**Seminar in Research and Research Methodology**

**INFO 5082**

**Research Project**

Sowmya Valandas

**Framingham Heart Study**

**INTRODUCTION**

As indicated by most recent insights from the World Health Organization (WHO), cardiovascular disease and stroke are the world's greatest killers, representing 15.2 million deaths in 2016. These sicknesses have remained the main sources of death universally in the most recent 15 years already it was. Around 787,650 Americans die from coronary illness every year—that is 1 in every 3 deaths. Cardiovascular infection is the main source of death in industrialized nations. This not only affected individuals but also their lifestyle. Moreover, it showed an impact on the country’s economy and public health costs. Diagnosis of heart disease is one of the costliest decisions in the health care industry.

Medical Information is the applied science which is a combination of both Medicine disciplines and information technology, which helped to improve both qualities of care and effectiveness. One of the largest sectors in the United States is Healthcare. Each year the US invests around $2.5 trillion in healthcare. However, each year about $200 billion is wasted as fraud or abuse. Data mining techniques help to reduce this wastage and improve decision making.

**PROBLEM STATEMENT**

Worldwide, a heart attack is one of the major causes of people’s death. Around 75 percent of the deaths occurred in developing countries. If proper measures are not taken, it is estimated that by 2030 about 20 million people will be dying from cardiovascular disease.The Healthcare industry gathers vast amounts of data. Data mining techniques are applied to this data to derive some insightful information. This can help the doctor on how to treat a specific patient. Further, it can help in reducing deaths due to heart diseases.

**REVIEW OF LITERATURE**

Many authors performed different algorithm to predict the heart attacks. Shanta Kumar and Kumaraswamy proposed an intelligent and effective heart attack prediction model using artificial neural networks. Anbarasi et al.in used genetic algorithm, their results were more accurate with feature selection method. They used Naïve Bayes, Clustering and Decision Tree methods. In the paper proposed by k.Rajeswari and colleagues used Neural Networks to predict the heart disease.

In this project, I used supervised learning algorithms like KNN, logistic, Random Forest and selected best model based on False Negative Rate.

**OBJECTIVE**

The main objective of this project is to create a Supervised Learning Algorithm like Logistic Regression, K-nearest neighbor’s (KNN), Random forest is implemented on the Framingham heart disease dataset and choose the best model to predict if a specific person has a chance to get a heart attack or not.

.

**DATA COLLECTION**

It is a process of collecting information from different sources considering the variables of interest which helps to answer the research questions and test hypotheses. In this project, we used the **‘Framingham Heart Study’**, Framingham is a city in Massachusetts. This dataset consists of the information of the residents, which is based on the various reasons which are responsible for coronary diseases. This study has begun back in 1948 at present it is the third-generation participants. Data is collected by using blood samples of Families of Framingham which help to understand the various factors that are responsible for the risk of heart disease. Scientists are now working to crack down the new reasons for cardiovascular diseases. In this project, we create a model with higher accuracy which can detect the patients who have the potential to get a heart attack.

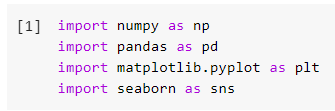
The dataset consists of around 4500 records of people who participated in this study and consists of 16 attributes that are responsible for cardiovascular disease. Attributes are listed below.

1. Gender (Binary): 0-Male, 1-Female
2. Age (Numeric): Patient’s age
3. Education: Education level of patient
4. Current Smoker (Binary): 1-Smoker, 0-Non-smoker
5. CigsPerDay (Numeric): Number of cigarettes smoked per day
6. BPMeds (Binary): 0-Not on blood pressure medication, 1-On blood pressure medication
7. PrevalentStroke (Binary): 0- Experienced stroke, 1- No stroke
8. PrevalentHyp (Numeric): 1- has hypertension, 0-No hypertension
9. Diabetes (Binary): 0-Has diabetes, 1-No diabetes
10. TotChol (Numeric): Cholesterol level in the body, represented in mg/dl.
11. sysBP (Numeric): Systolic measure of blood pressure (low blood pressure)
12. diaBP (Numeric): Diabolic measure of BP (high blood pressure)
13. BMI (Numeric): Body Mass Index.
14. HeartRate (Numeric): It is a measure of beats per minute.
15. Glucose (Numeric): Glucose content in blood, measured in mg/dl
16. Target(binary): 0-No heart attack, 1- heart attack.

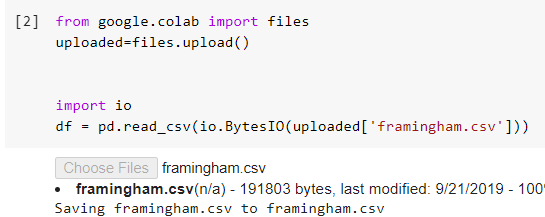
**Exploratory Data Analysis:**

It is an approach to analyze the dataset and retrieve the main characteristics. Mostly visualization models are used to explain the insights. This mainly focuses on model fitting, hypothesis testing, handling missing values, and changing the variables as required for the model.

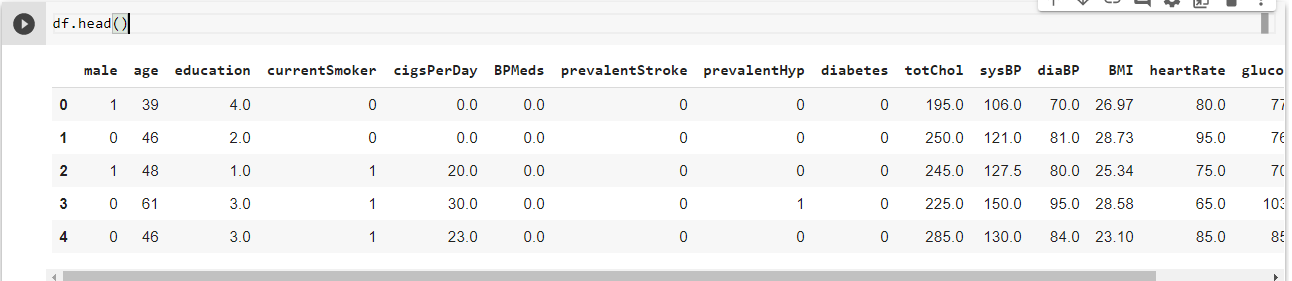
Firstly, load the file and read the file to see the attributes.



**Fig: 1** Importing libraries

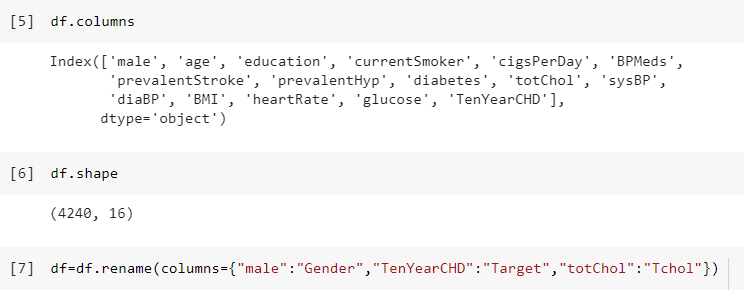


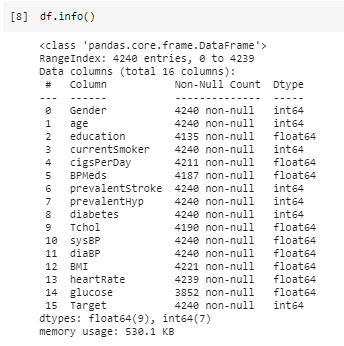
**Fig: 2** Reading excel file



**Fig: 3** viewing all the columns

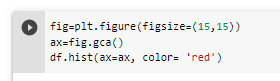
Renaming the columns ‘male’, ‘TenyearCHD’ and ‘totChol’ as ‘Gender’, ‘Target’, and ‘Tchol’ respectively.

 **Fig: 4** Renaming the columns



**Fig:5** type and non-null columns

To understand the distribution of each column, plotted a histogram for all the columns.



**Fig: 6** Histogram for all the variables

Chart

Description automatically generated **Fig: 7** Histogram for variables

Chart, waterfall chart, box and whisker chart

Description automatically generated

**Fig: 8** Histogram for variables

From the Fig:7 and Fig: 8 The majority of patients are in between 40 to 65 age. In most people who attended the test are men, whose count is around 2500. Further, 2000 people in the test are having glucose level’s normal, which ranges from 90 to 120mg/dl. Majority of people are not having any prevalent strokes. Half of the people are having the high cholesterol levels.

30% people are having normal diabolic pressure which is less than 80 mg/dl and 60% are having the normal systolic blood pressure.

**A picture containing text

Description automatically generated**

**Fig: 9** Percentage of people with and without disease

From Fig 9, we can say that number of people without heart disease are more than people with heart disease. People without heart disease are 84.81% and with heart disease are 15.19%.

**Chart, bar chart

Description automatically generated**

**Fig: 10** Bar plot of dependent variable

Fig 10 is the graphical representation of Target variable. ‘0’ represents who do not have heart disease and ‘1’ represents with heart disease. Around 3500 members do not have cardiovascular disease and about 500 members have heart disease. As above 75% do not have heart disease. This is the main detect to train the model as it predicts the majority class. Data is imbalanced, we use **SMOTE** method to balance the data. We will discuss it further below.

**Chart, bar chart

Description automatically generated**

**Fig: 11** Gender Vs Target variable histogram

From the fig 11, males have more tendency to get cardiovascular disease when compared to males. Male population in the dataset is more than females. However, both have same number of people with heart disease.

**A picture containing text

Description automatically generated**

**Fig: 12** Histogram for various variables Vs target variable

**Chart, bar chart, waterfall chart

Description automatically generated**

**Fig: 13** Histogram for current smoker, BPMeds Vs Target

From the above figure current smoker Vs Target, we can conclude that heart disease does not depend on smoking habit as people without smoking habit have similar chance of getting heart attack as people with smoking habit. From the graph BPMeds Vs Target, data under without BP medication is more. People with BP medication has less probability of getting a cardiovascular disease.

**Chart, bar chart

Description automatically generated**

**Fig: 14** Histogram for diabetes, prevalentHyp Vs Target

From the plot 14, we can say that patients who do not have diabetes has less chance of getting a heart attack. Similarly, people without prevalent hypertension has less chance of getting a heart attack.

**Text, letter

Description automatically generated**

**Fig: 15** number of young, middle, elderly age people

In the dataset, there are more data points under middle age, which ranges from 40 to 55 years of age.

**Text

Description automatically generated**

**Fig: 16** pie plot for age column

Above is the pie plot showing young, middle, and elderly age sectors. Sector under middle age has more area. Sector under young age is the smallest.

**Correlation:**

It shows the statistical relationship between two variables. It represents the degree to which pair of variables are related. Fig 20 is the heatmap to represent the correlation values between the variables. From the heatmap we can say that features ‘Education’ and ‘current smoker’ are irrelevant. ‘PrevalentHyp’ and ‘sysBP’ are highly related.

**Graphical user interface, table

Description automatically generated**

**Fig: 17** correlation for all the variables

**Graphical user interface, text

Description automatically generated**

**Fig: 18** heat map code

**Chart, scatter chart

Description automatically generated**

**Fig: 19** heat map for correlation

**Checking for NULL values:**

From the figure 20, we can observe that columns ‘education’, ‘cigsPerDay’, ‘BPMeds’, ‘Tchol’, ‘BMI’, ‘heartRate’ and ‘glucose’ have NULL values.

**A picture containing text, receipt, screenshot

Description automatically generated**

**Fig: 20** Null values

**Exploratory data analysis (EDA)**

**IMPUTING EMPTY VALUES:**

Now, replacing the empty values with median or 0

Graphical user interface, text, application, chat or text message

Description automatically generated

**Fig: 21** Imputing Null values

Replacing columns cigsPerDay, glucose, heartRate, BMI, and Tchol with median. Our solution for the problem is irrelevant to education so replace empty values with 0.

After replacing NULL values, checking for NULL values to see if there are any NULL values.

From the Fig: 22 it is shown that there are no NULL values.

Table

Description automatically generated

**Fig:22** checking for Null values

**Checking for Outliers:**

We used box plot for checking the presence of ‘Outliers’. From the Fig 25, we can summarize that features ‘Tchol’, ‘sysBP’, ‘diaBP’, ‘BMI’, ‘heartrate’, and ‘glucose’ have outliers. Further, we used ‘Inter Quartile Range’ method to remove outliers.

****

**Fig: 23** boxplot on the columns to find outliers

**Chart, box and whisker chart

Description automatically generated**

**Fig: 24** boxplot for different variables

**Removing the Outliers:**

Text

Description automatically generated

**Fig: 25** code for IQR to remove outliers

**Box plot after removing outliers:**

Used IQR method to remove the outliers. Further, plotted boxplot to check if there are any outliers after removing them using ‘Inter Quartile Range’ method.

**Chart, box and whisker chart

Description automatically generated**

**Fig: 26** boxplot after removing outliers

**Feature Selection: ‘SelectKBest’ method**

It is a feature selection method for supervised learning method. It helps to find the best independent variables for the target variable. This method reduces overfitting and improves accuracy of the model. Implemented this method to extract the top 10 best features. This method uses chi square value. Used bar plot to visualize the feature selection plot.

From the Figure, we found that ‘cigsPerDay’, ‘age’, ‘sysBP’, ‘Tchol’, ‘prevalentHyp’, ‘diaBP’, ‘Gender’ and ‘BMI’ are important for our model.

Text

Description automatically generated

**Fig: 27** Code to select top 10 best variables

Text, letter

Description automatically generated

**Fig: 28** Box plot for different variables based on score value

Chart

Description automatically generated

**Fig: 29** box plot for feature importance

**Data Analysis:**

Considering the top 10 features as final features from the ‘SelectKBest’ method.

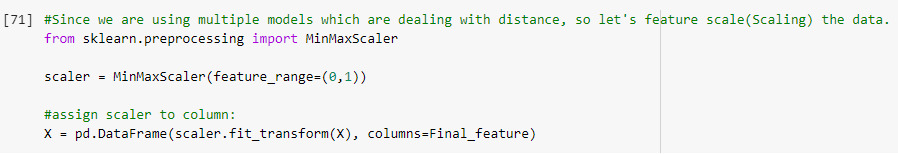
**Graphical user interface, application, Word

Description automatically generated**

**Fig: 30** final features

**MinMaxScaler:**

It transforms feature by scaling them in the given range. For our data, the range is from 0 to 1. As we are using multiple models which are dealing with distance, so let us feature scale (Scaling) the data. It helps to normalize the data (Independent variables) within a particular range.

****

**Fig: 31** minmax scaler

**Checking for class Imbalance:**

Checking class imbalance for the dependent variable ‘Target’. Class imbalance causes poor performance of the model. In our dataset we have a greater number of people who are not having cardiovascular disease. The below bar plot shows the class imbalance.

Chart, bar chart

Description automatically generated

**Fig: 32** barplot for class imbalance

**Splitting the data:**

Splitting the data as x\_train, x\_test, y\_train, y\_test.

Graphical user interface, text, application

Description automatically generated

**Fig: 33** splitting data as train and test

**SMOTE (Synthetic Minority Over Sampling Technique):**

One of the best approaches for dealing with the class imbalance dataset is oversampling the minority class. In our project we are using over sampling the minority class which is people with cardiovascular disease. In smote oversampling method there is no loss of data. It duplicates the minority class.

As shown below, we performed SMOTE on training Data set.

Text

Description automatically generated

**Fig: 34** Applying Smote to training dataset

The figure below shows the bar plot of dependent variable after applying SMOTE. There is no class imbalance.

Chart, bar chart

Description automatically generated

**Fig: 35** bar plot of Target variable after applying smote on train data set

**Performing Models:**

**Logistic Regression:**

Logistic Regression is a classification algorithm, used when the value of target variable is categorical. Imported all essential libraries from ‘sklearn’. In this I used ‘liblinear’ solver. It is a library used for large linear classification. Next, fit the model using the training dataset.

Graphical user interface, text, application

Description automatically generated

**Fig: 36** Logistic regression Importing libraries and model fitting

Calculating the accuracy for both training and testing dataset.

Accuracy for training dataset: 67%

Accuracy for testing dataset: 65%

Graphical user interface, text, application

Description automatically generated

**Fig: 37** accuracy for both train and test dataset

**Random Forest:**

It is a supervised learning model. Here I have taken ‘Max\_depth’ is 3 and ‘random state’ is 2.

Graphical user interface, text, application

Description automatically generated

**Fig: 38** Random forest Importing libraries and model fitting Text

Description automatically generated

**Fig: 39** accuracy for both train and test dataset

Calculating the accuracy for both training and testing dataset.

Accuracy for training dataset: 70%

Accuracy for testing dataset: 65%

Text

Description automatically generated

**Fig: 40** Accuracy**, Precision**

**K Nearest Neighbor (KNN):**

It is a supervised model can be applied for both classification and regression. I have used leaf size 30, metric is ‘minkowski’, p value 2 and n\_neighbors as 5.

**Graphical user interface, text, application

Description automatically generated**

**Fig: 41** ImportingKNN and fitting the training dataset

**Text

Description automatically generated**

**Fig: 42** Accuracy results for KNN

Calculating the accuracy for both training and testing dataset.

Accuracy for training dataset: 88%

Accuracy for testing dataset: 66%

**Data Visualization and Results Report:**

**Confusion Matrix:**

It is a table used to describe the performance of the model. In this project the ‘False negative’ rate should be low because the model predicts as the specific person will not suffer from heart attack when he will suffer from heart attack.

**A picture containing table

Description automatically generated**

**Fig: 43** confusion matrix

**Graphical user interface, text, application

Description automatically generated**

**Fig: 44** Specificity and precision

**AUC/ROC curve:**

The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes. More the Area under the curve better the model.

**Chart

Description automatically generated**

**Fig: 45** ROC/AUC curve

**Logistic Regression**

Text, letter

Description automatically generated

**Fig:46** Accuracy score for Logistic regression

Chart, treemap chart

Description automatically generated

**Fig: 47** Confusion matrix result for Logistic Regression

Here, the False Negative rate is low (47). FN is the most dangerous. This model is best as it has less FNR.

Text, letter

Description automatically generated

**Fig: 48** AUC cure code for Logistic Regression

Chart, line chart

Description automatically generated

**Fig: 49** AUC curve for Logistic Regression

Area under the curve is 0.665.

**Random Forest:**

Chart, treemap chart

Description automatically generated

**Fig: 50** Confusion matrix for Random forest

FN value is 55, it is less than all other values. It is a good model. However, Logistic Regression has much lower FN value.

Chart, line chart

Description automatically generated

Fig: 51 AUC curve for Random forest

Area under the curve is 0.638.

**K Nearest Neighbor (KNN):**

**Chart, treemap chart

Description automatically generated**

**Fig:** 52 Confusion matrix for KNN

FN value is 71 and it is lower than other values. Hence, the model is good. However, Logistic Regression has much lower FN value.

**Chart, line chart

Description automatically generated**

**Fig:** 53 AUC curve for KNN

Area under the curve is 0.598.

**Conclusion:**

As False Negative error is the most dangerous, Logistic Regression got less FNR when compared to other models. Therefore, Logistic Regression using ‘liblinear’ solver is the best model compared to other two. It got an accuracy of 67%.

**Future work:**

Collecting more data from other states improves the accuracy of the model. We can implement Hyperparameter tuning to find the best model which has low type 2 error.

**Bibliography**

Kyle E. Walker\*, Sean M. Crotty. Classifying high-prevalence neighborhoods for cardiovascular disease in Texas. Applied Geography, 2014; 57: 22-31, 2014.

<https://www.webmd.com/diabetes/qa/what-are-normal-blood-sugarlevels>

Jesmin N, Tasadduq I, Kevin ST, Yi-Ping Ph Ch. Association rule mining to detect factors which contribute to heart disease in males and females. Expert Systems with Application, 2013; 40(4): 1086-1093.

Niti Guru, Anil Dahiya and Navin Rajpal, “Decision Support System for Heart Disease Diagnosis using Neural Network”, Delhi Business Review,Vol.8, No.1, pp.99

A.V Senthil Kumar. Generating Rules for Advanced Fuzzy Resolution Mechanism to Diagnosis

Heart Disease. International Journal of Computer Applications, 2013; 77(11): 6-12.

Nitesh V. Chawla Department of Computer Science and Engineering University of Notre Dame IN 46530, USA DATA MINING FOR IMBALANCED DATASETS: AN OVERVIEW.

Very Simple Classification Rules Perform Well on Most Commonly Used Datasets Robert C. Holte (holte@csi.uottawa.ca) Computer Science Department, University of Ottawa, Ottawa, Canada K1N 6N5