ILLINOIS INSTITUTE OF TECHNOLOGY

HEART DISEASE PREDICTION AND ANALYSIS

Tejaswini NandaKumar

A20468281

(email id: tnandakumar@hawk.iit.edu)

Prof. Lulu Kang

Statistical Learning MATH 569

ABSTRACT

The Heart Disease is one of the most major cause for the increase in the death rate. It is estimated that 17.9 million people died from cardiovascular disease in 2016, representing 31% of all global deaths according to World health organization., out of these 80% are due to coronary heart diseases. It is found common in both men and women leading to the death of the individual. Diagnosis and prediction of heart related issues requires more precision and correctness as a small reluctance can lead to the fatigue or death of the person. A large volume of data is collected related to the different causes of heart disease and a decision is made which checks the assumption, determines the correlation among the explanatory variables and an efficient model is designed to analyze the causes of heart disease and a solution can found out earlier to save lives of individual. The Logistic regression model is designed to make sure that it efficiently predicts the causes of the heart disease and valid analysis is carried out to show that the prediction is accurate.

1. INTRODUCTION

The heart is the most major organ that pumps oxygenated blood throughout the body via arteries for the normal and healthy functioning of all other organs in our body, and removes the deoxygenated blood via veins. Cardiology is one of the most important and yet very difficult field of heath care. The Cardiovascular disease (CVD) is a class of disease that involves the heart or blood vessels, it includes coronary artery diseases such as angina and myocardial infraction that may be caused due to high blood pressure, diabetes, lack of exercise, high blood cholesterol, obesity. Cardiovascular disease is the number one cause of death

globally more people die annually from CVD's than any other cases irrespective of the gender. Hence heart disease can prove lethal if not detected in early stages. Considering this I chose heart disease analysis topic where the main objective is to construct a predictive model that will help to determine heart disease problem in a person beforehand.

The dataset used in this project is heart.csv obtained from Kaggle.com which is a open source site for various real life datasets, The Cleveland Heart Disease Data found in the UIC machine learning repository that contains of 14 variables (column) measured on 1025 individuals observation (rows). This consists of a 'target' variable that classifies the presence and absence of heart disease, represented by 1 and 0 indicating presence of heart disease and absence of heart disease respectively, the binary classification problem found in the dataset, exploring the logistic regression model which is capable to predict the target value based on the variables, which gives a prediction accuracy of 90% based on the logistic regression model.

2. DATASETS:

The first step to the working of the project is to collect the datasets and analyzing the obtained data set. The datasets made use are:

i. age - The person's age in years

ii. sex - person's gender 1 represents male and 0 for female.

iii. cp - chest pain, as there are different types of chest pain

- 0 - typical angina
- 1 - atypical angina
- 2 - non-anginal pain
- 3 - asymptomatic pain

iv. trestbps – the resting blood pressure measured in mm Hg

v. chol – Cholesterol level of a person measured in mg/dl

vi. fbs – fasting blood sugar level greater than 120 mg/dl

- 1 – true if greater than 120mg/dl
- 0 – false if less than 120mg/dl

vii. restcg – resting electrocardiographic measurement

- 0 – normal
- 1 – having ST-T wave abnormality
- 2 – showing definite left ventricular hypertrophy

viii. thalach – person's maximum heart rate achieved

ix. exang – exercise induced angina

- 0 - no
- 1 – yes

x. oldpeak – ST depression induced by exercise relative to rest

xi. slope – peak exercise ST segment

- 0 – upsloping
- 1 – flat
- 2 - downsloping

xii. ca – number of major vessels colored by fluoroscopy 0 – 4 vessels.

xiii. thal – Thallium stress test level

- 0 to 1- normal
- 2 – fixed defect
- 3 – reversable defect

xiv. target – Heart disease

- 0 – No
- 1 – Yes

Data quality issues:

This dataset is from 1988 from different parts of the world. There is no much information about the collection of this data, but this is a medical data which contains data of the patients suffering from cardiovascular problems and their types, however this dataset is clean.

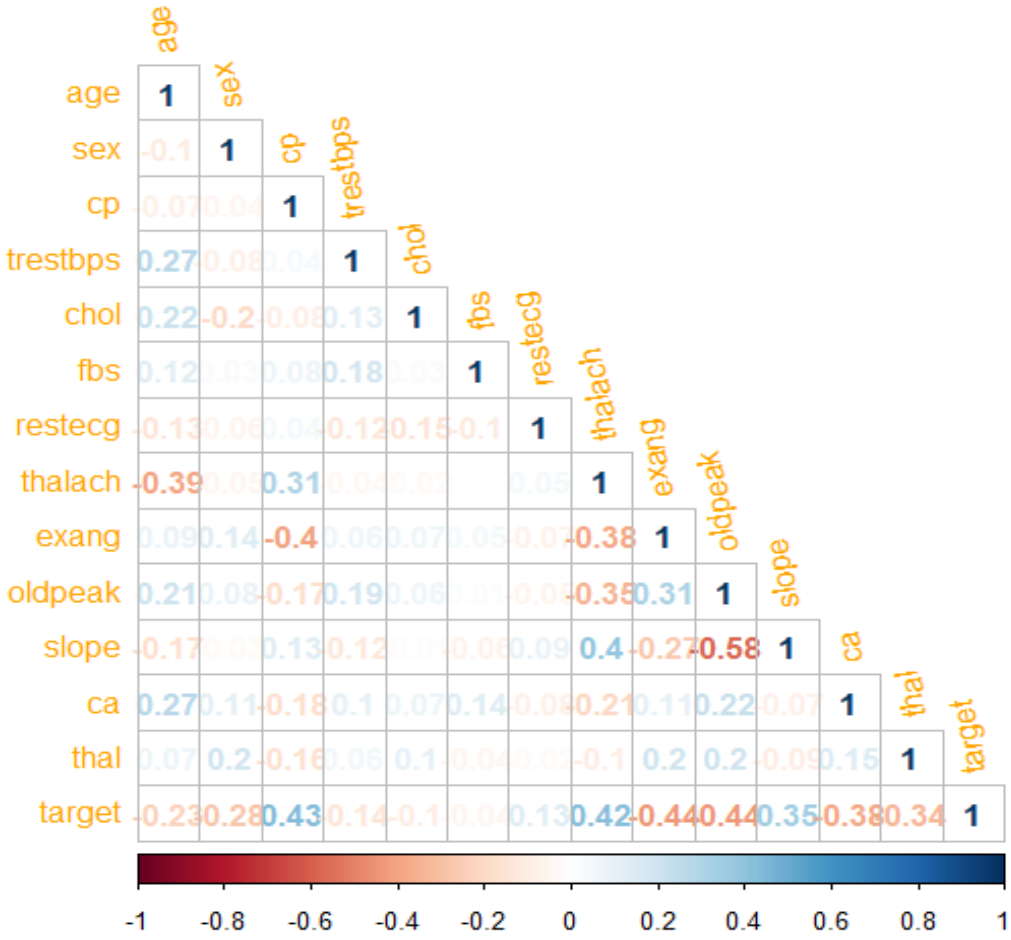
3. METHODOLOGY

The logistic regression model considered for predicting heart disease, should not have high correlation among the predictors, hence to find the collinearity between the regressor and target variable correlation test is performed, using the heterogeneous data correlation analysis, the following results are inferred. The positive correlations are depicted in blue and negative correlations in red, and the intensity of color directly is related to the grade of the correlation coefficients.

- we notice that there is a negative correlation of -0.28 between the target and the variable sex
- negative correlation of -0.44 between target and variable exang
- negative correlation of -0.44 between target and variable oldpeak
- negative correlation of -0.38 between target and variable ca
- negative correlation of -0.34 between target and variable thal
- there is a negative correlation of -0.4 between variable exang and cp

- there is a negative correlation of -0.58 between variable oldpeak and slope
- a positive correlation of 0.43 between variable taget and cp
- positive correlation of 0.42 between variable thalach and target
- positive correlation of 0.35 between target and variable slope.

The regressor variables sex, cp, thalach, exang, oldpeak, slope, ca and thal these variables show moderate positive or negative correlations; hence these variables play important role in predicting the target variable.



By plotting the relationship between the variables and target variable we can infer certain analysis of the type and cause of heart disease.

Relationship between the target and the regressor variables:

The relationship between the various regressor variables and the target variables are represented in the graphical model. The different graphs used here are boxplot, bar plot and dot plot, where the dot plots are used for univariate data, bar plot to compare different groups and box plot a way of summarizing a set of data measured on an interval scale.

1. target and chest pain (cp):

The cp type 0, 1, 2, 3 which are typical angina, atypical angina, non-anginal pain and asymptomatic types respectively.

The most predominant type of chest pain that is observed in the patients are “Atypical angina”, the next will be due to non-anginal pain and asymptomatic can also be expected to be the cause of chest pain. As the chest pain increases there is higher probability that the patient will have a heart problem.

Angina also called Angina Pectoris is a type of chest pain that is caused by reduced blood flow to the heart is a symptom of coronary artery disease. Angina pectoris is often described as squeezing, pressure, heaviness, tightness or pain in the chest. This is due to the narrowing of the arteries by the deposition of the fat. The Atypical angina is the pain that does not meet the criteria for angina known as atypical chest pain, when the heart muscle does not get an adequate supply of oxygenated blood. If the chest pain cannot be considered as angina, then that person is suffering from Atypical chest pain.

From the graph it is clearly understood that most of the persons are suffering from atypical angina.

2. target and thalach:

The normal resting heart rate should be 60 – 100 beats per minute, if heart beats fewer than 60 minutes is called bradycardia and if heart beats faster than 100 times a minute is called tachycardia

3. target and thal:

From the plot it is seen that, the level 2 is higher than among the 4 levels. “thal” is the Thallium stress test level. Thallium stress test also called cardiac or nuclear stress level is a nuclear imaging test that shows how well the blood flows into your heart when a person is at rest or when he/she is doing exercise. This test is carried out by introducing a small quantity of radioactive liquid called radioisotope into one of the veins, this radioisotope will flow through the bloodstream and end up in the heart. Once the radiation enters in the heart a special camera called gamma camera can detect the radiation and reveal any issues in the heart muscles. This test shows the size of heart chamber, ventricular function, if there’s any damage in the heart muscles.

4. target and ca:

There are 5 vessels in our heart indicated by 0 to 4 in the graph with 0 and 4 vessels leading to the major incident to have a heart failure. The 5 heart vessels are Superior vena cava, Inferior vena cava, pulmonary artery, pulmonary vein, and the aorta. In cardiac catheterization, fluoroscopy is used to help the healthcare provider to see the blood flow through coronary arteries to check for arterial blockages. Another test that detects any problem in heart like blockages are by using the Angiography exams, which treats the blood vessel diseases and conditions.

5. target and oldpeak:

The oldpeak refers to the ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment (1=upcoming, 2=flat, 3=downsloping). The ST depression seen in the ECG

below the baseline, which is often a sign of myocardial ischemia of which coronary insufficiency is a major cause of heart disease.

6. target and slope:

The slope of level 2 which is downsloping or horizontal ST depression >= 0.5mm at the J-point in >=2 contiguous leads indicates myocardical ischaemis.

7. slope and oldpeak:

There is more incidence of heart disease for patients with a slope of 2 and low oldpeak. An electrocardiogram (ECG or EKG) records the electrical signal from your heart to check for different heart conditions. Electrodes are placed on your chest to record your heart's electrical signals, which cause your heart to beat. The ECG detects the cardiac abnormalities by measuring the electrical activity generated by the heart as it contracts. The machine that records the patients ECG.

Flat, downsloping or depressed ST segments may indicate coronary ischemia. ST elevation indicates transmural myocardial infraction.

8. target and restecg:

restecg of 1 in the graph indicates having ST-T wave abnormality, looks very predominant in the people with few exceptions where people with normal level of 0 also are found to suffer from heart disease.

9. target and exang:

The level 0 indicates no exercise angina, a count with patients with heart disease compared to level 1.

10. exang and cp:

Typical angina also expresses exercise induces angina. That results in heart disease.

11. target and gender:

From the graph it can be seen that, the women are more prone to heart disease compared to the men. Hence, the gender plays a role in the symptoms, treatments and outcomes of coronary artery disease (CAD). A simple linear regression model indicates that the features that had a moderate correlation with the target variable also have significant p-values. This points towards a possible reduction in the needed number of variables when shaping an optimal model for prediction.

This graph shows the overall relationship between the target and all the other regressor variables.

4. IMPLEMENTATION:

The implementation of this project is carried out in a way such that, the heart disease model is divided into two one is the original model with all the significant and non-significant variables present and the other is the reduced model for which the heterogenous correlation is applied to find out the presence of non-significant variables, and they are removed. The analysis is carried out on both the models to find which is the best model that is helpful in Heart Disease Analysis.

LOGISTIC REGRESSION

I have made use of the Logistic regression model, to find out the best model to predict the heart disease state. Logistic regression models the probability of the default class. The Logistic regression model basically works on using the logistic function that mainly deals with the model that has a binary dependent variable here ‘target’ variable is the binary dependent variable, which outputs with two possible values either 0 or 1, the relationship between the regressors. The logistic regression being a predictive analysis is used to describe the data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio level independent variable. Here, it explains the relation between one dependent target variable with the predictor variables.

The assumptions for binary logistic regression that has to be considered are, it has to be made sure that there should be no high correlations (multicollinearity) among the predictors. The other point to be considered is when selecting the model for logistic regression analysis is model fit, it has to be made sure that adding independent variables to logistic regression model will always increase the amount of variance, adding more and more variables to the model can result in overfitting which reduces the generalizability of the model beyond the data on which the model is fit.

In this project I have brought the relationship between the regressors and the target variables along with the best model to predict the heart disease state.

The methodology is to compare between two models and to come out with the best model. The first is the original model which contains all the regressor variables, and the relationships between target is observed. The second model is based on the result of correlation (multicollinearity) where some of the regressor variables are not favorable to use in the comparison of the target and regressor variables, the variables like chol, fbs, and restecg are not significant for the model, hence these variables are removed and a significant model is formed, which is called the reduced model. Hence, a comparison of the original and reduced models is made to find the best among them.

Original Logistic Regression Model:

In this model includes all regressor variables in the regression equation with an AIC of 652.82. On applying Stepwise Algorithm, a stepwise regression algorithm is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. It can be performed using 3 main approaches they are: Forward selection, Backward elimination, Bidirectional elimination. Here backward elimination is considered where some of the non-significant variables are removed and a reduced model is formed. The AIC value of reduced model is 662.94, which seems to be more than the original model.

Reduced Logistic Regression Model:

This model contains some of the missing regressor variables, that are non-significant. The modified datasets contain 8 elements: sex, ca, thal, trestbps, exang, oldpeak, slope, cp. The result of the reduced logical regression model has an AIC value of 662.94, which is larger than the original model.

a). The original logistic regression model with AIC 652.82

b). The reduced logistic regression model with AIC 662.94

```
Call:
glm(formula = target ~ ., family = "binomial", data = heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8582  -0.2917   0.0718   0.4167   3.1908

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.081901   2.028691  -0.040 0.967797
age          0.026846   0.013950   1.924 0.054297 .
sex1        -1.992347   0.314204  -6.341 2.28e-10 ***
cp1          0.886380   0.308803   2.870 0.004100 **
cp2          2.006394   0.286281   7.008 2.41e-12 ***
cp3          2.409722   0.391965   6.148 7.86e-10 ***
trestbps    -0.024979   0.006537  -3.821 0.000133 ***
chol        -0.005462   0.002307  -2.367 0.017914 *
fbs1         0.380096   0.319620   1.189 0.234356
restecg1     0.397268   0.217975   1.823 0.068374 .
restecg2    -0.800417   1.536998  -0.521 0.602530
thalach      0.021692   0.006525   3.324 0.000886 ***
exang1      -0.750331   0.248746  -3.016 0.002557 **
oldpeak     -0.403411   0.132156  -3.053 0.002269 **
slope1      -0.595618   0.472076  -1.262 0.207057
slope2       0.799689   0.504500   1.585 0.112941
ca1         -2.334076   0.286781  -8.139 3.99e-16 ***
ca2         -3.597039   0.444870  -8.086 6.19e-16 ***
ca3         -2.288131   0.532138  -4.300 1.71e-05 ***
ca4          1.565677   0.930256   1.683 0.092363 .
thal1        2.796813   1.466219   1.908 0.056456 .
thal2        2.404646   1.421542   1.692 0.090727 .
thal3        0.991243   1.423972   0.696 0.486359
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1420.24  on 1024  degrees of freedom
Residual deviance:  606.82  on 1002  degrees of freedom
AIC: 652.82

Number of Fisher Scoring iterations: 6
```

(a). original model

```
Call:
glm(formula = target ~ ., family = "binomial", data = new_heart)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.95984  -0.32421   0.07717   0.43687   3.02162

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.033603   2.335768   1.299 0.194026
exang1      -0.826035   0.239940  -3.443 0.000576 ***
slope1      -0.814416   0.441517  -1.845 0.065098 .
slope2       0.760010   0.475434   1.599 0.109918
trestbps    -0.021459   0.005977  -3.590 0.000331 ***
oldpeak     -0.486974   0.126013  -3.864 0.000111 ***
thal1        2.743904   2.192343   1.252 0.210721
thal2        2.395335   2.157882   1.110 0.266982
thal3        0.955658   2.158966   0.443 0.658022
sex1        -1.704993   0.287189  -5.937 2.91e-09 ***
cp1          1.032386   0.301753   3.421 0.000623 ***
cp2          2.201356   0.278656   7.900 2.79e-15 ***
cp3          2.588232   0.383516   6.749 1.49e-11 ***
ca1         -2.354291   0.270340  -8.709  < 2e-16 ***
ca2         -3.241437   0.415447  -7.802 6.08e-15 ***
ca3         -2.403082   0.518453  -4.635 3.57e-06 ***
ca4          1.547255   0.874993   1.768 0.077010 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1420.24  on 1024  degrees of freedom
Residual deviance:  628.94  on 1008  degrees of freedom
AIC: 662.94

Number of Fisher Scoring iterations: 6
```

(b). reduced model

ANOVA Test:

The ANOVA or Analysis of Variance test is a statistical method that separates the observed variance data into different components to use for additional tests. The Anova is a statistical test that is used to check if the means of two or more groups are significantly different from each other.

The test for significance is performed using the ANOVA test, where the p-value is calculated using the chi-square distribution. The chi-square distribution is used to test the goodness of fit of the observed distribution, which tell how much difference exists between the observed value and the expected value. Hence, a low chi-square value means there is high correlation between the two sets of data.

On performing the ANOVA test on the original and reduced model:

original model:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: target

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      1024    1420.24
age       1   55.175      1023    1365.07 1.103e-13 ***
sex       1  105.254      1022    1259.81 < 2.2e-16 ***
cp        3  270.021      1019     989.79 < 2.2e-16 ***
trestbps  1   14.147      1018     975.64 0.0001691 ***
chol      1   10.564      1017     965.08 0.0011530 **
fbs       1    0.417      1016     964.66 0.5182896
restecg   2   11.196      1014     953.47 0.0037047 **
thalach   1   64.315      1013     889.15 1.060e-15 ***
exang     1   19.526      1012     869.62 9.925e-06 ***
oldpeak   1   59.080      1011     810.54 1.514e-14 ***
slope     2   14.117      1009     796.43 0.0008602 ***
ca        4  143.420      1005     653.01 < 2.2e-16 ***
thal      3   46.193      1002     606.82 5.160e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

reduced model:

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: target

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      1024    1420.24
exang     1  206.159      1023    1214.08 < 2.2e-16 ***
slope     2   91.979      1021    1122.10 < 2.2e-16 ***
trestbps  1   13.199      1020    1108.90 0.0002801 ***
oldpeak   1   73.720      1019    1035.18 < 2.2e-16 ***
thal      3  125.740      1016     909.44 < 2.2e-16 ***
sex       1   22.247      1015     887.20 2.397e-06 ***
cp        3   99.423      1012     787.77 < 2.2e-16 ***
ca        4  158.833      1008     628.94 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The result on performing the ANOVA test on the original and reduced model, the p-value chi-square distribution indicates that we fail to reject the null hypothesis which estates that both the models are equal. This implies that the reduced model is better than the original model with all the regressor variables.

5. ACCURACY EVALUATION:

On performing the accuracy evaluation, the original data Logistic regression modeling yielded an accuracy of 89% and the reduced dataset resulted with an accuracy of 90%.

Both the models are more specific than sensitive, which means that they predict better false negative cases.

a). Original dataset accuracy

b). Reduced dataset accuracy:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 86  8
         1 13 97

               Accuracy : 0.8971
                 95% CI : (0.847, 0.9351)
    No Information Rate : 0.5147
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7936

 Mcnemar's Test P-Value : 0.3827

            Sensitivity : 0.8687
            Specificity : 0.9238
         Pos Pred Value : 0.9149
         Neg Pred Value : 0.8818
             Prevalence : 0.4853
         Detection Rate : 0.4216
   Detection Prevalence : 0.4608
      Balanced Accuracy : 0.8962

       'Positive' Class : 0
```

(a). original model

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 88  8
         1 11 97

               Accuracy : 0.9069
                 95% CI : (0.8584, 0.943)
    No Information Rate : 0.5147
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8134

 Mcnemar's Test P-Value : 0.6464

            Sensitivity : 0.8889
            Specificity : 0.9238
         Pos Pred Value : 0.9167
         Neg Pred Value : 0.8981
             Prevalence : 0.4853
         Detection Rate : 0.4314
   Detection Prevalence : 0.4706
      Balanced Accuracy : 0.9063

       'Positive' Class : 0
```

(b). reduced model

As we are analyzing a human disease, a False Negative (the opposite error where the test result incorrectly fails to indicate the presence of a condition when it is not), is more dangerous than a False Positive (an error in binary classification in which a test result incorrectly indicates the presence of a condition such as a disease when the disease is not present) indicates the case when the prediction results in 1 when in reality is 0.

6. CONCLUSION:

The heart disease analysis method where, two models are tested original model with all the regressor variables and a reduced model with few of the regressor variables missing. The variables slope, exang, trestbps, oldpeak, thal, sex, cp and ca are the features that play a vital role in driving the prediction of the heart disease condition. Some of these variables are just symptoms, some of them are measurable medical variables and some of them can be either measurable or symptomatic. In my opinion this model is mostly good to uncover the significance of each variable in connection with the presence of heart disease, however it can make some suggestions to diagnose symptoms in order to identify life threatening problems faster. The reduced model variables allowed for a prediction accuracy of 90% based on the logistic regression model. Hence, the reduced model with only the significant variables can be chosen as the best model for Heart Disease Analysis. So, with models like this we should always be careful not to make any false negative decisions as this might lead to death of a person.

References:

[1]. Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach, p. 3277 2009.

[2]. Heart disease statistics, https://www.world-heart-federation.org/world-heart-day/world-heart-day-2019.

[3]. Prediction of Heart Disease and classifiers sensitivity analysis by Khaled Mohamad Almustafa.

[4]. Prediction of Heart Disease Using Machine Learning IEEE paper by Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Kailas Devadkar p. 18132580, 2018.

[5]. Prediction of Cardiovascular Disease Using Machine Learning Algorithms by Kumar G Dinesh, K Arumugaraj, Kumar D Santhosh, V Mareeswari, 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT).