IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 1, January 2020

# Detailed Analysis of Coronary Artery Disease (CAD): Data Analytics and Machine Learning

**Shruthi Krishna Murthy[1], Priyanka P A[2], Deepak C R[3], Varshini G[4]**

Student, BE, Department of IEM, BIT, Bangalore, India[1]

Student, MBA (Business Analytics), CMS Business School, Bangalore, India[2]

Student, BE, Department of CSE, Brindavan College of Engineering, Bangalore, India[3]

BBA, Jain College CGS, Bangalore, India[4]

**Abstract:** Heart disease, also known as cardiovascular disease, is a general term for a variety of conditions that affect the heart and blood vessels. It is a chronic disease that can lead to serious events including heart attack and death. Heart disease is one of the leading causes of death in Canada and worldwide. The most common form of heart disease is Coronary Artery Disease (CAD) caused by atherosclerosis. Some heart attacks cause very little damage to the heart muscle and the heart can still pump strongly. Some heart attacks are larger and the muscle damage causes a weak heart. There are several heart tests that measure the strength of the heart such as an echocardiogram (an ultrasound of the heart that looks at the pumping strength of the heart and how the heart valves work), nuclear scans such as a MUGA scan, or a ventriculogram which is commonly done during an angiogram. After many lab tests and investigations data has been tabulated and ready to be analysed. Any acquired/ given data can be analysed and conclusions drawn accordingly. The acquired or given data usually exists in its crude or raw state. In our assignment, the acquired data consists of many physiological parameters which directly or indirectly lead to this disease. Data pre-processing helps to format the data into useful form by removing redundancy and noise, eliminating missing and non-numerical values, and also by normalization. Data analysis and visualization are carried out to improve the statistical analysis of given data. Logistic regression is carried out on the data since it contains lot of columns with categorical values. Accuracy, precision, and f1 score of the model have been measured. Various conclusions can be drawn from this interdependent data set and can be stored as historical data for future analysis.

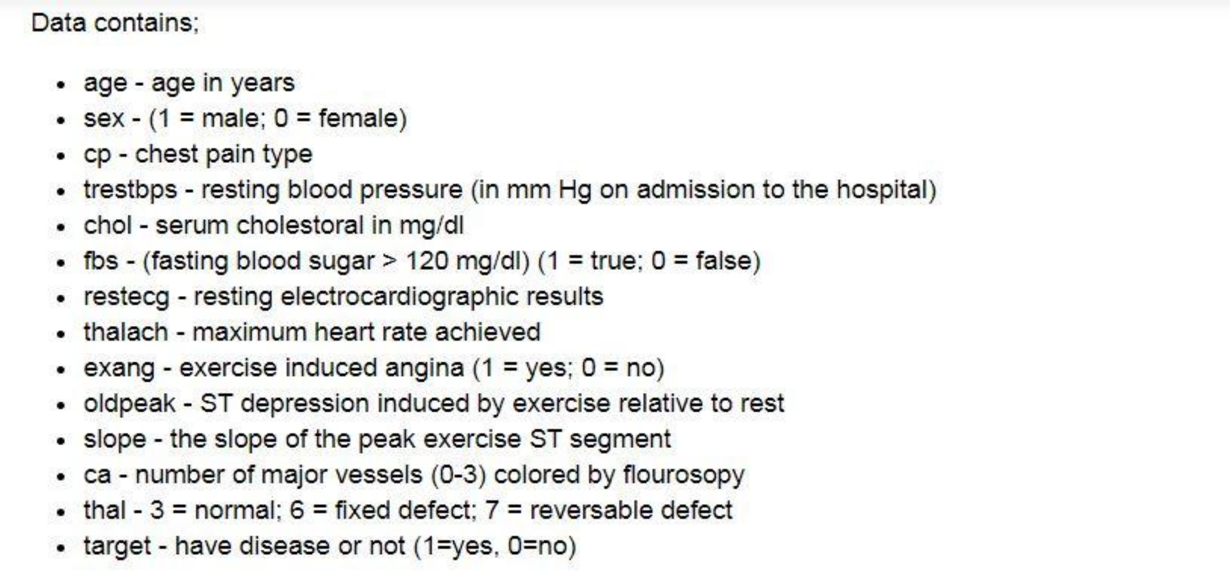
**Keywords:** Coronary Artery Disease (CAD), Machine Learning, Data pre-processing, Logistic regression, accuracy, precision, and f1 score, atherosclerosis, physiological parameters, echocardiogram, MUGA and ventriculogram

## I. INTRODUCTION

As of 2010, CAD was the leading cause of death globally resulting in over 7 million deaths. This increased from 5.2 million deaths from CAD worldwide in 1990. It may affect individuals at any age but becomes dramatically more common at progressively older ages, with approximately a tripling with each decade of life. Males are affected more often than females. The heart muscle, like every other part of the body, needs its own oxygen-rich blood supply. Arteries branch off the aorta and spread over the outside surface of the heart. The Right Coronary Artery (RCA) supplies the bottom part of the heart. The short Left Main (LM) artery branches into the Left Anterior Descending (LAD) artery that supplies the front of the heart and the Circumflex (Cx) artery that supplies the back of the heart. Over time, plaque builds up on the inside wall of arteries. Plaque is made of several substances including cholesterol. This build up is called atherosclerosis or hardening of the arteries. It can start at an early age and is caused by a combination of genetic and lifestyle factors that are called risk factors. Atherosclerosis can cause a narrowing in the arteries to various parts of the body such that blood flow is slowed or blocked. Poor blood flow to the brain can cause a stroke. Poor blood flow to the arms or legs is called Peripheral Artery Disease (PAD). Poor blood flow to the heart is called Coronary Artery Disease (CAD) and can cause angina or a heart attack. If the heart is starving for blood and not getting enough oxygen for more than 20 minutes, then a part of the heart muscle dies causing some permanent damage. This is called a heart attack or Myocardial Infarction (MI). Heart attacks are confirmed with 3 blood tests and a test that shows the electrical activity of the heart called an Electrocardiogram (ECG). Our analysis consists of many attributes as shown in figure 1. Coronary artery disease has a number of well determined risk factors. These include high blood pressure, smoking, diabetes, lack of exercise, obesity, high blood cholesterol, poor diet, depression, family history, and excessive alcohol. About half of cases are linked to genetics. Smoking and obesity are associated with about 36% and 20% of cases, respectively. Smoking just one cigarette per day about doubles the risk of CAD. Lack of exercise has been linked to 7–12% of cases. Exposure to the herbicide Agent Orange may increase risk. Rheumatologic diseases such as rheumatoid arthritis, psoriasis, and psoriatic arthritis are independent risk factors as well. If these parameters are well assessed at the beginning and well treated, can easily save thousands of lives. Many lives have already been

DOI 10.17148/IJARCCE.2020.9129

saved by systematic follow-up procedures. Post Heart attacks can also be a challenging situation and we have considered the same.



Data contains;

- age - age in years
- sex - (1 = male; 0 = female)
- cp - chest pain type
- trestbps - resting blood pressure (in mm Hg on admission to the hospital)
- chol - serum cholestoral in mg/dl
- fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg - resting electrocardiographic results
- thalach - maximum heart rate achieved
- exang - exercise induced angina (1 = yes; 0 = no)
- oldpeak - ST depression induced by exercise relative to rest
- slope - the slope of the peak exercise ST segment
- ca - number of major vessels (0-3) colored by flourosopy
- thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
- target - have disease or not (1=yes, 0=no)

Figure 1 shows the attributes of our interest.

## II. PROBLEM STATEMENT

Data has to be acquired from reports. Data analysis and visualization needs to be carried out for statistical and graphical analysis of the acquired data. Logistic regression needs to be carried out on the data set (categorical). Accuracy, precision, and f1 score of the model to be measured. Conclusions to be drawn from the prepared report.

## III. METHODOLOGY

A. Importing Libraries [2]

Figure 2 shows the Python code to import libraries. We have used three libraries

- ‘numpy’ is a package for scientific computing with Python. This library is imported as ‘np’ and will be used throughout the project.
- ‘pandas’ is for data manipulation and analysis. panadas is an open source, BSD- licenced library providing easy-to-use data structures and data analysis tools. pandas is imported as pd.
- ‘matplotlib.pyplot’ is a collection of command style functions that make matplotlib work like MATLAB. It is imported as plt
- ‘seaborn’ is a Python data visualization library based on matplotlib for attractive and informative statistical graphics.

B. Importing data

Figure 3 shows the Python code to import data from respective directory/ file and assigning it to DataFrame df. The data stored in CSV format is being imported. [3] [4]

C. Checking for NaN

It is very essential in data pre-processing to check for NaN. In this attempt we could identify few NaN.

D. Manipulating NaN values

It is essential to remove the NaN values. This can be done by

- Removing the entire column containing many NaN values
- Forward fillna method
- Backward fillna method
- Mean method

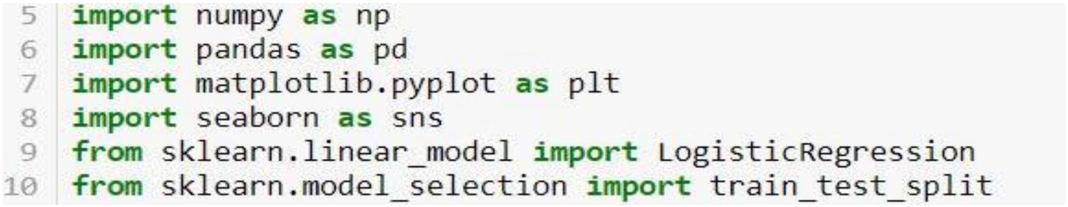
E. Data Visualization and data exploration needs to be carried out, which is explained in the next section.

F. Plotting a Heatmap

Correlation between the fields of the recorded data is analysed by plotting a heatmap. The values may be negative or positive and the magnitude plays a key role in designing various predictive models in AI. Figure 4 shows the heatmap of CAD

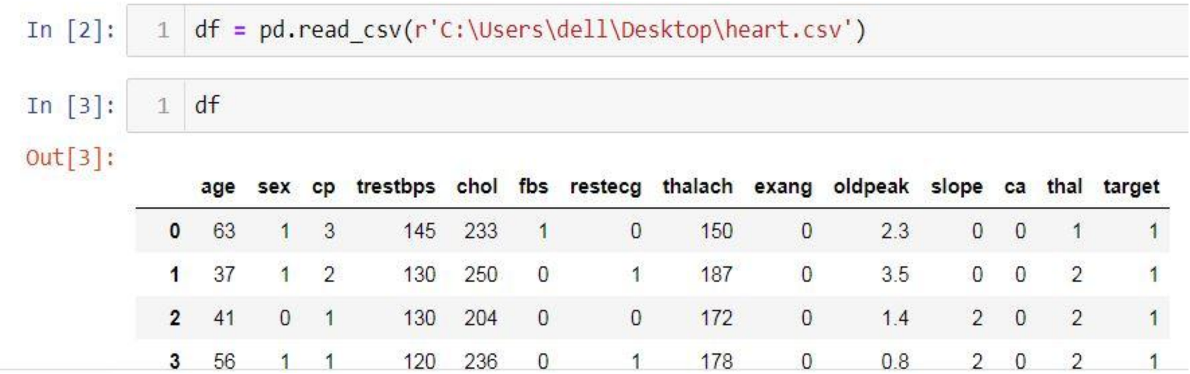
G. Splitting the data into train and test sets. Figure 5 shows the python code to split the data set into train and test data.

H. Applying logistic regression on the split data. Figure 6 shows logistic regression on given data set. Figure 7 shows an effort to normalize the dataset.



```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

Figure 2 shows the Python code to import libraries.



```
In [2]:  df = pd.read_csv(r'C:\Users\dell\Desktop\heart.csv')

In [3]:  df

Out[3]:
     age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak  slope  ca  thal  target
  0   63    1   3       145   233    1        0      150      0      2.3      0   0     1       1
  1   37    1   2       130   250    0        1      187      0      3.5      0   0     2       1
  2   41    0   1       130   204    0        0      172      0      1.4      2   0     2       1
  3   56    1   1       120   236    0        1      178      0      0.8      2   0     2       1
```

Figure 3 shows the technique of forward fillna method.

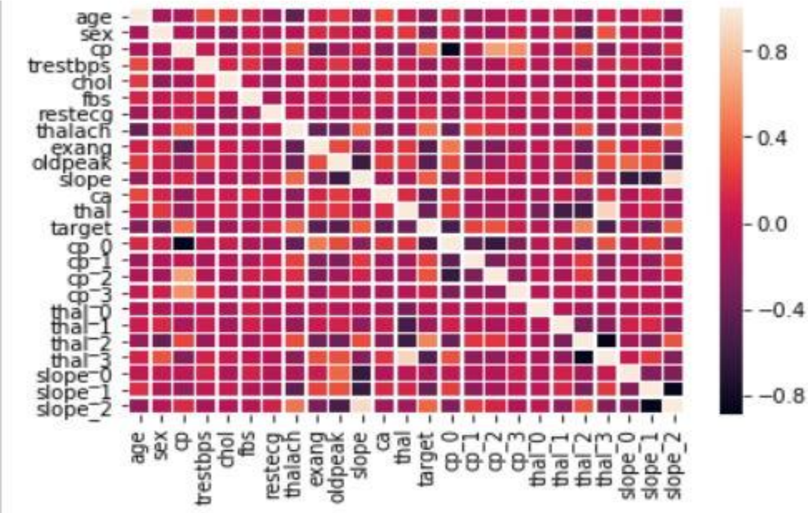
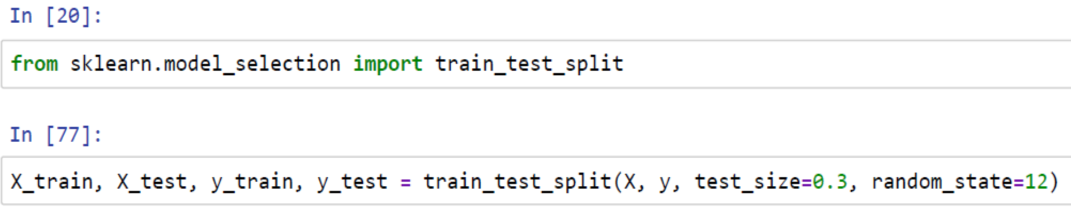


Figure 4 shows the heatmap of the assignment.

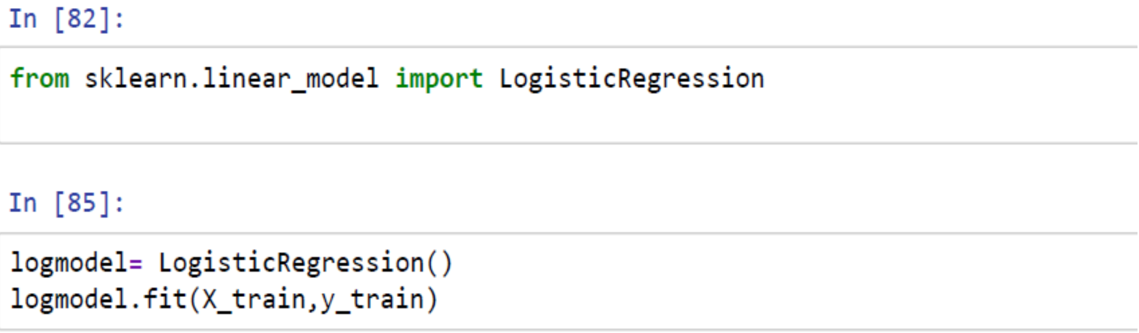


```
In [20]:
from sklearn.model_selection import train_test_split

In [77]:
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=12)
```

Figure 5 shows the python code to split the data set into train and test data.

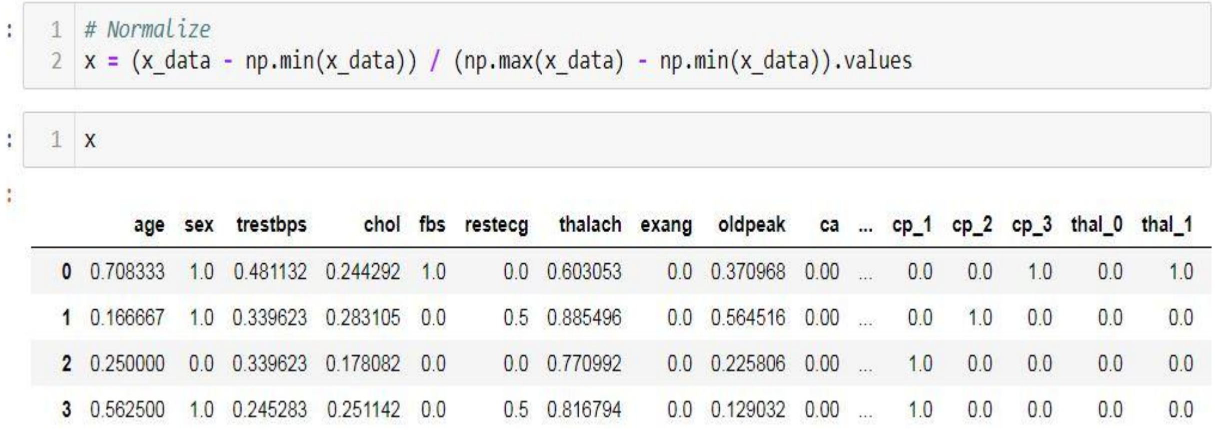
DOI 10.17148/IJARCCE.2020.9129



```
In [82]:
from sklearn.linear_model import LogisticRegression

In [85]:
logmodel= LogisticRegression()
logmodel.fit(X_train,y_train)
```

Figure 6 shows logistic regression on given data set.



```
# Normalize
x = (x_data - np.min(x_data)) / (np.max(x_data) - np.min(x_data)).values
```

```
x
```

| | age | sex | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | ca | ... | cp_1 | cp_2 | cp_3 | thal_0 | thal_1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.708333 | 1.0 | 0.481132 | 0.244292 | 1.0 | 0.0 | 0.603053 | 0.0 | 0.370968 | 0.00 | ... | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 |
| **1** | 0.166667 | 1.0 | 0.339623 | 0.283105 | 0.0 | 0.5 | 0.885496 | 0.0 | 0.564516 | 0.00 | ... | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| **2** | 0.250000 | 0.0 | 0.339623 | 0.178082 | 0.0 | 0.0 | 0.770992 | 0.0 | 0.225806 | 0.00 | ... | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **3** | 0.562500 | 1.0 | 0.245283 | 0.251142 | 0.0 | 0.5 | 0.816794 | 0.0 | 0.129032 | 0.00 | ... | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Figure 7 shows an effort to normalize the dataset.

## IV. DATA VISUALIZATION

Data visualization is an integral part of data analytics and Machine Learning. When there is a huge data set, manual analytics becomes almost impossible. Data visualization plays a vital role in analysis in such situation. It involves use of various plots – bar graph, pie charts, box plots, line graphs and many more. Figure 8 and figure 9 includes a scatter graph of Max heart rate vs. Age and a plot of FBS respectively. Figure 10 shows the count plot of age. Figure 11 shows the hue plot of sex v/s. Frequency. Figure 12 shows the line plot of age v/s. fbs.

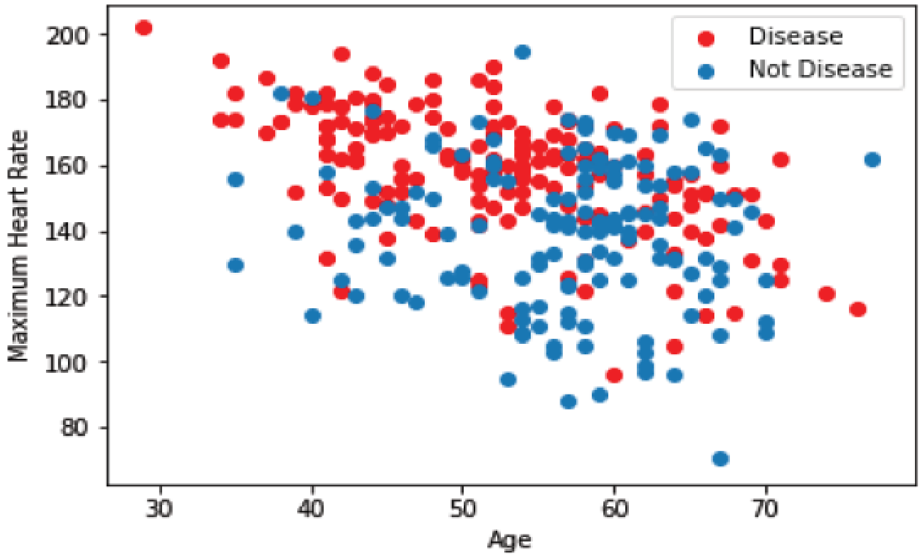


Figure 8 shows the scatter plot of age v/s. maximum heart rate.

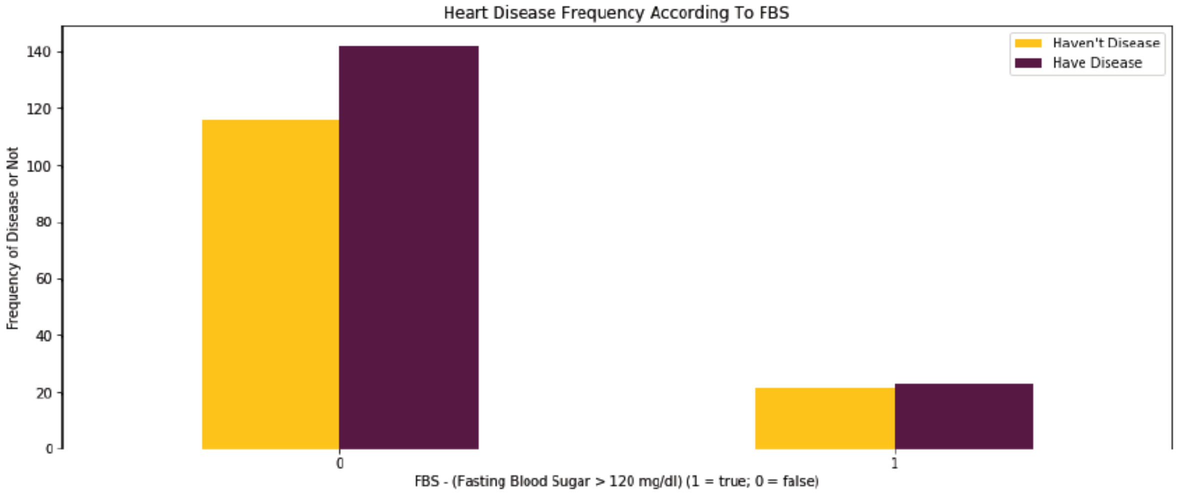
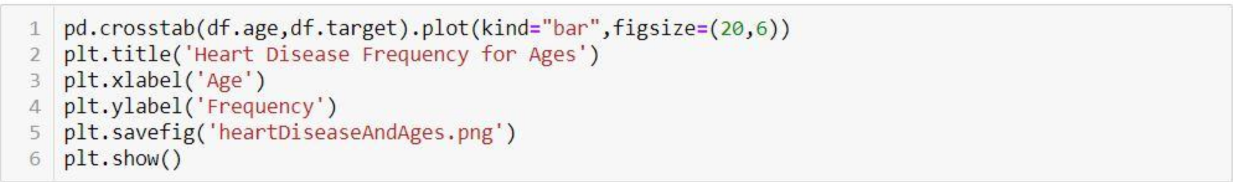


Figure 9 shows a bar graph of FBS.



```
pd.crosstab(df.age,df.target).plot(kind="bar",figsize=(20,6))
plt.title('Heart Disease Frequency for Ages')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.savefig('heartDiseaseAndAges.png')
plt.show()
```

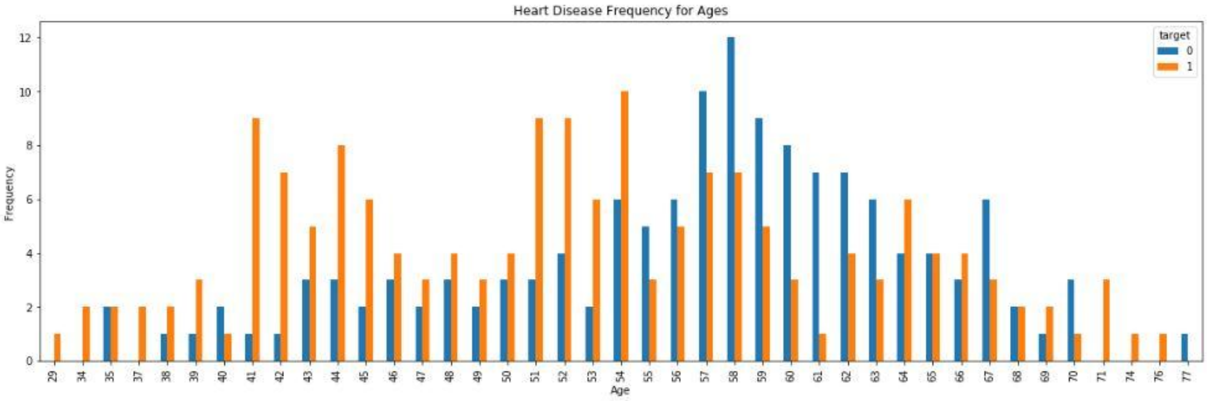


Figure 10 shows the count plot of age.

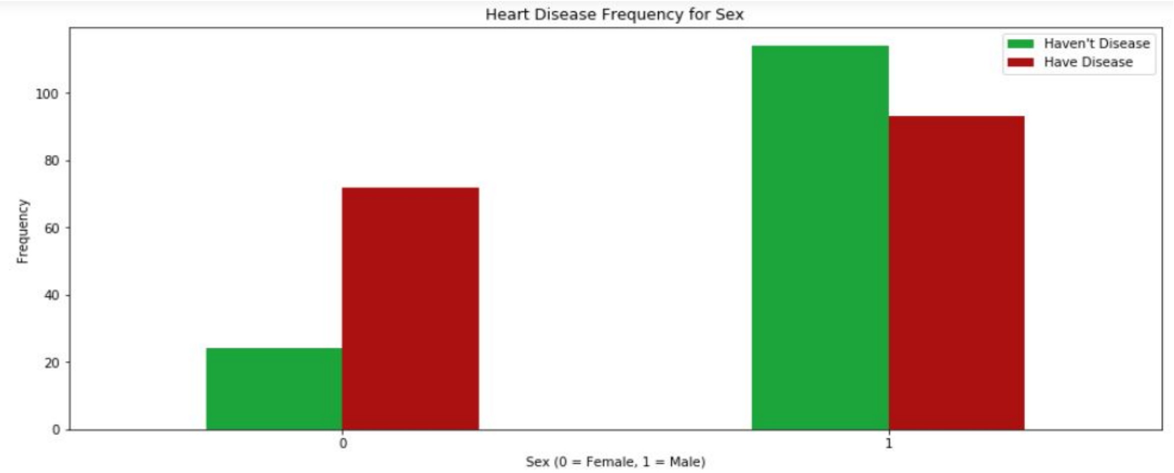


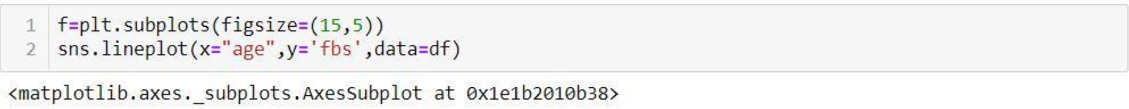
Figure 11 shows the hue plot of sex v/s. Frequency.

DOI 10.17148/IJARCCE.2020.9129

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 1, January 2020



```
f=plt.subplots(figsize=(15,5))
sns.lineplot(x="age",y='fbs',data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1e1b2010b38>
```

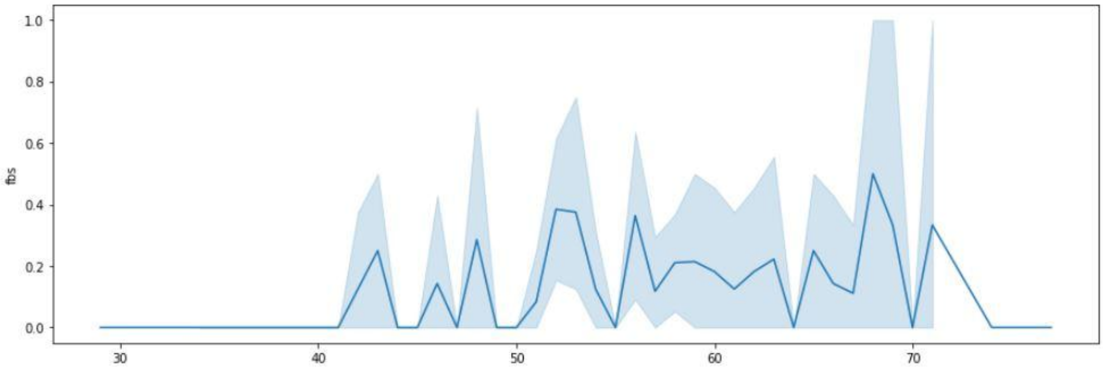
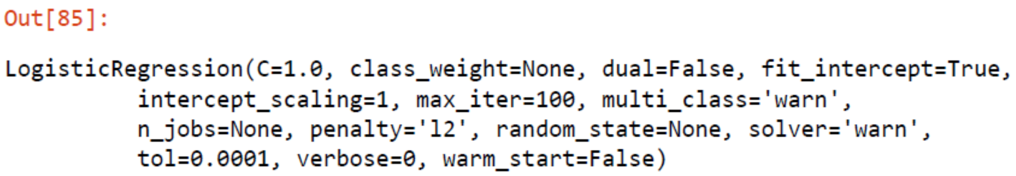


Figure 12 shows a line plot of age v/s. fbs.

## V. RESULTS

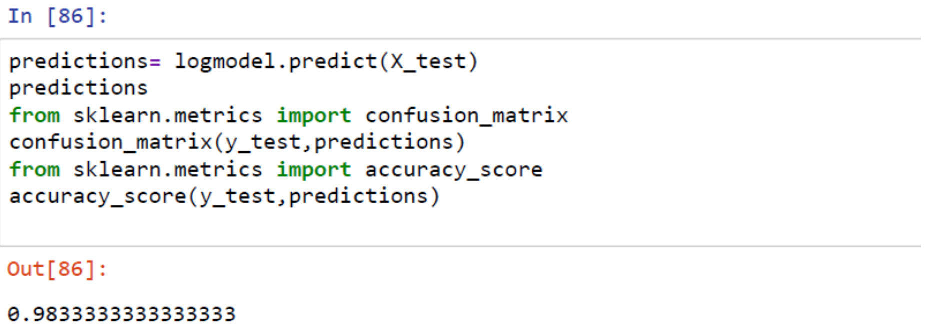
After analysing the heatmap and figuring out the correlation between different columns/ physiological parameters, Logistic regression needs to be carried out to create a prediction model. Figure 13 shows the results of logistic regression model. Figure 14 shows the Accuracy score of the designed model. From this data, precision, f1 score and reliability can be calculated. Figure 15 shows the confusion matrix of Logistic Regression to calculate accuracy, precission and f1 score.



```
Out[85]:

LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='warn',
          n_jobs=None, penalty='l2', random_state=None, solver='warn',
          tol=0.0001, verbose=0, warm_start=False)
```

Figure 13 shows the results of logistic regression model



```
In [86]:

predictions= logmodel.predict(X_test)
predictions
from sklearn.metrics import confusion_matrix
confusion_matrix(y_test,predictions)
from sklearn.metrics import accuracy_score
accuracy_score(y_test,predictions)

Out[86]:

0.9833333333333333
```

Figure 14 shows the Accuracy score of the designed model.

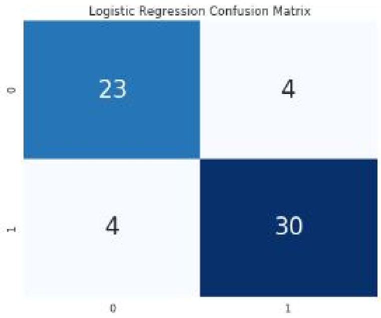


Figure 15 shows the confusion matrix of Logistic Regression to calculate accuracy, precission and f1 score.

Copyright to IJARCCE

DOI 10.17148/IJARCCE.2020.9129

## VI. CONCLUSIONS

21st century is the era of data explosion and manual analysis of such huge data is almost impossible. In our assignment we have used data analysis and Machine Learning algorithms to make predictions in seconds. Also the accuracy of these data can be verified. Data analysis and visualization was carried out for statistical and graphical analysis of the acquired data. Logistic regression was carried out on the data set (categorical). Accuracy, precision, and f1 score of the model was measured.

## REFERENCES

[1]. Interactions between kidney disease and diabetes- dangerous liaisons- Roberto Pecoits-Filho, Hugo Abensur, Carolina C.R. Betônico, Alisson Diego Machado, Erika B. Parente, Márcia Queiroz, João Eduardo Nunes Salles, Silvia Titan and Sergio Vencio- 2016- article 50.

[2]. The Python Standard Library — Python 3.7.1rc2 documentation https://docs.python.org/3/library/

[3]. Data Warehousing Architecture and Pre-Processing- Vishesh S, Manu Srinath, Akshatha C Kumar, Nandan A.S.- IJARCCE, vol 6, issue 5, May 2017.

[4]. Data Mining and Analytics: A Proactive Model - http://www.ijarcce.com/upload/2017/february-17/IJARCCE%20117.pdf

[5]. A comparative analysis on linear regression and support vector regression- DOI: 10.1109/GET.2016.7916627- https://ieeexplore.ieee.org/abstract/document/7916627

## OUR GUIDE

**VISHESH S (BE, MBA, PGDIB, (MTech))** born on 13th June 1992 hails from Bangalore (Karnataka) and has completed B.E in Telecommunication Engineering from VTU, Belgaum, Karnataka in 2015. He also worked as an intern under Dr. Shivananju BN, former Research Scholar, Department of Instrumentation, IISc, Bangalore. His research interests include Embedded Systems, Wireless Communication, BAN and Medical Electronics. He is also the Founder and Managing Director of the corporate company Konigtronics Private Limited. He has guided over a thousand students/interns/professionals in their research work and projects. He is also the co-author of many International Research Papers. He has recently completed his MBA in e-Business and PG Diploma in International Business. Presently Konigtronics Private Limited has extended its services in the field of Software Engineering and Webpage Designing. Konigtronics also conducts technical and non-technical workshops on various topics. Real estate activities are also carried out under the guidance of Siddesh B S BE (civil). Vishesh S along with his father BS Siddesh has received various awards and applauses from the scientific and entrepreneurial society. He was appointed as the MD of Konigtronics Pvt Ltd (INC. on 9th Jan 2017) at an age of 23 years. His name is indexed in various leading newspapers, magazines, scientific journals and leading websites & entrepreneurial forums. He is also the guide for many international students pursuing their Masters.