Heart attack analysis and predection

A Project

by

Team Runtime Terror

Bedre M Samarth

2nd Year

Computer Science Engineering

Vidyavardhaka College of Engineering

Mohammed Kaif Khan Ghori

2nd Year

Computer Science Engineering

Vidyavardhaka College of Engineering

Mohammed Rayid

2nd Year

Mechanical Engineering

Vidyavardhaka College of Engineering

Aaqib Mahamood

2nd Year

Mechanical Engineering

Vidyavardhaka College of Engineering

Hamdan Khan S

2nd Year

Mechanical Engineering

Vidyavardhaka College of Engineering

Anirudh G

2nd Year

Electronics and Communication Engineering

Vidyavardhaka College of Engineering

Anirudha C A

2nd Year

Electronics and Communication Engineering

Vidyavardhaka College of Engineering

Adithya N Bharadwaj

2nd Year

Electronics and Communication Engineering

Vidyavardhaka College of Engineering

Submitted to the CSRBOX team

as the final project deliverable for

IBM SkillsBuild Data Analytics Innovation Camp

**Project Summary**

Cardiovascular diseases (CVDs) are the leading cause of death globally. World Health Organization has estimated that 17.9 million deaths occur worldwide, every year due to heart diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project intends to identify the factors that play a dominant role in causing CVDs and try to predict who is more likely to have a CVD in future, using different machine learning models. The main purpose of the project is to predict who is susceptible to have a CVD in the future and make adequate lifestyle changes to reduce the likelihood of getting a CVD.

The data was collected from the open-source platform Kaggle which contained 303 rows and

14 columns out of which 13 were characteristic variables and 1 was the target variable.

The data was already preprocessed. Distribution plots were visualized, correlation matrices were plotted, pivot tables were formulated for the given data set. The exploratory data analysis was done and the following result were found.

1. Mostly people aged between 55-59 are present in the data
2. Males typically have less chance of heart attack
3. Females have higher chance of attack
4. People aged between (50-58) are more in number with most chance of a heart attack
5. People who have chest pain after exercise have typically less chances of a heart attack
6. People having Chest pain type typical angina (1) have less chances of a heart attack
7. People having Chest pain type non-anginal (3) have higher chances of a heart attack
8. Maximum heart rate achieved (thallach) with 160-164 and 170-174 have higher chances of a heart attack
9. People having Thall as type 3 have higher chances of a heart attack
10. As resting blood pressure of a person increasres, the risk of heart disease also increases
11. Risk of heart disease increases with age.

The outliers in the data were found and then the data was normalized. The feature building on the data was done.

The data was split in 80-20 fashion wherein 80% of the data was used for training and 20% of the data was used for testing. Four Classification models were used to test the accuracy and predict whether a person has a chance of heart attack or not. The accuracy of the four models were found.

1. Logistic Regression – 92%
2. Navies Baies – 89%
3. Random Forest – 89%
4. Decision Tree – 86%

The model can be used to predict the chances of a person having heart attack.

# TABLE OF CONTENTS

CHAPTER

1. INTRODUCTION 1-6

Topic Brief 1

Purpose of the Study 2

Hypotheses 2

Significance of the Study 3

Method of Procedure 4

Collection of Data 4

Treatment of the Data 4

Data Source 4

Definitions of Terms 5

Limitations 6

Assumptions 6

2. PRESENTATION OF FINDINGS (or DATA) 6

3. SUMMARY OF THE STUDY AND THE FINDINGS, CONCLUSIONS, ESTIMATIONS, PREDICTIONS, FUTURE COURSE OF ACTION 15

REFERENCES 22

APPENDICES 23

Appendix

Chapter 1

INTRODUCTION

**Topic Brief**

Cardiovascular diseases (CVDs) are the leading cause of death globally. World Health Organization has estimated that 17.9 million deaths occur worldwide, every year due to heart diseases. Cardiovascular diseases (CVDs) have now become the leading cause of mortality in India. A quarter of all mortality is attributable to CVD. Ischemic heart disease and stroke are the predominant causes and are responsible for >80% of CVD deaths. The Global Burden of Disease study estimate of age-standardized CVD death rate of 272 per 100 000 population in India is higher than the global average of 235 per 100 000 population.

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project intends to identify the factors that play a dominant role in causing CVDs and try to predict who is more likely to have a CVD in future, using different machine learning models.

The dataset is available on Kaggle. The dataset contains the following attributes:

Age: Age of the patient

Sex: Sex of the patient

exang: exercise induced angina (1 = yes; 0 = no)

ca: number of major vessels (0-3)

cp: Chest Pain type chest pain type

Value 1: typical angina Value

2: atypical angina Value

3: non-anginal pain Value

4: asymptomatic

trtbps : resting blood pressure (in mm Hg)

chol : cholestoral in mg/dl fetched via BMI sensor

fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

rest\_ecg : resting electrocardiographic results

•Value 0: normal

•Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)

• Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria thalach : maximum heart rate achieved

target: 0= less chance of heart attack 1= more chance of heart attack

**Purpose of the Study**

The main purpose of the project is to predict who is susceptible to have a CVD in the future and make adequate lifestyle changes to reduce the likelihood of getting a CVD. As an early prognosis may help in reducing the risk complications. Which in turn may decrease the fatality rate in patients having CVD. And also, to pinpoint the prominent factors which lead to CVDs.

**Hypotheses**

Cardiovascular Diseases are chronic diseases that occur by long-term cumulative effects of risk factors. Besides, a large number of people die from acute cardiovascular events without prior symptoms. And about two-thirds of deaths caused by CVD occur in out-of-hospital conditions. It is therefore important to develop effective risk prediction approaches for screening individuals who are at high risk of developing CVD for timely prevention and treatment at an early stage before obvious symptoms happen.

Several prediction models have been proposed to estimate a 10-year risk of developing CVD. The models are expressed as multivariate regression equations using risk factors as variables. The most influential model is the Framingham Risk Score (FRS), which predicts coronary heart disease (CHD) using traditional risk factors as follows: age, diabetes, smoking, systolic blood pressure (SBP), treatment for hypertension, total cholesterol, and high-density lipoprotein (HDL) cholesterol. Other similar risk-scoring algorithms, such as ATP-III, SCORE, PROCAM, QRICK, Reynolds Risk Score, and MUCA, are accomplished through the incorporation of factors into the FRS or recalibration of the Framingham functions to the local subjects. These prediction models have become primary tools in the prevention of CVD in clinical practice. Based on these models, individuals classified in low-risk stratum, intermediate stratum, and high-risk stratum are recommended for lifestyle modification, further risk stratification or drug therapy, and more intensive preventive interventions, respectively.

**Significance of the Study**

The objective of this project is to review current evidences regarding the physiological parameters for CVD prediction, as well as to discuss the potential implications for promoting CVD prevention and treatment in the future.

**Method of Procedure**

**Collection of Data**

1. We collected the data from the open-source platform ‘Kaggle’
2. The data in Kaggle was originally taken from the University of California, Irvine’s Machine Learning Repository

**Treatment of Data**

The data has 303 rows and 14 columns present, 13 of which are characteristic variables (input data) and one of them is a target variable (which we want our model to be able to predict).

The original dataset contained 76 attributes but only 14 of them have been chosen experimentation and machine learning purpose. The data has been preprocessed by the University of California, Irvine. There are no empty or Null values in the data. The dataset is also pretty balanced as the number of people with a chance of heart attack is same as the number of people not having a chance of heart attack.

**Data Source**

The data was collected by the University of California, Irvine from 4 different hospitals from around the world namely

1. *Hungarian Institute of Cardiology, C/O: Dr Andras Janosi, MD.*
2. *University Hospital, Zurich, Switzerland, C/O: Dr William Steinbrunn, MD*
3. *University Hospital, Basel, Switzerland, C/O: Dr Matthias Pfisterer, MD*
4. *V.A. Medical Centre, Long Beach and Cleveland Clinic Foundation, C/O: Dr Robert Detrano, MD ,Ph.D.*

The data we’ve collected is mainly from the open-source platform Kaggle

The links to the data source is as follows

1. <https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>
2. https://archive.ics.uci.edu/ml/datasets/heart+disease

**Definitions of Terms**

The data description is as follows

1. Age: Age of the patient
2. Sex: Sex of the patient (1=Male, 0=Female)
3. Exang: Exercise induced Angina (1=Yes, 0=No)
4. Ca: Number of Major Vessels
5. Cp: Chest pain type
   1. 1. Typical Angina
   2. 2. Atypical Angina
   3. 3. non-Anginal
   4. 4. Asymptomatic
6. Trtbps: Resting Blood Pressure
7. Chol: Cholesterol fetched via BMI sensor
8. Fbs: Fasting Blood Sugar
9. Old peak: ST depression induced by exercise relative to rest
10. Rest\_ecg: resting electrocardiographic results
    1. 0. Normal
    2. 1. ST abnormalities
    3. 2. Showing probable or definite Left Ventricular hypertrophy by Estes’ Criterion
11. Slp: Slope of the peak exercise ST segment
    1. 2= Upsloping
    2. 1=Flat
    3. 0=Down sloping
12. Thalach: Maximum heart rate achieved
13. Thall:
    1. 2=Normal
    2. 1=Fixed Defect
    3. 3=Reversable Defect
14. Target:
    1. 0= Less chances of a heart attack
    2. 1= More chances of a heart attack

**Limitations and Delimitations**

There are No Limitations or Delimitations in the dataset .

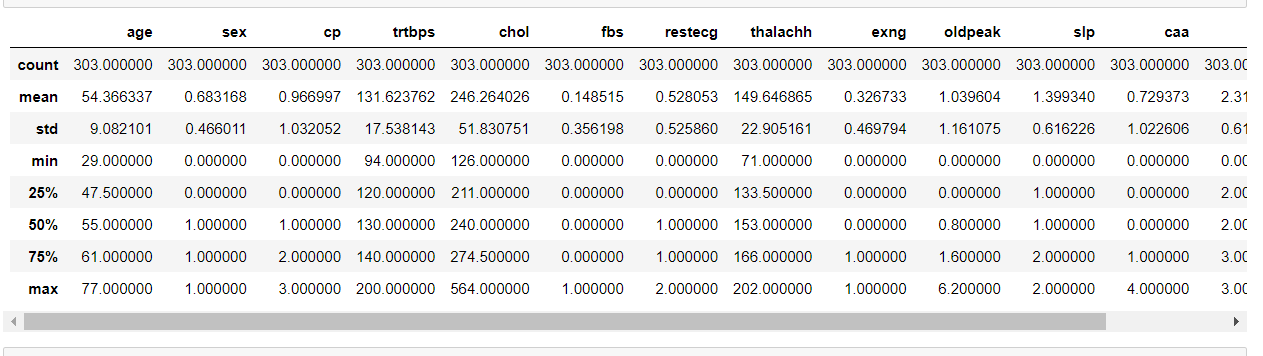
**Assumptions**

There are no Assumptions which we have made.

Chapter 2

PRESENTATION OF FINDINGS

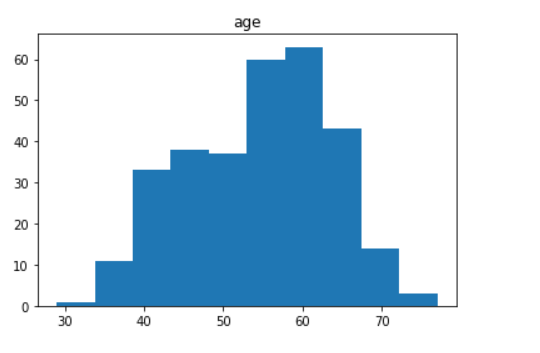
The Analysis started with importing of libraries and reading of data into pandas data frame. The dataset consisted of 303 row and 14 columns. The data did not have any empty or Null values .All the items in the dataset were mainly of Int or Float type and there was no object type of data. The data was then described using pandas describe function ().

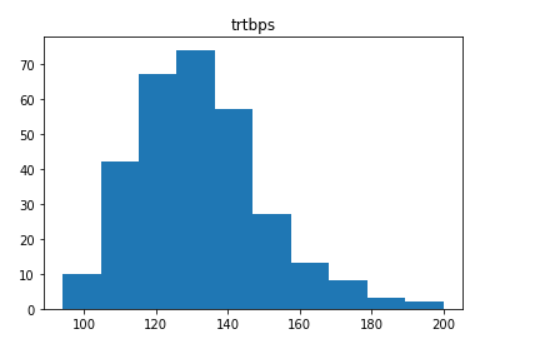
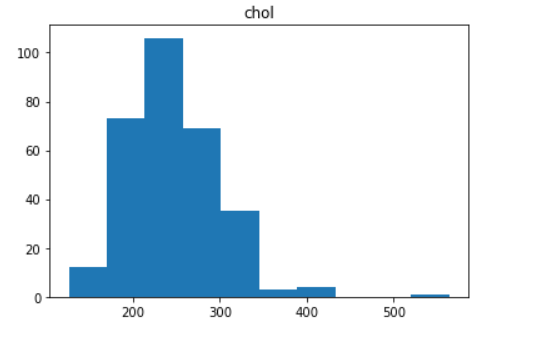
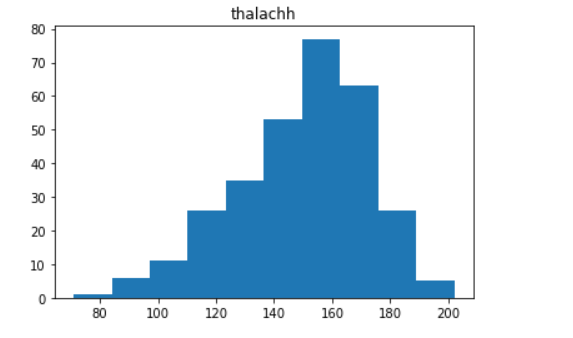


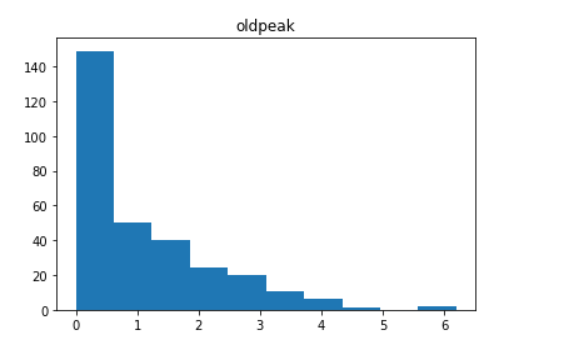
The data set contained two types of columns. One type is the columns containing numerical data and the other type contains Categorical data.

The two types of columns were split into separate data frames.

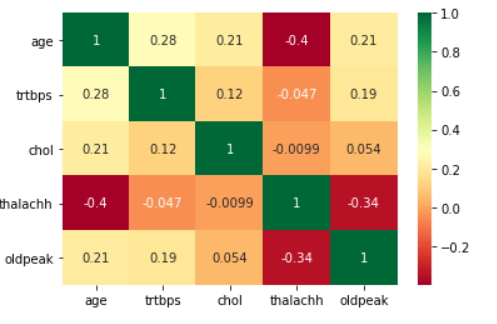
The numerical data was then visualized to see how the data was distributed.

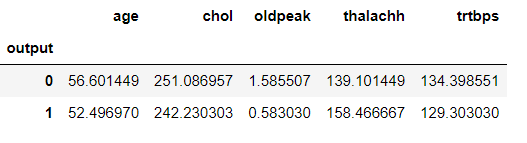


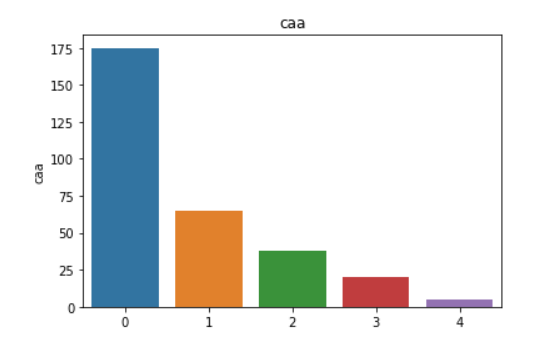
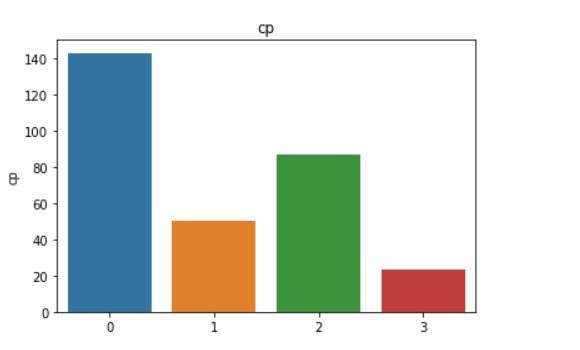


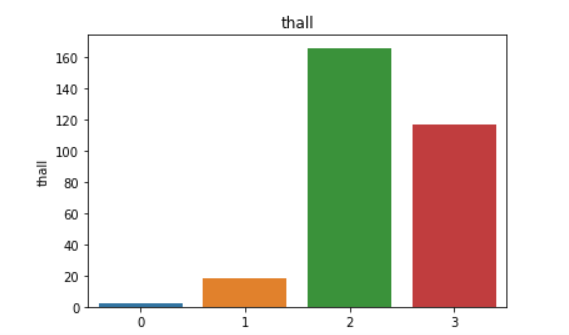
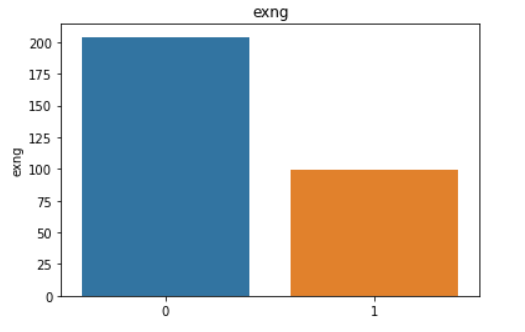
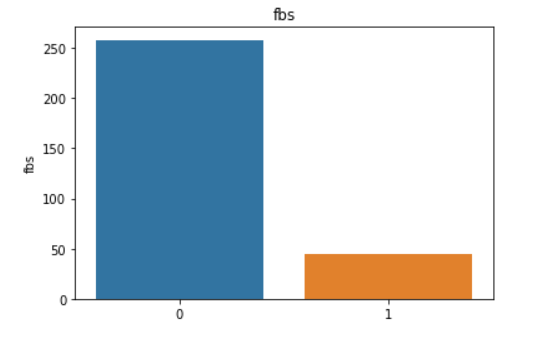


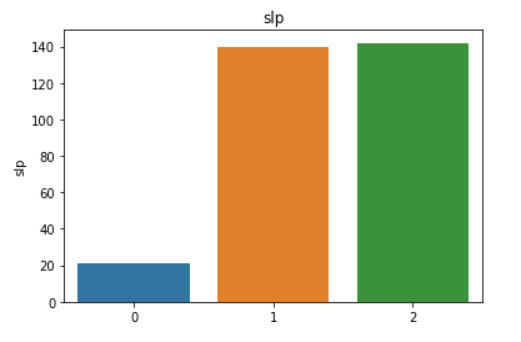
The correlation matrix for the numerical data was found as follows. The pivot table for the data was also found as follows.

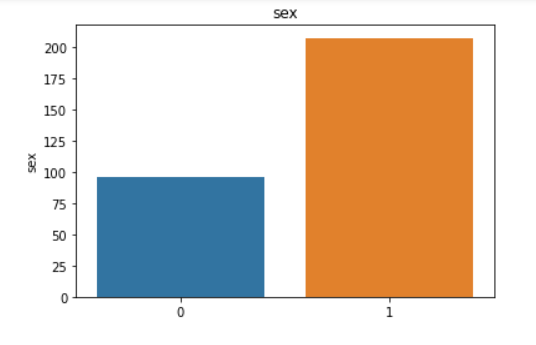


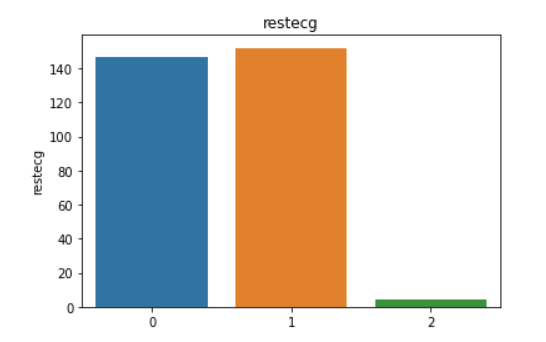
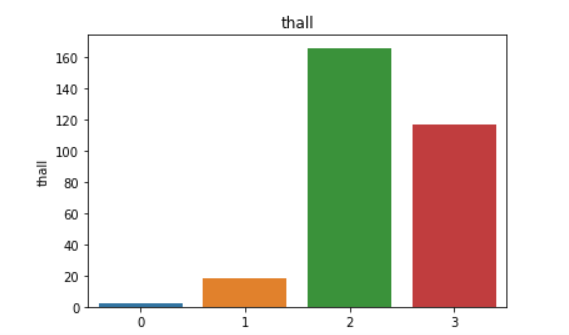


The categorical data was visualized next and the distribution was found

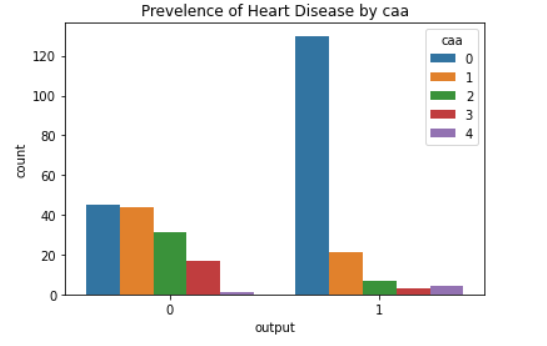
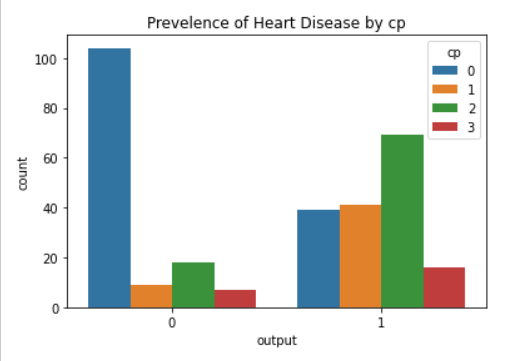


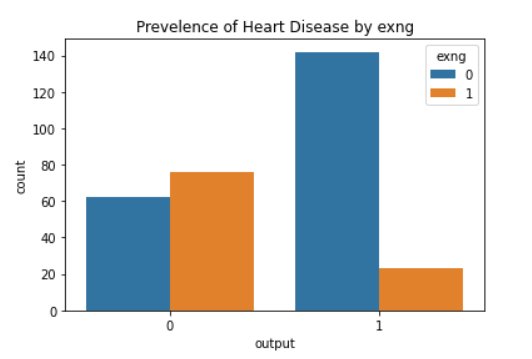
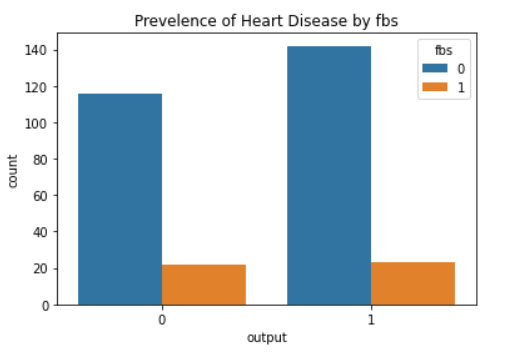
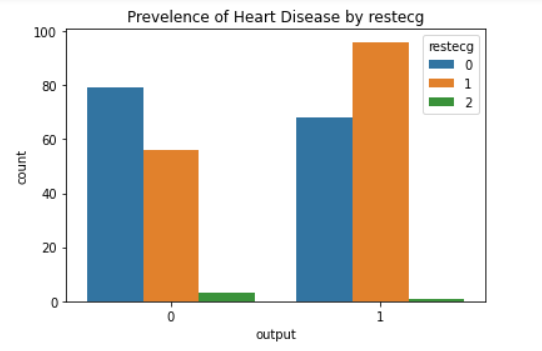
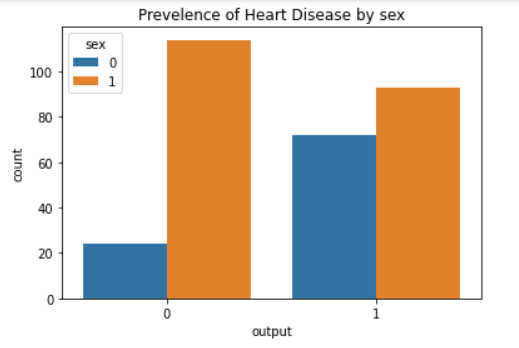


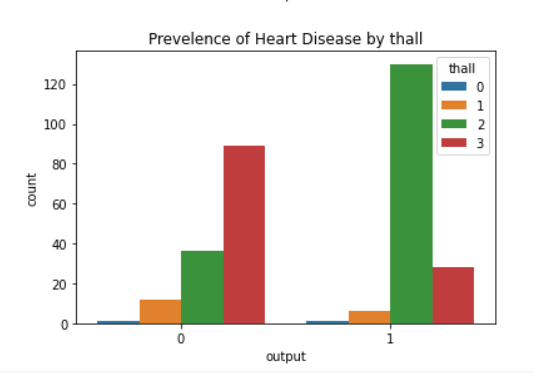
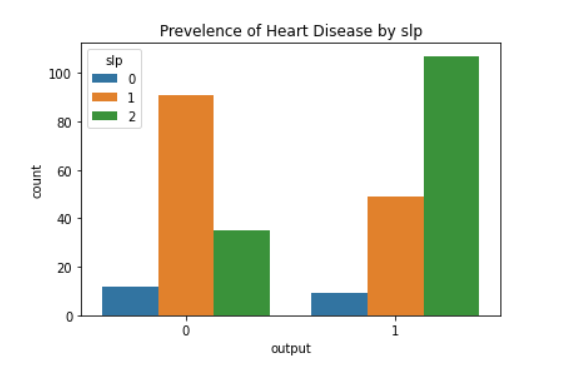


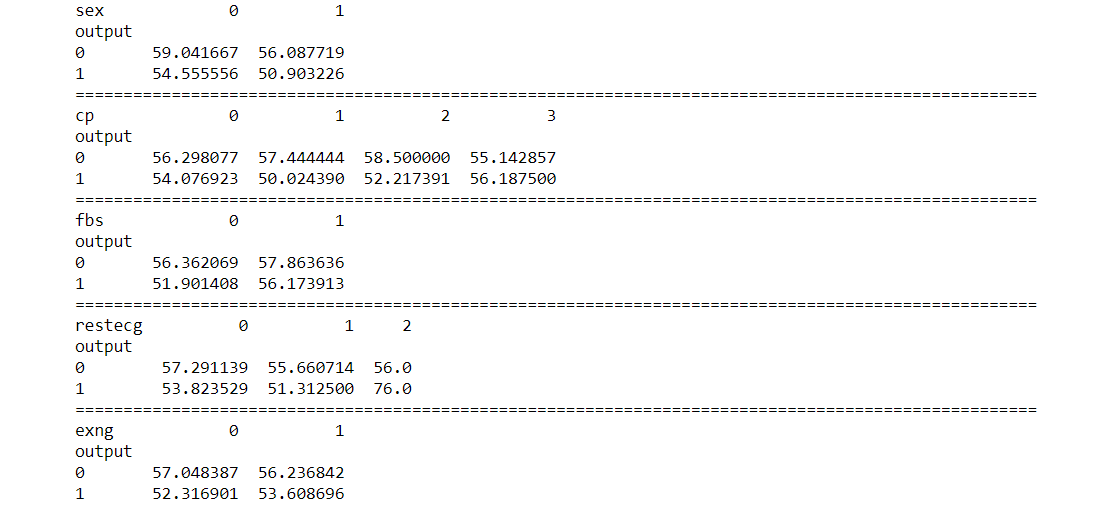


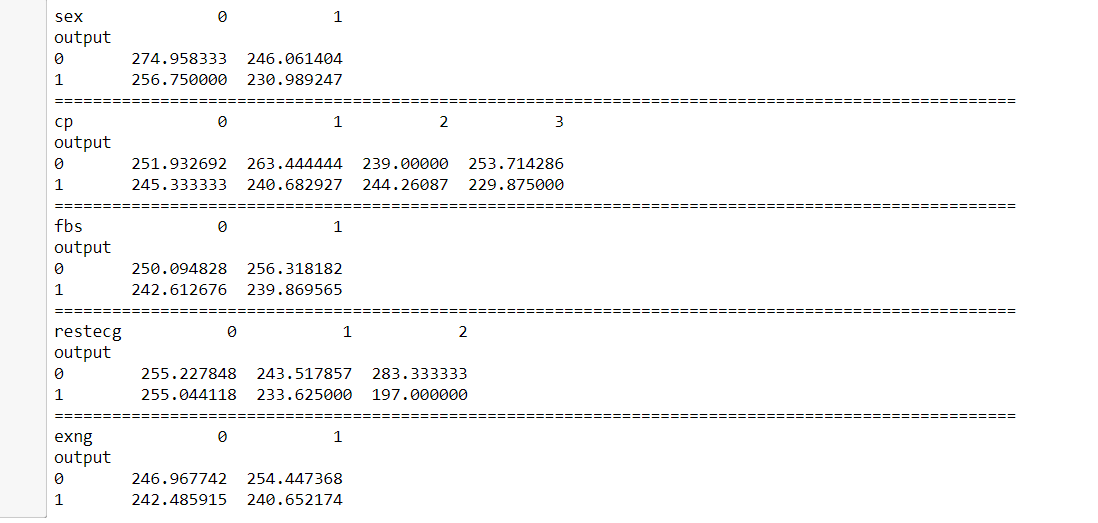
The prevalence of heart attack for each categorical variable was then visualized and the following graphs were obtained

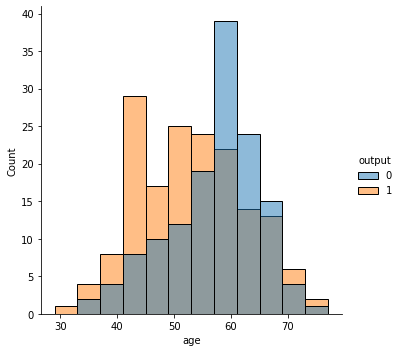
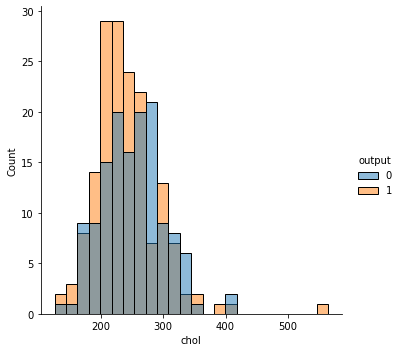


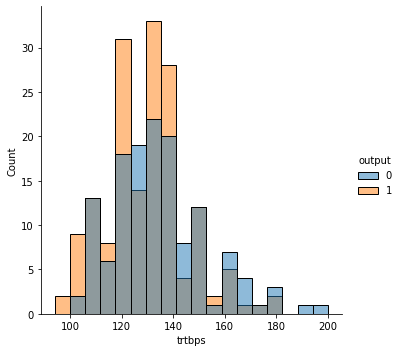


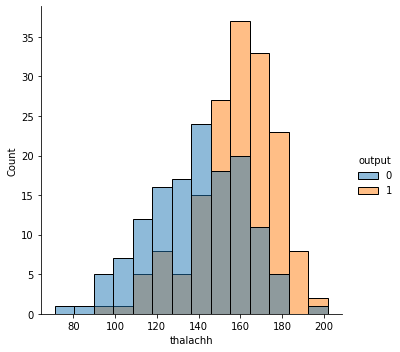


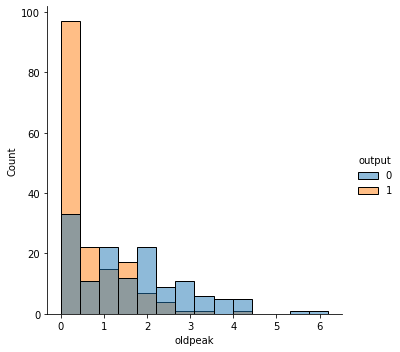
The categorical column’s pivot tables were formulated to inspect the relation among the variables closely 



The relation between the numerical variables and the output variable was visualized as the following plots

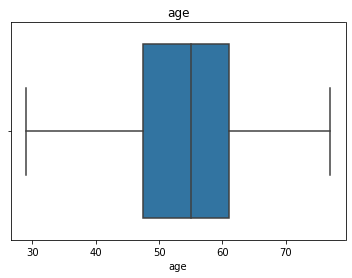
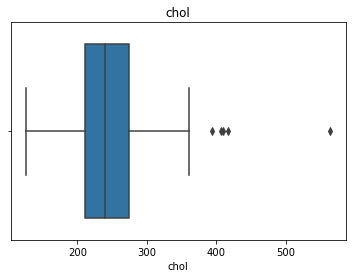


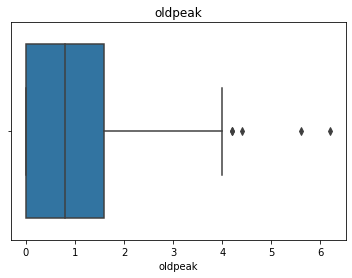
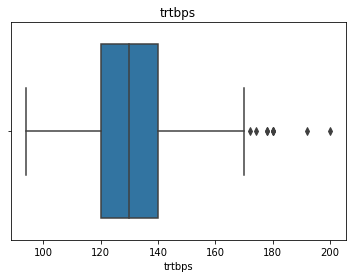


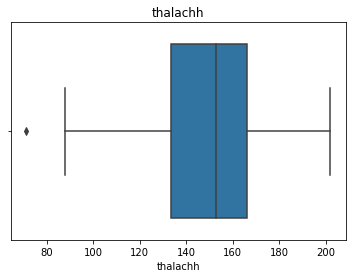


After studying the plots and the pivot table the following conclusion were made

1. Mostly people aged between 55-59 are present in the data
2. Males typically have less chance of heart attack
3. Females have higher chance of attack
4. People aged between (50-58) are more in number with most chance of a heart attack
5. People who have chest pain after exercise have typically less chances of a heart attack
6. People having Chest pain type typical angina (1) have less chances of a heart attack
7. People having Chest pain type non-anginal (3) have higher chances of a heart attack
8. Maximum heart rate achieved (thallach) with 160-164 and 170-174 have higher chances of a heart attack
9. People having Thall as type 3 have higher chances of a heart attack
10. As resting blood pressure of a person increasres, the risk of heart disease also increases
11. Risk of heart disease increases with age.

Before proceeding to the model building and the model testing, the outliers in the data were identified.



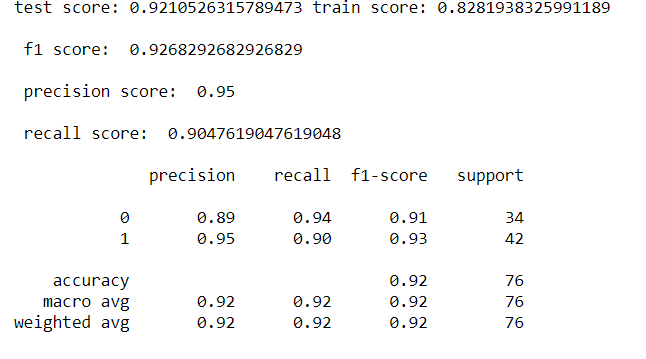


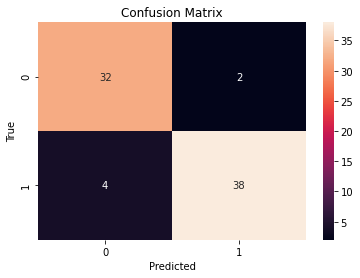
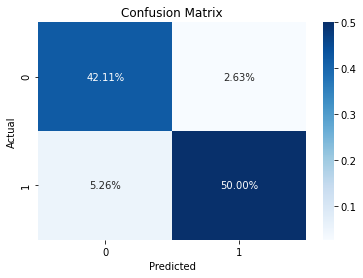
The data was then normalized to get a better accuracy in the model building. Feature engineering was also done on the data to make it sophisticated.

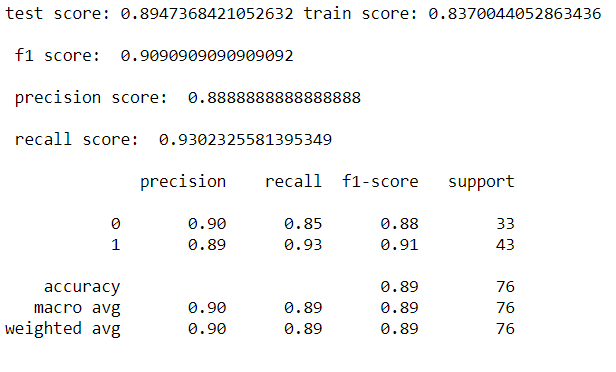
As the output variable to be predicted was a categorical variable, The supervised learning model of classification were chosen. The chosen algorithms were

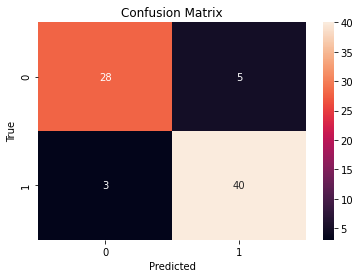
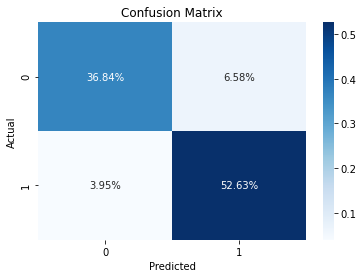
1. Logistic Regression
2. Navies Baies
3. Random Forest
4. Decision Tree

The data was split in 80-20 fashion for the model building with 80% as the train data and remaining 20% as the test data.

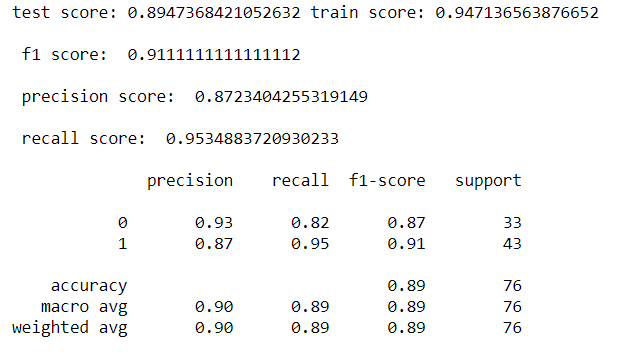
The accuracy, prediction and the confusion matrix of the Logistic Regression model is as follows

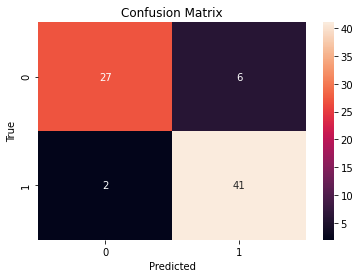
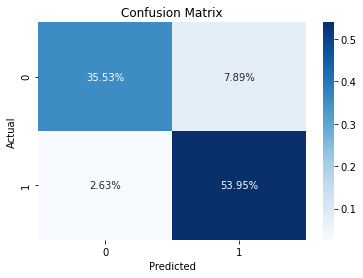


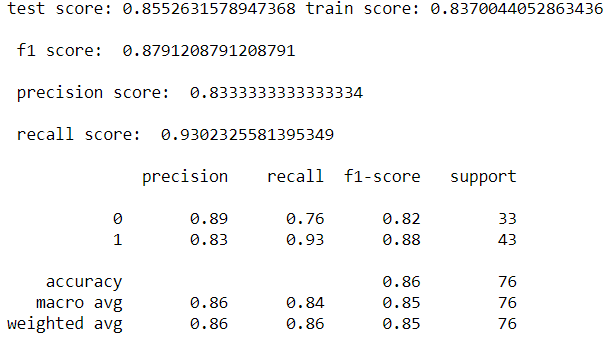
The accuracy, prediction and the confusion matrix of the Navies Baies model is as follows

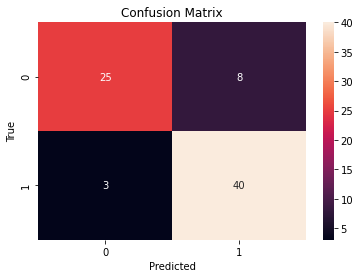
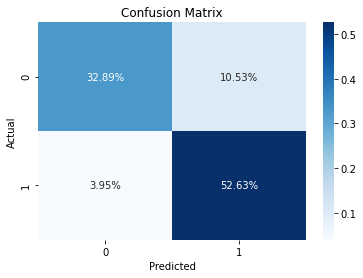


The accuracy, prediction and the confusion matrix of the Random Forest model is as follows





The accuracy, prediction and the confusion matrix of the Decision Tree model is as follows



The accuracy that was given by these 4 models are as follows

1. Logistic Regression – 92%
2. Navies Baies – 89%
3. Random Forest – 89%
4. Decision Tree – 86%

REFERENCES

The following are the links which were referred for the project

1. <https://www.hindawi.com/journals/cmmm/2013/27269>
2. <https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(10)60309-1/fulltext>
3. <https://www.sciencedirect.com/topics/nursing-and-health-professions/cardiovascular-parameters>
4. <https://www.ahajournals.org/doi/10.1161/01.CIR.97.18.1837>
5. <https://www.javatpoint.com/regression-vs-classification-in-machine-learning>
6. <https://medium.com/@dtuk81/confusion-matrix-visualization-fc31e3f30fea>
7. <https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset>
8. <https://medium.com/@dtuk81/confusion-matrix-visualization-fc31e3f30fea>
9. <https://archive.ics.uci.edu/ml/datasets/heart+disease>

APPENDICES

The codes for the project is as shown

