195122.

Classification report of Logistic Regression is-



	precision	recall	f1-score	support
0	0.85	0.72	0.78	102
1	0.76	0.87	0.81	103
accuracy			0.80	205
macro average	0.80	0.79	0.79	205
weighted average	0.80	0.80	0.79	205

Figure 5.1: Classification Report

Precision — What percent of your predictions were correct?

Precision: - Accuracy of positive predictions.

This means that the logistic regression model is built have an accuracy of 85% in predicting 0 values (that is the person has no chance of heart attack) and have an accuracy of 76% in predicting 1 values (that is the person has high chance of heart attack).

Area under the ROC curve is given by the graph-

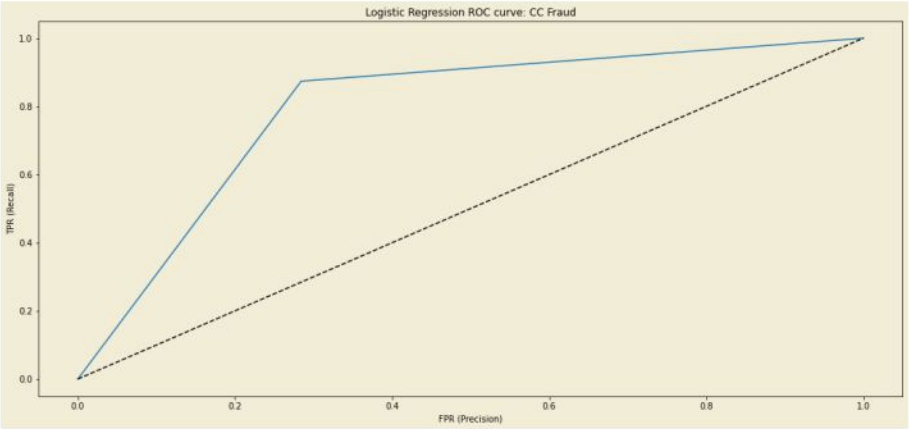


Figure 5.2: ROC Curve
Area= 0.7947363411383972

5.2 K Nearest Neighbour Model

Here aim is to fit a KNN model for the given data.

This model gives an accuracy of 0.824390243902439.

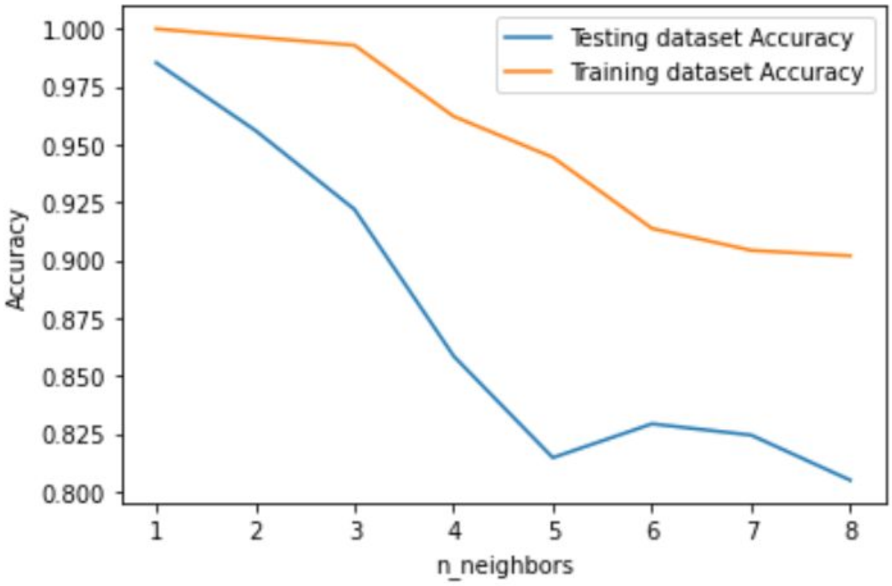
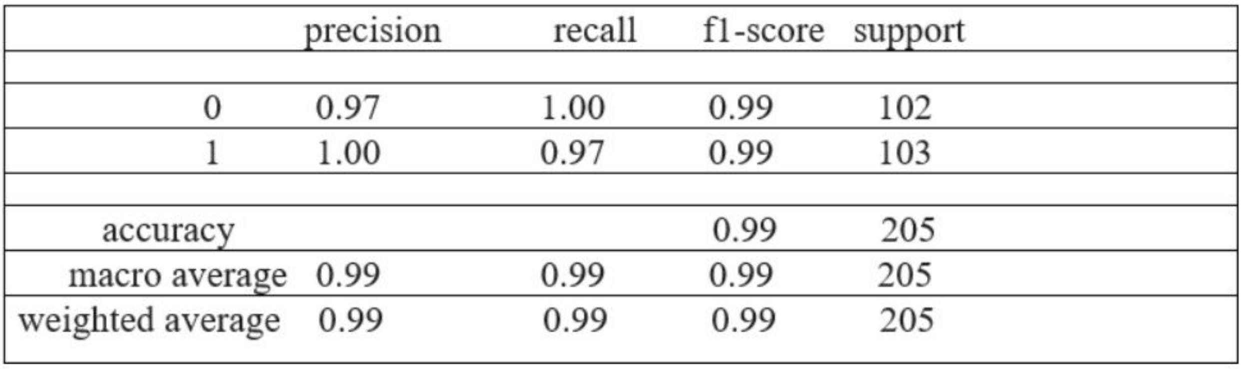


Figure 5.3: Test data-Training data accuracy comparison
The test data set have less accuracy compared to the training data set.

5.3 Random Forest classifier model

Here we aim to fit a random forest model for the given data.
This model gives an accuracy of 0.9853658536585366.
The classification report of Random Forest model is given by-



	precision	recall	f1-score	support
0	0.97	1.00	0.99	102
1	1.00	0.97	0.99	103
accuracy			0.99	205
macro average	0.99	0.99	0.99	205
weighted average	0.99	0.99	0.99	205

Figure 5.4: Classification Report

Precision — What percent of your predictions were correct?
Precision: - Accuracy of positive predictions.

This means that the logistic regression model is built have an accuracy of 97% in predicting 0 values (that is the person has no chance of heart attack) and have an accuracy of 100% in predicting 1 values (that is the person has high chance of heart attack).
Area under the ROC curve is given by the graph-

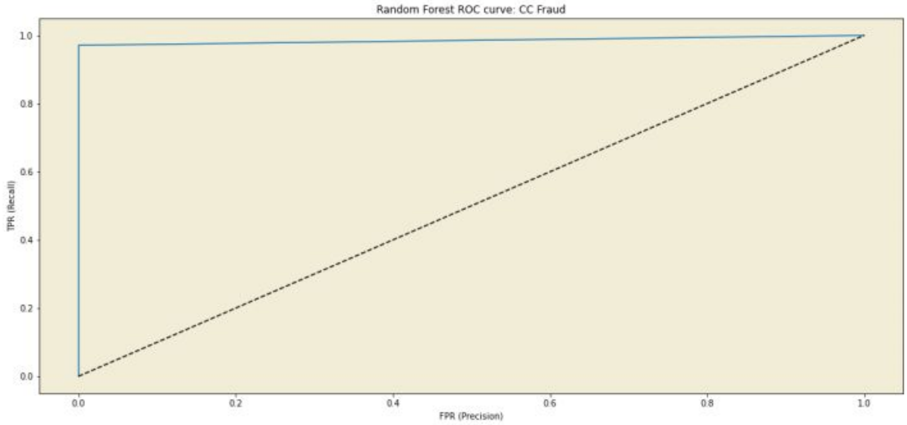


Figure 5.5: ROC Curve
Area= 0.9854368932038835

So among the three models that we built Random Forest classifier model is better, that is the one with better accuracy. So we continue our prediction using Random Forest classifier model.

5.3.1 Feature Importance of Random Forest Classifier model

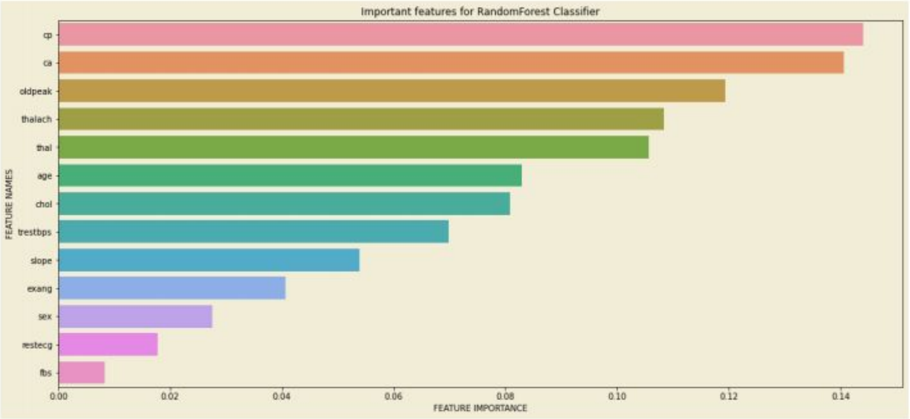


Figure 5.6: Feature Importance

- **Feature importance graph gives the relevance of each feature in deciding the outcome in that corresponding model.**
- **From the above graph it is clear that cp (Chest-pain type) is the most important feature in the Random Forest Classifier model.**
- **fbs (Fasting blood sugar) is the least important feature in this model.**

5.3.2 Random Forest Confusion Matrix

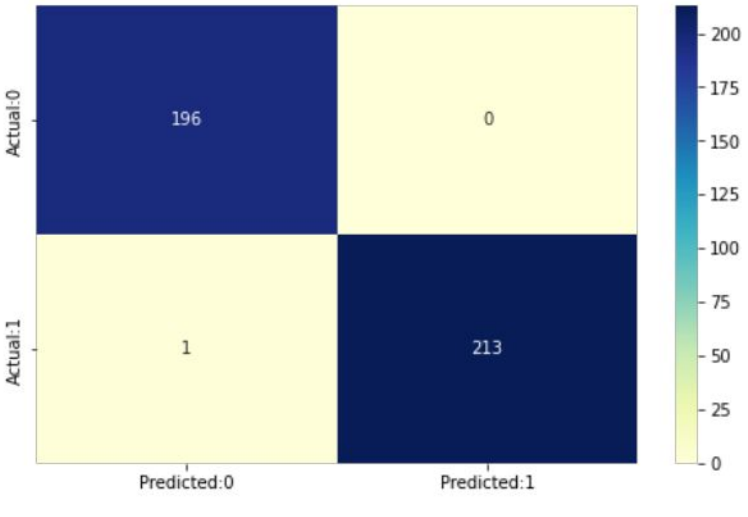
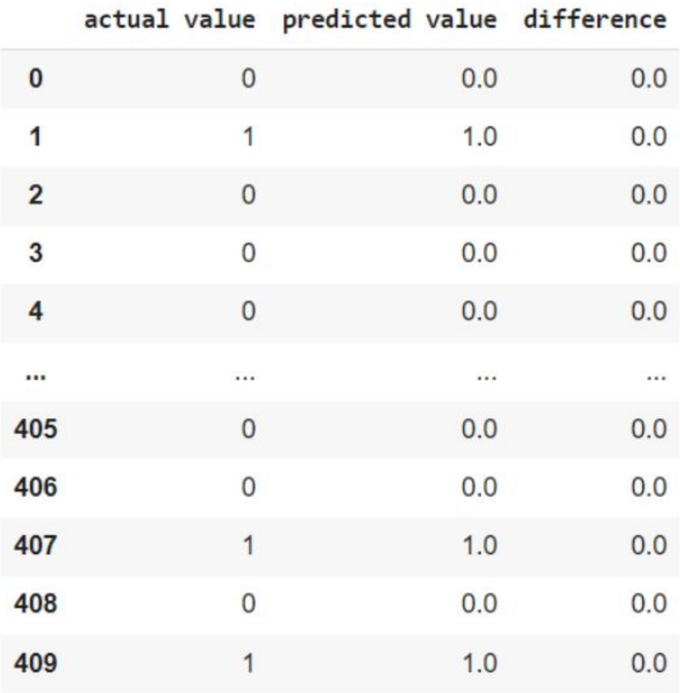


Figure 5.7: Confussion Matrix

The accuracy of the model = TP+TN/(TP+TN+FP+FN) = 0.99756097560976
The Miss classification = 1-Accuracy = 0.0024390243902439046
Sensitivity or True Positive Rate = TP/(TP+FN) = 0.9953271028037384
Specificity or True Negative Rate = TN/(TN+FP) = 1.0
Positive Predictive value = TP/(TP+FP) = 1.0
Negative Predictive Value = TN/(TN+FN) = 0.9949238578680203
Positive Likelihood Ratio = Sensitivity/(1-Specificity) = infinity
Negative Likelihood Ratio = (1-Sensitivity)/Specificity =0.004672897196261627

5.3.3 Prediction using Random Forest Classifier model



	actual value	predicted value	difference
0	0	0.0	0.0
1	1	1.0	0.0
2	0	0.0	0.0
3	0	0.0	0.0
4	0	0.0	0.0
...	...	...	...
405	0	0.0	0.0
406	0	0.0	0.0
407	1	1.0	0.0
408	0	0.0	0.0
409	1	1.0	0.0

Figure 5.8: Prediction

These is a raw representation of the table giving predicted values for the target variable by fitting the Random Forest Classifier model on the test data. We can see that only few predicted values deviate from the actual values of the target variable, this gives the model an accuracy of 98.5%.

Chapter 6

Conclusion

The early detection of risks of heart attack or heart disease in a person can really be helpful in making decisions on lifestyle changes among high risk patients and thus reduce the complications. This can be a great step in reducing the severity of the patient's condition and thus making the treatment more easier and can also help in reducing treatment costs by providing Initial diagnostics in time.

Univariate and bivariate analysis were conducted on the data under study using exploratory data analysis. Three machine learning models were built which includes logistic regression model, K-nearest neighbour model, random forest classifier model and there accuracies were compared. Among these random forest model have the best performance with an accuracy of 98.5% and it is used for the prediction. Important features for the random forest model were also determined.

There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system and thus in future we can use this system for the analysis of different data sets.

Chapter 7

References

1. C. Beyene, P. Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”,

2. Beyene, C., Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics,
118(Special Issue 8), 165–173.

3. Heart Disease Prediction
Singha Taqdees, Department of Software Engineering Fatima Jinnah Women University, The Mall, Rawalpindi, Pakistan Kanwal Dawood, Department of Software Engineering Fatima Jinnah Women University, The Mall, Rawalpindi, Pakistan Nayab Akhtar, Department of Software Engineering Fatima Jinnah Women University, The Mall, Rawalpindi, Pakistan.

4. David, H. B. F., Belcy, S. A. (2018). Heart Disease Prediction Using Data Mining Techniques, 6956(October), 1817–1823.

5. M. A. K. S. H. K. M. a. V. P. M Marimuthu, ”A Review on Heart Disease Prediction usingMachine Learning and Data Analytics Approach”