Capstone Project

Cardiovascular Risk Prediction

Presented by:

Chetan Chavan
Samata Parulekar

Contents:

- Problem Statement.
- Variables in Dataset.
- Process flow of Project.
- Visualization Graphs.
- Correlation Heat map.
- Different algorithms to check accuracy
- Conclusion.

Problem Statement:

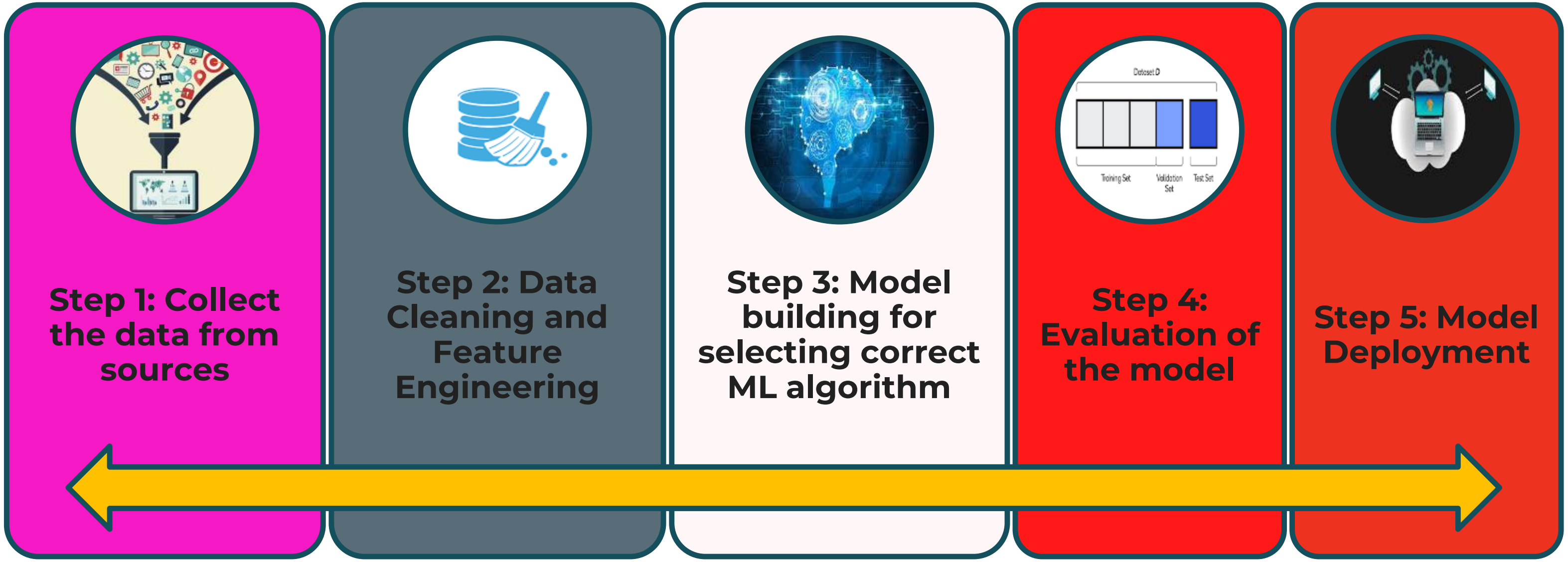
- The dataset acquired contain various information related to human body health as well as body condition. The information incorporates the data containing various classification based values as explain in the dataset. **The main purpose of this project is to understand the dataset, visualize and prepare a classification based model to predict whether a person is prone to 10 year coronary heart disease or not.**

Variables in Dataset

- **Sex :** male / female ("M" or "F")
- **Age :** Age of the person
- **Education :** Education level of the person
- **is_smoking :** whether the person is smoking or not (YES or No)
- **CigsPerDay:** No. of cigarettes person used to have per day.
- **BPMeds:** whether the person taking medicine for Blood Pressure or not
- **prevalentstroke:** whether the person had heart stroke previously or not
- **prevalenthyp:** whether the person having prevalent hypertension issue or not
- **Diabetes:** whether patient having diabetic or not
- **TotChol:** total cholesterol level in the body of person
- **sysBP:** systolic blood pressure of person
- **diaBP:** diastolic blood pressure of person

- **BMI:** Body Mass Index of persons
- **HeartRate:** heart rate of the person
- **Glucose :** glucose level of the person
- **tenyearchd :** 10-year risk of coronary heart disease CHD

Process flow of project



Step 1: Collect the data from sources

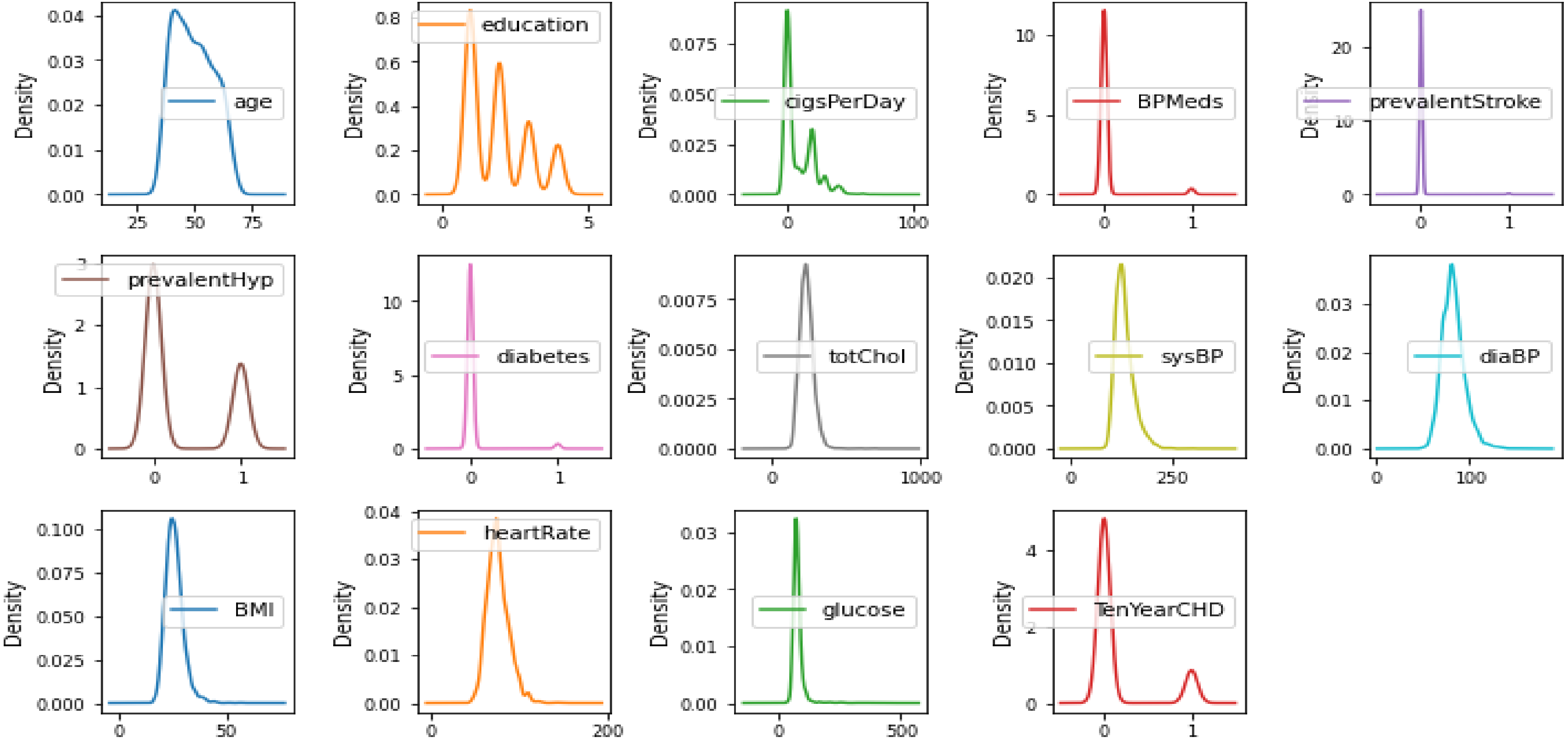
Step 2: Data Cleaning and Feature Engineering

Step 3: Model building for selecting correct ML algorithm

Step 4: Evaluation of the model

Step 5: Model Deployment

Plotting graphs to check the importance of various factors



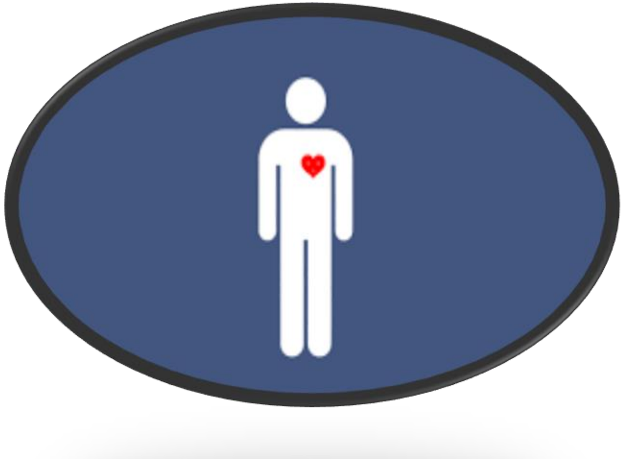
1) From the plots it is clearly visible that age group from 35 to 60 has been considered for heart disease studies

2) People having 2 to 10 cigarettes are less likely to occurs

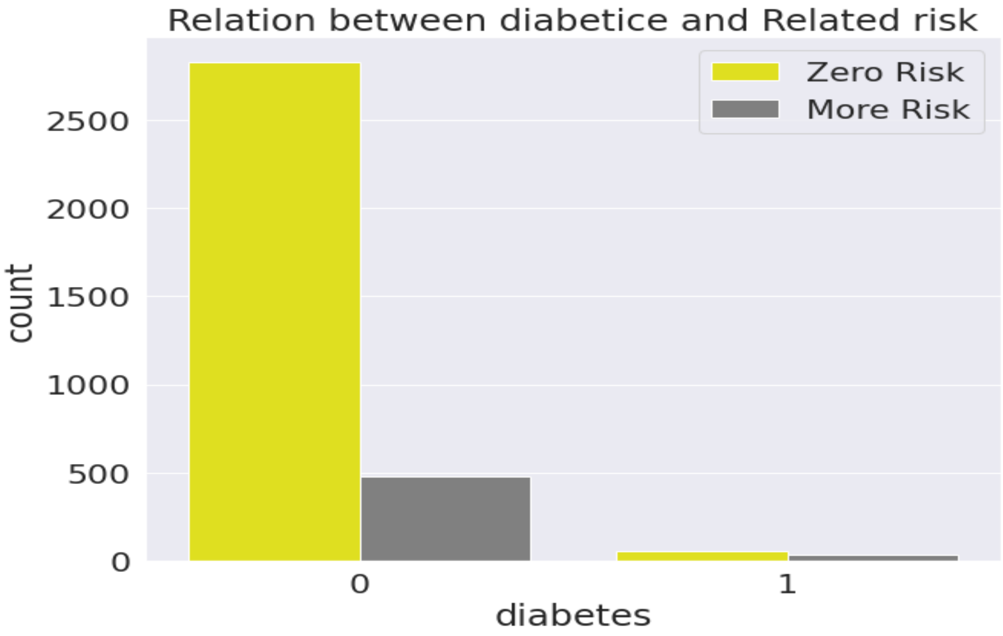
3) Not much people suffered from prevalent stroke i.e. previously occurred stroke

4) Not much people suffered from prevalent hypertension

5) Remaining observations are telling about health conditions such as diabetes, total cholesterol, BMI, etc.

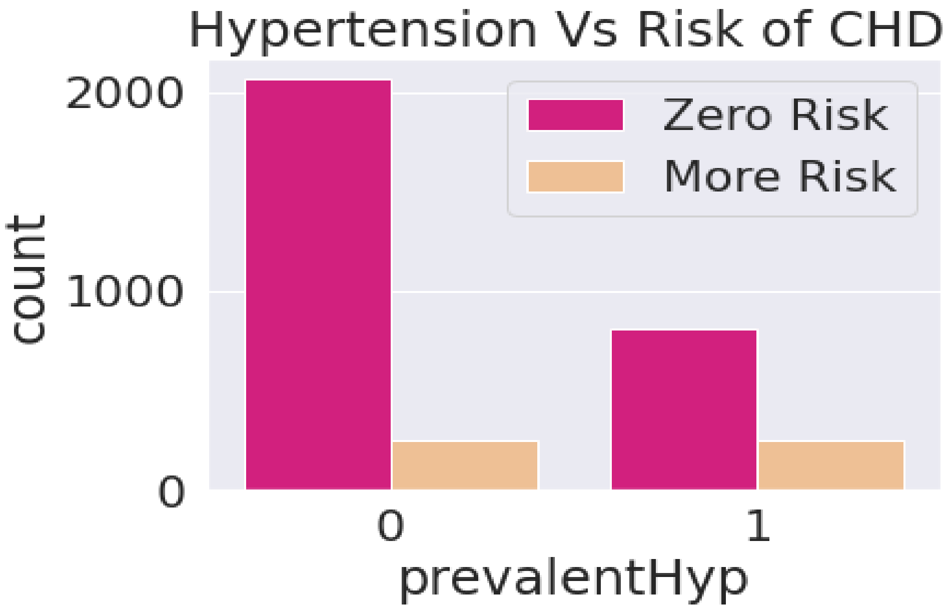


a) Relation between Diabetes and risk factor



From above visualization, it is clearly seen that males having diabetics are more prone to CHD

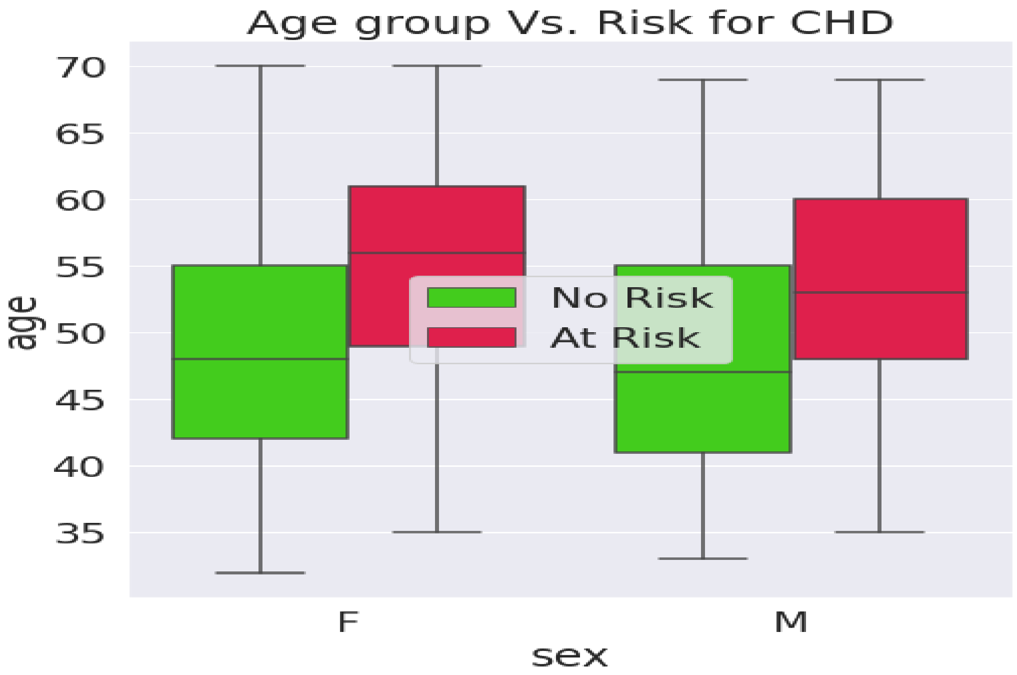
b) Prevalent stroke and Risk of CHD



From bar graph, most of the men with hypertension are not at risk but yes some are at serious risk.

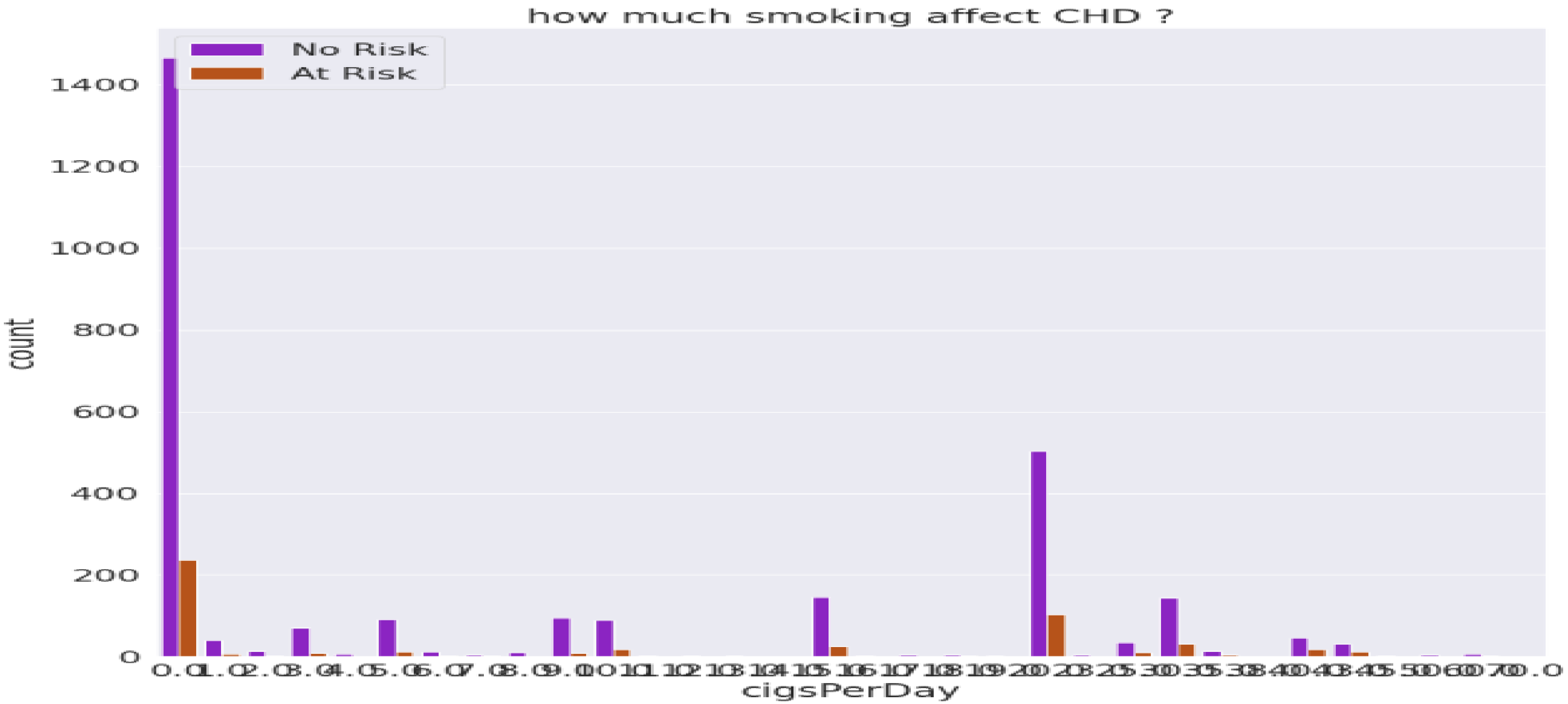
All in all, numbers of men at risk are way less than those who are at no risk.

c) Age group and risk of CHD relation



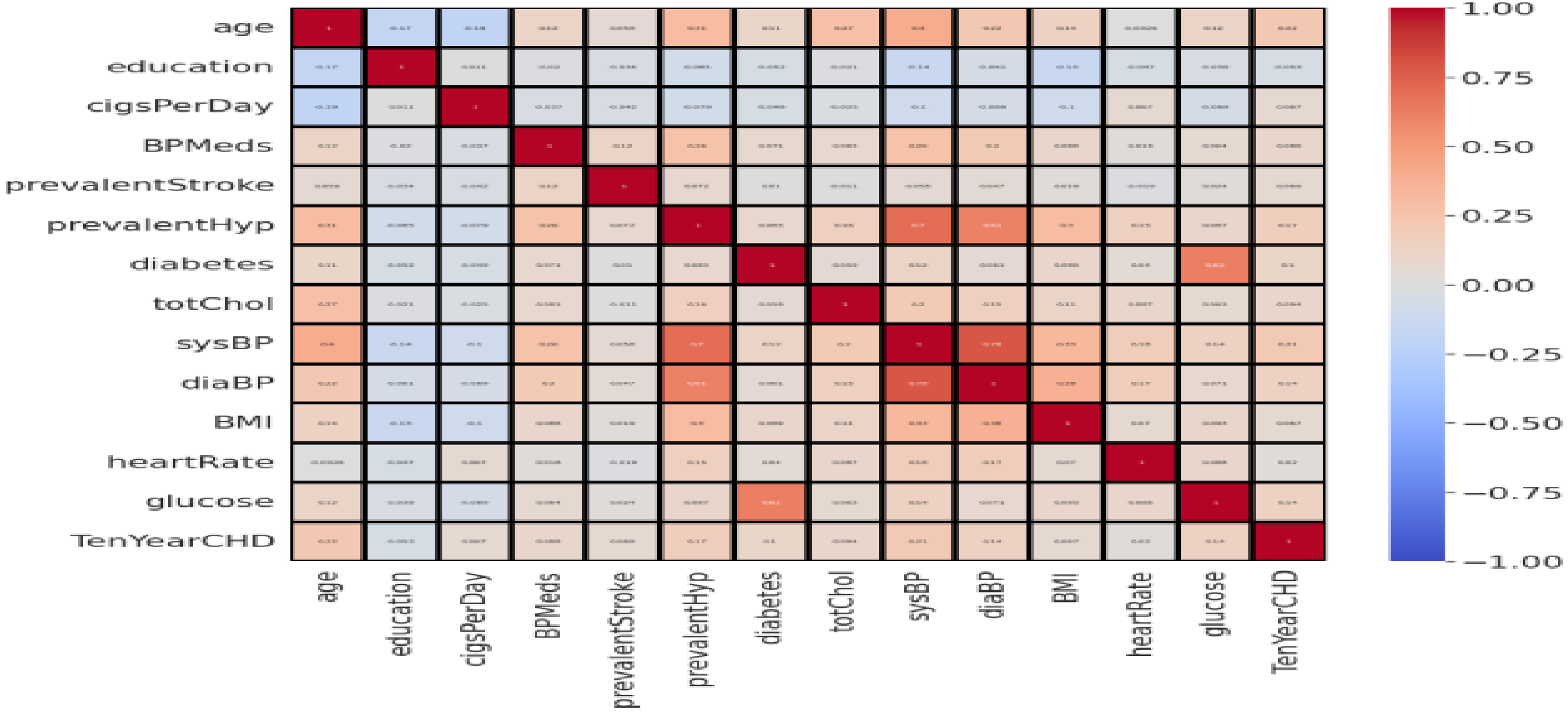
The age group between 47 to 60 for males and 48 to 62 for females are more prone to CHD

d) Number of Cigarettes per day Vs. CHD risk



For males as well as females CHD does not much related to Cigarettes' per day but still those having more than 2 cigarettes per day are more prone to CHD

e) Correlation Matrix



From correlation matrix it is clearly visible that not many factors are correlated to each other but some independent variables are related to each other glucose level and diabetes.

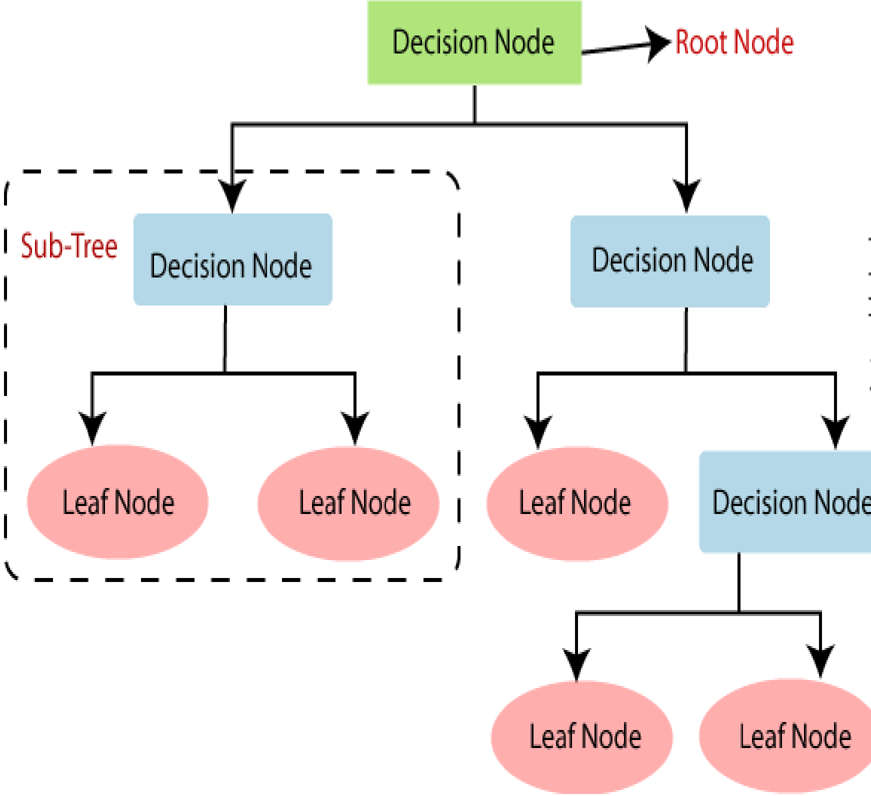
Model Selection an developing the algorithm

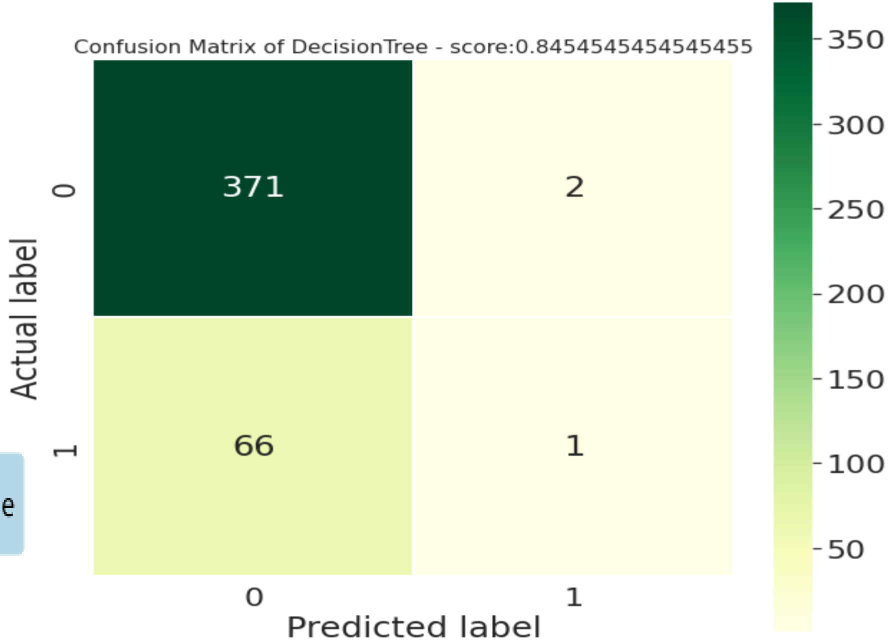
a) Decision Tree Classifier:

- Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

- The decisions or the test are performed on the basis of features of the given dataset. The confusion matrix obtain from the Decision Tree algorithm is shown in next slide





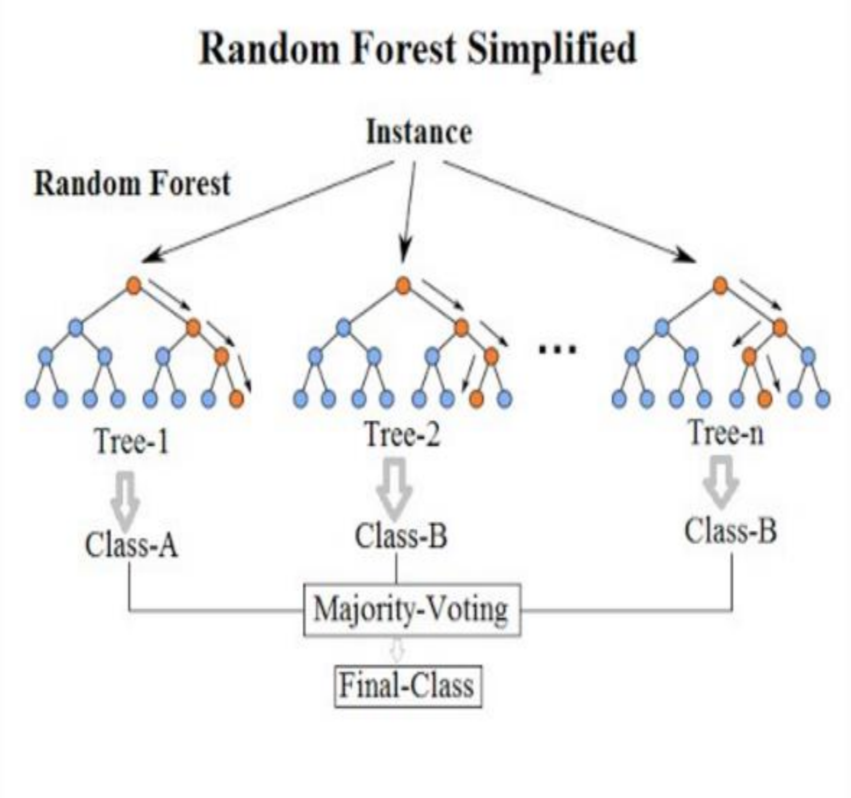
From decision tree classifier, we are able to achieve the accuracy of about 73.18%.

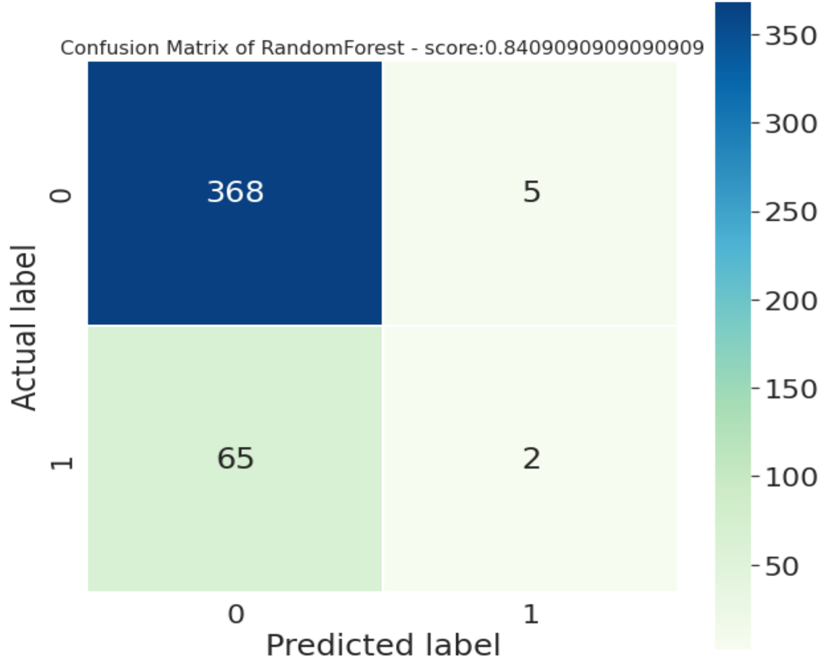
b) Random Forest Classifier

- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

- As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. The confusion matrix obtained from this algorithm is shown in next slide.




Confusion Matrix of RandomForest - score:0.8409090909090909

Actual label \ Predicted label	0	1
0	368	5
1	65	2

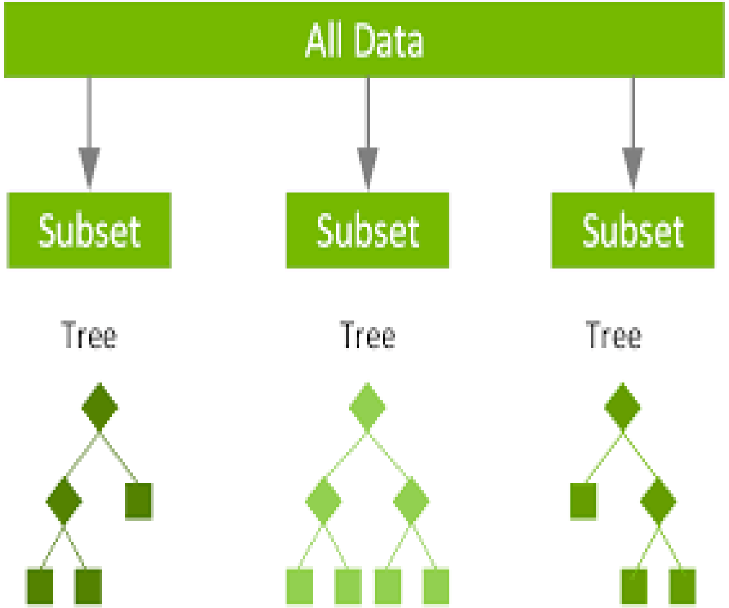
From random forest classifier, we are able to achieve the accuracy of about 84.09%.

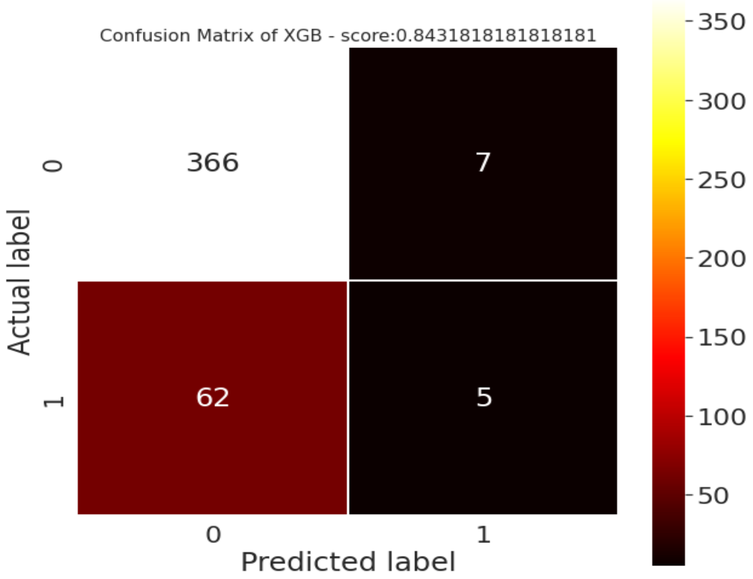
c) XGBoost Classifier

- A loss function should be improved, which implies bringing down the loss function better than the result.

- To make expectations, weak learners are used in the model Decision trees are utilized in this, and they are utilized in a jealous way, which alludes to picking the best-divided focuses in light of Gini Impurity and so forth or to limit the loss function.

- The additive model is utilized to gather every one of the frail models, limiting the loss function. Trees are added each, ensuring existing trees are not changed in the decision tree. Regularly angle plummet process is utilized to find the best hyper boundaries, post which loads are refreshed further. The confusion matrix obtained for this model is shown in next slide.





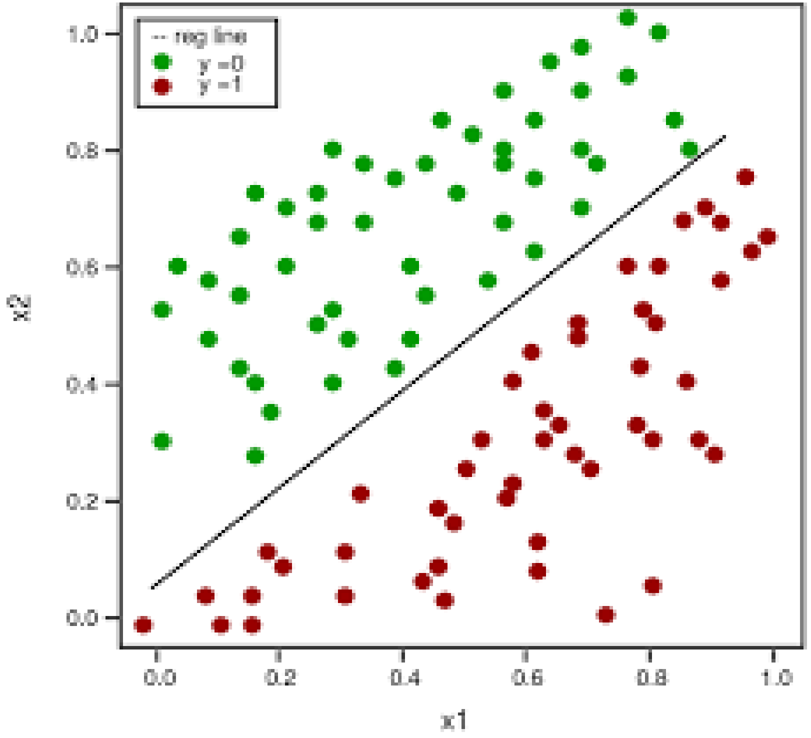
From xg boost algorithm, we are able to achieve the accuracy of about 84.31%.

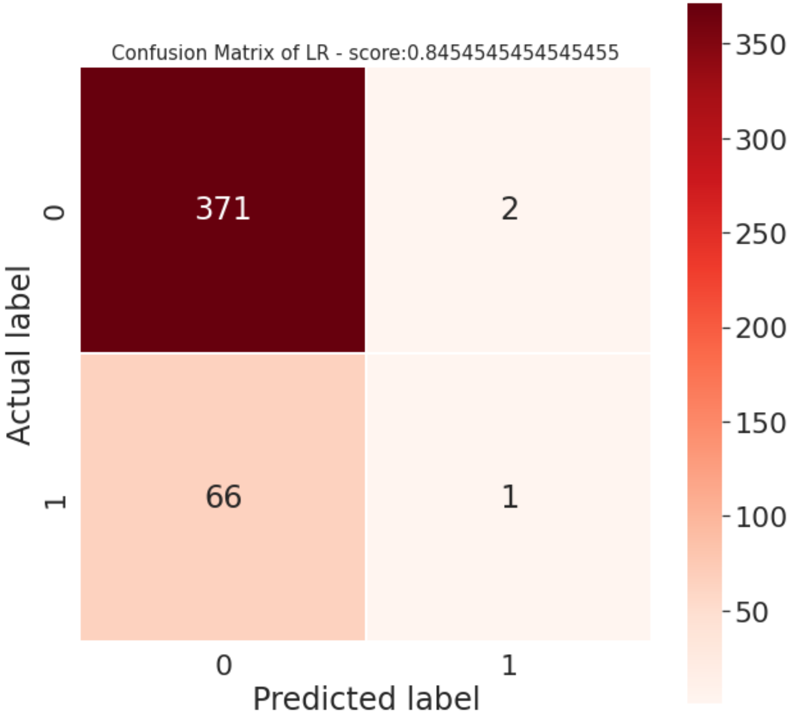
d) Logistic Regression

- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. The confusion matrix obtained is given as follows:

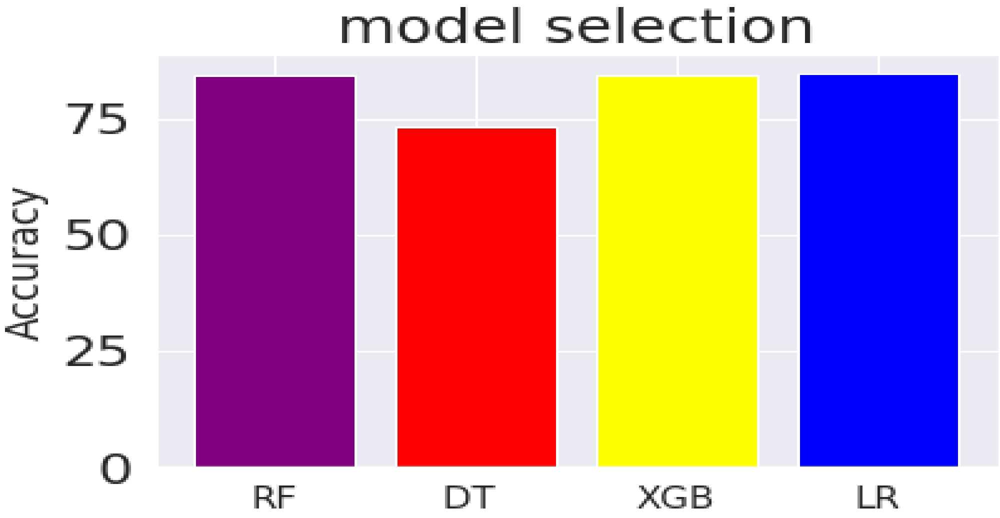




From this algorithm, we are able to achieve the accuracy of about 84.54%.

f) Comparison Plot

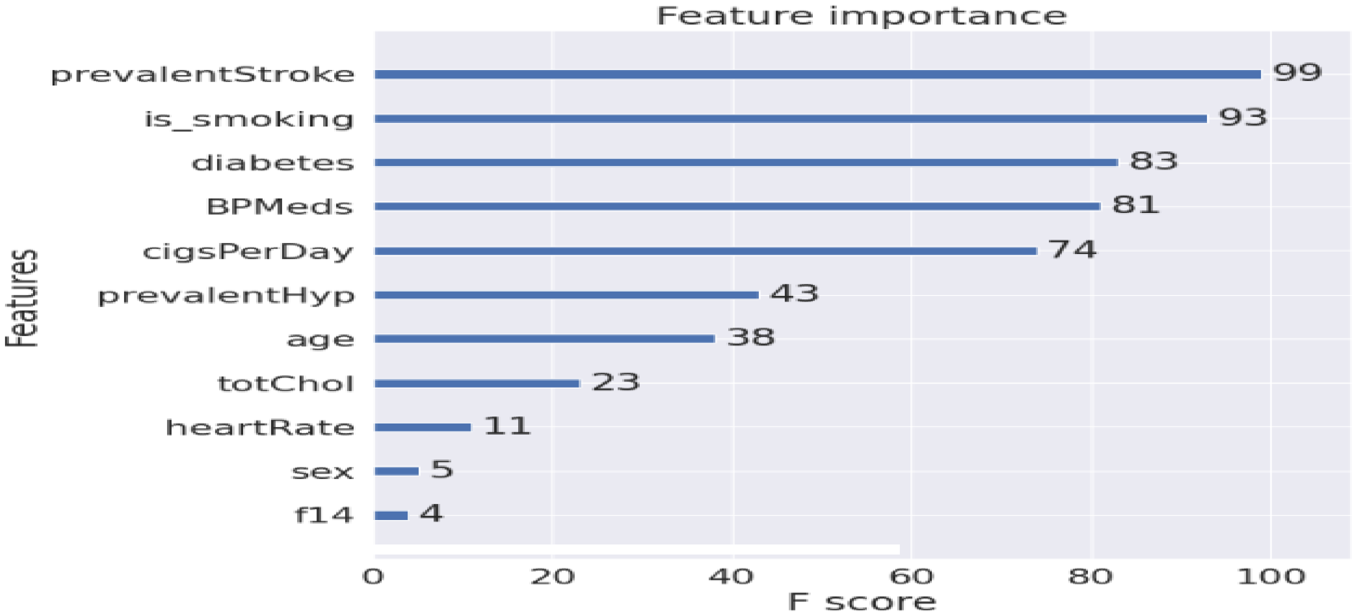
Graphical representation for comparison of all algorithms is given as follows:



Based on graph Logistic regression gives best accuracy.

g) Important features:

After preforming various operations we can conclude that most serious factors for 10 year CHD are prevalent stroke, smoking and Diabetes. Whereas type of gender and heart rate are the least bothered factors as shown in following fig.



Conclusion:

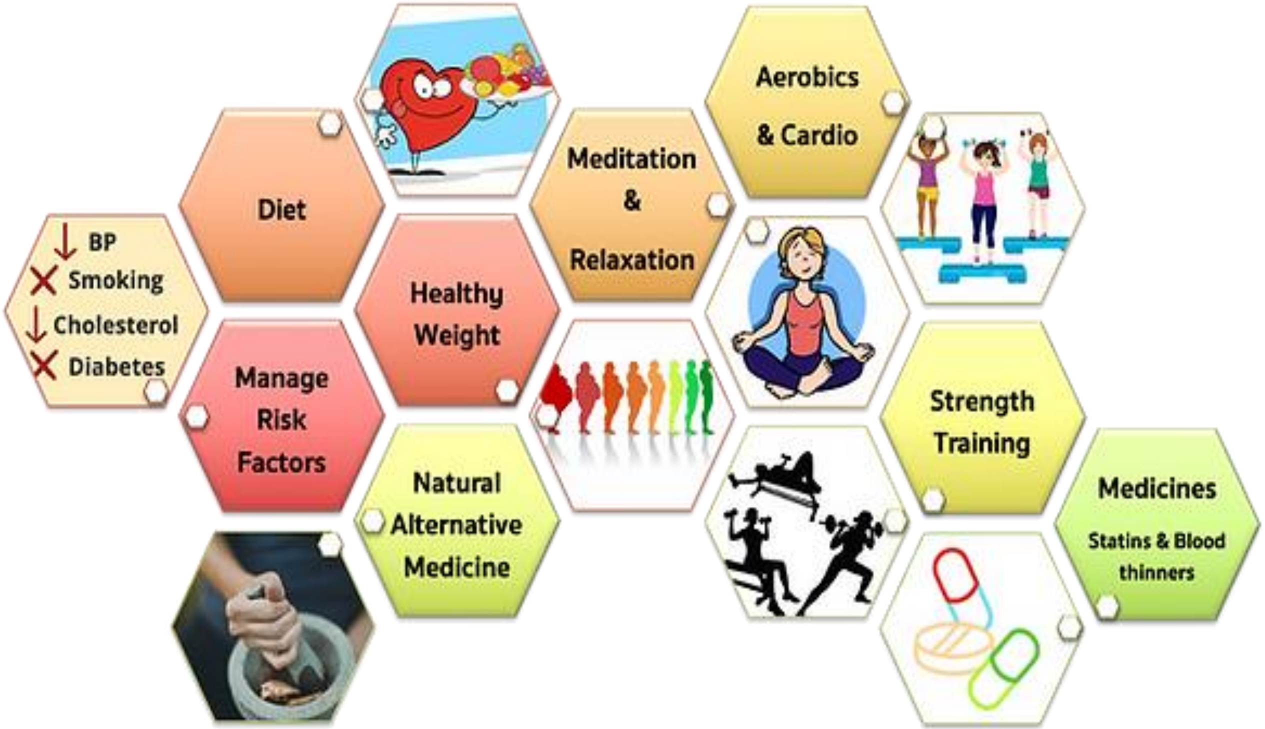
- A cardiovascular risk prediction model is being prepared with the help of various classification techniques such as Logistic regression, Random Forest technique, XGBoost, Decision tree.

- The designed model can predict the 10 year risk of Coronary Heart Disease (CHD) for the individual. It is completely based on the previous and present health as well as medical condition.

- Amongst all XGBoost classifier and Logistic regression gives best accuracy score of **84.31** and **84.54%** respectively. Whereas, Random Forest classifier gives the accuracy of around **84.09%**.

- Decision Tree proven to be less effective model for the project as it gives the efficiency as **73.18%**. The performance results are varying from **70%** to **84 %** and the reason could be no proper pattern of data, not relevant data or not enough data.

Thank You.....


↓ BP
✗ Smoking
↓ Cholesterol
✗ Diabetes
Diet
Manage Risk Factors
Healthy Weight
Natural Alternative Medicine
Meditation & Relaxation
Aerobics & Cardio
Strength Training
Medicines
Statins & Blood thinners