# Cardiovascular Risk Prediction

**Pavithra K**

**Data science trainee**

**AlmaBetter, Bangalore**

## Abstract:

Life depends on the functioning of the heart; thus, the heart is involved in all death, but this does not account for its prominence in causing death. To some degree, as medical science advances, more people are saved from other illnesses only to die from one of the unsolved and uncontrolled disorders of the cardiovascular system. However, changes in lifestyle and diet, including the adoption of more sedentary lifestyles and the consumption of fried foods and foods high in sugar, have resulted in increases in the incidence of otherwise preventable cardiovascular-related illness and death.

Exploratory data analysis on cardiovascular study on residents of the town of Framingham, Massachusetts helped to understand the major demographic, behavioral, and medical risk factors like age, hypertention, smoking(cigsperday), gender etc., We used different algorithms like logistic regression, decision tree, random forest, XGBoost and Support vector machine to predict whether the patient has a 10-year risk of future coronary heart disease. Used recall as evaluation metric to analyze the model prediction.

## Introduction:

Coronary heart disease is a type of heart disease where the arteries of the heart cannot deliver enough oxygen-rich blood to the heart. It is the leading cause of death in the United Kingdom. Three main risk factors have been identified: cigarette smoking, a high level of cholesterol in the blood (hypercholesterolemia), and high blood pressure (hypertension). Important as these risk factors are, they are found only in about one-half of those who experience heart attacks. The proportion of persons with any or all these three risk factors is greater in young and middle-aged adults than in older adults. It is impossible to incriminate any one of these risk factors over another, since the manifestations of coronary heart disease are undoubtedly due to many independent and interdependent influences, but the coexistence of the three greatly increases the risk of developing the disease.

According to the World Health Organization, as many as 80% of all heart attacks and strokes are preventable. The majority of deaths due to CVD are precipitated by risk factors such as high blood pressure, high cholesterol, obesity, or diabetes, which can, to a large extent, be prevented or controlled through the consumption of a healthy diet, regular exercise and avoiding tobacco. Keeping an eye on your blood pressure, cholesterol levels and blood sugar levels is also very important.

Our aim to build a machine learning algorithm that predict the goal is to predict whether the patient has a 10-year risk of future coronary heart disease. To minimize the risk of having it by maintaining healthy life style

## Problem statement:

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

## Data description:

The dataset provides the patients' information. It includes over 3,390 records and 17 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors

## Methodology followed:

1. **Importing dataset:** Here we use pandas libraries to import data.
2. **Exploratory Data analysis: Exploratory Data Analysis (EDA)** is a critical component involved while working with data. EDA is used to comprehensively understand the data and discover all its characteristics, summarizing data, identifying patterns and relationships, and detecting outliers, typically by employing visual techniques. This is an important step in data analysis that can be used on both qualitative and quantitative data. This makes it possible for you to understand your data more thoroughly and find interesting patterns in it**.**

- **Understanding about data source**: To utilize the data in well manner the knowledge of data source and the data collection process is important.to what purpose data has been collected for research purpose or for randomly collected**.**
- **Understanding about attributes of data:** This includes understanding the dataset's data type, what each column represents, and any other relevant information. This understanding is critical for properly performing an EDA because it will help you know what to look for and how to analyze the data.
- **Data cleaning:** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. It includes following steps**:**

- ✓ **Dealing with missing values:** Sometimes we may find some data are missing in the dataset. We know that we can use the isnull().sum() and notnull().sum() functions from the pandas library to determine count of null values present. Based on amount of data present either we can drop the missing rows, columns or replace missing values with calculated mean, mode or median of respective**.**

- ✓ **Dealing with duplicate rows:** Sometimes we may find some data are missing in the dataset. We know that we can use the isnull().sum() and notnull().sum() functions from the pandas library to determine count of null values present. Based on amount of data present either we can drop the missing rows, columns or replace missing values with calculated mean, mode or median of respective**.**

- ✓ **Dealing with incorrect format:** It is important to check our data is in correct formats (int, float, text, or other). We can use following pandas function, df.dtype() to check data types and .astype() to change the data type df.dtypes- will return data types in our data frame**.**

- ✓ **Dealing with Anomalies/Outliers:** Outliers are unusual values in your dataset, handling them is important because they can distort statistical analyses and violate their assumptions. Outlier are very informative it is important to know reason of cause of outliers it may cause due to Data entry errors/ Measurement errors/ instrument errors/Sampling errors/Data processing error/Natural novelties in data

- **Data Analysis:**

  - ✓ **Univariate analysis:** The univariate analysis of data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it. It is done using Histogram, Pie chart, Bar chart, Boxplot.

  - ✓ **Bivariant analysis:** As its name indicates this type of analysis includes two variables for analysis. Through this we can find the how two variables are related to each other

  - ✓ **Multivariant analysis:** Multivariate analysis is a more complex form of statistical analysis technique and used when there are more than two variables in the data set. Using pairplot, heatmap on correlation matrix

**5. Encoding of categorical column:** Feature encoding of categorical variables binary because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

**6. Binning:** one of the popular techniques of feature engineering, "binning", can be used to normalize the noisy data. This process involves segmenting different features into bins.

7. **Transformation:** Transform helps in handling the skewed data, and it makes the distribution more approximate to normal after transformation. It also reduces the effects of outliers on the data, as because of the normalization of magnitude differences, a model becomes much robust.

### 8.Splitting the dataset into training and testing dataset.

### 7._Building Classification models:

For modelling we tried various classification models like:

1.Logistic Regression

2.Decision Tree

3.Random Forest

4.XGBoost Classifier

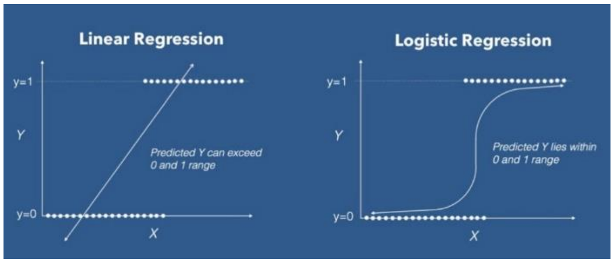
5. Support vector machine

## 1.Logistic Regression:

Logistic regression is supervised and parametric machine learning classification algorithm. This is interpretable algorithm. It is a statistical method used for mainly for binary classification by measuring the relationship between categorical dependent variable and independent variable by using log of odds. However, it can be used for multiclass classification by one v/s other technique. It is useful for linearly separable data.

Since linear regression has limitation on outlier and not able to predict value range only between 0-1. To overcome these drawbacks a sigmoid function is introduced which give output in range 0-1 and works well with outlier when dimension of data is low.

$$y = \frac{1}{1+e^{-(w_o + w_1 x)}}$$


Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed 0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within 0 and 1 range

A ***decision boundary*** is a threshold that we use to categorize the probabilities of logistic regression into discrete classes. A decision boundary could take the form (default value is 0.5):

y = 0 if predicted probability < 0.5

y = 1 if predicted probability > 0.5

**Assumptions of logistic regression:**

- **target variable should be binary**
- **no multicollinearity between independent variable**
- **there should not be extreme outlier**
- **There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable**
- **The Sample Size is Sufficiently Large**

The cost function checks what the average error is between actual class membership and predicted class membership in logistic regression we use log loss or also called as binary cross entropy (maximum likelihood function) as cost function.

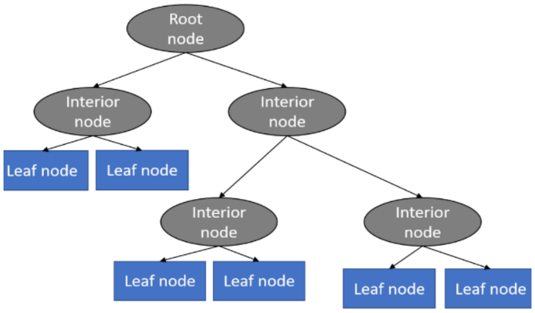
$$J(\theta) =- \frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}log(h_\theta(x^{(i)})) + (1-y^{(i)})log(1-h_\theta(x^{(i)}))\right]$$

We use gradient descent as optimization function to update the model parameter.

## 2.Decision Tree:

Decision tree is supervised and non-parametric machine learning classification algorithm. This is interpretable algorithm. This algorithm used for regression as well as for classification algorithm.

” Learning simple decision rules in the form of tree inferred from the data features”. It has tree like structure consisting of nodes and branches. Where internal node represents the features of dataset, branches represent decision rules and leaf node represents outcome. The leaf node may have continuous (Regression) as well as categorical values(classification).



Major points to be known earlier:

- **Root Node:** Root node is from where the decision tree starts.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

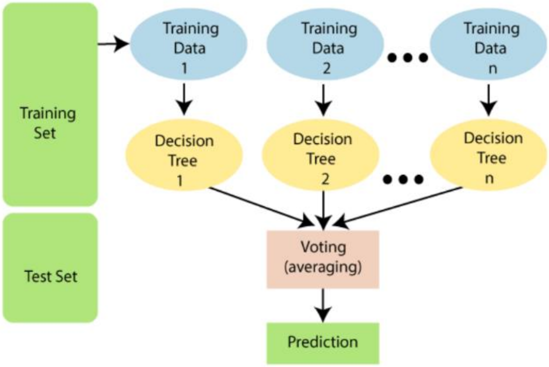
A decision tree before starting usually considers the entire data as a root. Then on condition, it starts splitting by means of branches or internal nodes and decides until it produces the outcome as a leaf. Only one important thing to know is it reduces impurity present in the attributes and simultaneously gains information to achieve the proper outcomes while building a tree.

To select proper attribute at each node we use *algorithm* like information gain**(using entropy** is a measure of impurity present in the data**)**, gini index**,** etc. These algorithms will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e, the attribute with a high value (in case of information gain) is placed at the root.

Decision trees have an advantage that it is easy to understand, lesser data cleaning is required, non-linearity does not affect the model's performance and the number of hyper-parameters to be tuned is almost null.

## 3.Random Forest

Supervised machine learning algorithm used for regression as well as for classification technique. But, mainly used for classification technique It is based on the concept of **ensemble learning**. Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly (row replacement method) selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the predicted output of all the model. **The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**

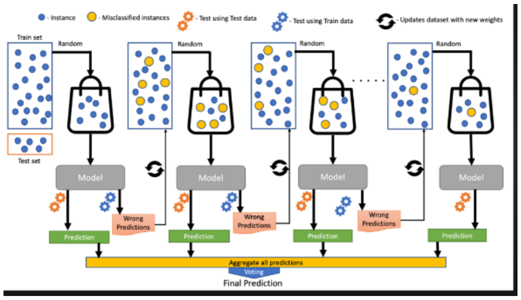


- It works well even if it contains missing or null values in it.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

## 4. XGBoost

The XGBoost algorithm is supervised machine learning algorithm used for regression as well as for classification (binary and multiclass), and ranking problems technique. The XGBoost algorithm performs well in machine learning competitions because of its robust handling of a variety of data types, relationships, distributions, and the variety of hyperparameters that you can fine-tune. Before moving ahead we should know about boosting.

Boosting is an ensemble learning technique to build a strong classifier model from several weak classifier in series. Boosting algorithms play a crucial role in dealing with bias-variance trade-off. Unlike bagging algorithms, which only controls for high variance in a model, boosting controls both the aspects (bias & variance) and is more effective.



- 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 considering 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.

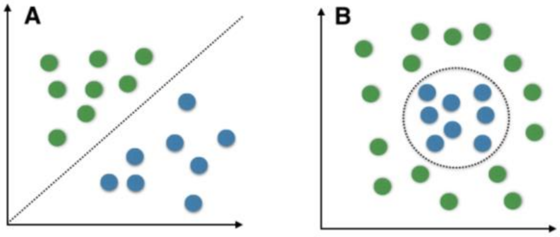
Trees are added each, ensuring existing trees are not changed in the decision tree. Regularly angle plummet process is utilized to find the best hyper boundaries, post which loads are refreshed further.

## 5. Support vector machine:

Support vector machine is supervised machine learning algorithm used for classification and/or regression technique. Most of the time used for classification but, sometime it is most useful in regression technique. It works well with linearly and non-linearly separable data. It Is mainly used for binary classification. If The data is multiclass problem it uses one verses other technique to classify the classes.

The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three. It uses Hinge loss as cost function

One reasonable choice as the best hyperplane is the one that represents the largest separation or margin between the two classes. So, we choose the hyperplane whose distance from it to the nearest data point on each side is maximized. If such a hyperplane exists it is known as the maximum-margin hyperplane/hard margin. There is another type of margin i.e, soft margin used to overcome outlier problem.



**Types of SVMs:**

**Simple SVM:** Typically used when data is linearly separable.

**Kernel SVM:** It is used when data is non linearly separable. It uses kernel tricks to convert lower dimensional data to higher dimensional data where it easy to separate two classes linearly. It is more flexibility for non-linear data because you can add more features to fit a hyperplane instead of a two-dimensional space. Different kernels are Polynomial kernel, RBF, Sigmoid, etc.,

**Evaluation metrics:**

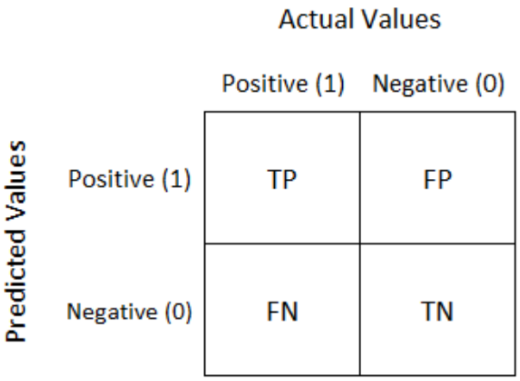
### Accuracy

Accuracy simply measures how often the classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.

$$\textbf{\textit{Accuracy}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

### Confusion Matrix:

A confusion matrix is defined as the table that is often used to describe the performance of a classification model on a set of the test data for which the true values are known.



| | | Actual Values | |
|---|---|---|---|
| | | Positive (1) | Negative (0) |
| Predicted Values | Positive (1) | TP | FP |
| | Negative (0) | FN | TN |

It is extremely useful for measuring the Recall, Precision, Accuracy, and AUC-ROC curves.

**True Positive: Correctly predicted as Positive class**

**True Negative: Correctly predicted as Negative class**

**False Positive: observation is negative class but, predicted as positive**

**False Negative: observation is Positive class but, predicted as negative.**

### Recall:

Recall explains how many of the actual positive cases we were able to predict correctly with our model. It is a useful metric in cases where False Negative is of higher concern than False Positive. It is important in medical cases where it doesn't matter whether we raise a false alarm but the actual positive cases should not go undetected!

Recall for a label is defined as the number of true positives divided by the total number of actual positives.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

### Precision:

Precision explains how many of the correctly predicted cases actually turned out to be positive. Precision is useful in the cases where False Positive is a higher concern than False Negatives. Precision is number of true positives divided by the number of predicted positives.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

👉 **F1_Score**: It gives a combined idea about Precision and Recall metrics. It is maximum when Precision is equal to Recall.

**F1 Score is the harmonic mean of precision and recall.**

$$F1 = 2.\,\frac{Precision\ \times Recall}{Precision + Recall}$$

F1 Score could be an effective evaluation metric in the following cases:

- When FP and FN are equally costly.
- Adding more data does not effectively change the outcome
- True Negative is high

### 👉 AUC-ROC curves:

The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR (True Positive Rate) against the FPR(False Positive Rate) at various threshold values and separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes.

Greater the AUC, the better is the performance of the model at different threshold points between positive and negative classes. This simply means that When AUC is equal to 1, the classifier able to perfectly distinguish between all Positive and Negative class points. When AUC is equal to 0, the classifier would be predicting all Negatives as Positives and vice versa

### Hyperparameters tunning for better Result:

Hyperparameters tunning for logistic regression decision tree, random forest, XGBoost, and support vector machine algorithms is necessary for using best parameter to obtain the better result.

### Cardio Vascular Risk Prediction:

Here we are discussing insight obtained from exploratory data analysis on cardio vascular dataset. The dataset is from an ongoing cardiovascular study on residents of Framingham Massachusetts.The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD)

- **Importing the dataset: Cardio vascular risk prediction dataset in .csv (**comma-separated values file) form is imported using pandas' library.

- **Identifying the Data Source and Data Collection:** Cardio vascular risk prediction dataset by researcher. Early Prediction of risk of having cardio vascular disease will Maximize the awareness of maintaining healthy habit in patient which result in minimizing the possibility of having heart problem can be reduced.

- **Dataset Information:** The dataset provides the patients' information. It includes over 3,390 records and 17 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors

☞ **Id-** unique id for observation

☞ **Education –** education level of patient indicated in numbers 1,2,3, and 4

☞ **Demographic:**

- **Sex**: male or female ("M" or "F")
- **Age:** Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

☞ **Behavioral**

- **is_smoking:** whether the patient is a current smoker ("YES" or "NO")
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day. (can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

☞ **Medical ( history)**

- BP Meds: whether the patient was on blood pressure medication (Nominal)
- **Prevalent Stroke**: whether or not the patient had previously had a stroke (Nominal)
- **Prevalent Hyp:** whether or not the patient was hypertensive (Nominal)
- **Diabetes**: whether or not the patient had diabetes (Nominal)

☞ **Medical(current):**

- **Tot Chol**: total cholesterol level (Continuous)
- **Sys BP:** systolic blood pressure (Continuous)
- **Dia BP**: diastolic blood pressure (Continuous)
- **BMI:** Body Mass Index (Continuous)
- **Heart Rate**: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
- **Glucose**: glucose level (Continuous)

☞ **Predict variable (desired target)**

- 10-year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”) - DV

- **Data Cleaning cardio vascular risk prediction dataset:**

❖ **Dealing with missing values: The dataset includes missing data as percentagewise is listed as follows:**

| | | |
|---|---|---|
| 1) | Glucoe | 8.97 |
| 2) | education | 2.57 |
| 3) | BPMeds | 1.30 |
| 4) | totChol | 1.12 |
| 5) | cigsPerDay | 0.65 |
| 6) | BMI | 0.41 |
| 7) | heartRate | 0.03 |

we used k- nearest neighbor algorithm to impute the missing value for categorical and numerical features separately.

❖ **Dealing with duplicate rows:** Our **dataset** does not contain any duplicate rows. So, no need to deal with duplicate values.

❖ **Dealing with incorrect format:** we renamed target feature name.

❖ **Dealing withAnomalies/Outliers:** On conducting univariant analysis using boxplot on continuous data we come to know that few feature suffer from skewness that is due to extreme outlier present in the data. We used capping method to deal with outlier i.e, by using median value of each feature. Here we listed skweness after capping

1) age....... 0.226
2) cigsPerDay .......1.116
3) totChol ........ 0.238
4) sysBP ...... 0.59
5) diaBP......... 0.3
6) BMI ....... 0.294
7) heartRate........ 0.288
8) glucose ........ 0.365

❖ **Data Analysis (Univariant and bivariant analysis on cardio vascular dataset).**

☞From analysis of target variable " TenYearCHD" we observed that data set is biased we need to carefully deal with biased dataset.

☞ in comparison with male female less likely to get cardiovascular risk. Also, we observed that male who smoke has higher risk of having cardiovascular risk than female who smoke. As no of cigsperday increases the cardiovascular risk also increase in males

☞ As people get older the risk of getting cardiovascular disease is high male in comparison with female. **Binning is used to** Convert age feature into categorical. Only for reference.

☞ we observed that the people who completed their education level 1 is higher in count, than people who completed level 2(intermediate) or level 3(graduates) or level 4(post graduate. Education level does not affect in having cardio vascular risk but it makes to increase awareness of taking care of health.

☞People with BP medication, patient had previously had a stroke, and prevalentHyp have chance of having cardio vascular risk is 50%,80%,and 40% respectively.

☞People who have diabetes have 60-70% chance of having cardio vascular disease risk.

☞ People with total cholesterol range 220-270 have higher risk of having cardiovascular risk

☞ People with bmi between 23-28 has higher risk of having heart problem.

❖ **Based on correlation metrics:** Here we have observed that glucose, heartrate are not significantly correlated with cardio risk.

cigsPerDay -- is_smoking and "diaBP" and "sysBP" these two pair of features are highly correlated with each other. We used Varience influence Factor to drop the highly correlated feature

- **Converting categorical columns into numerical columns**. categorical features is_smoking and sex is converted into numerical feature.
- Feature selection done using vif for logistic regression. Also, Id and education features are dropped
- **Splitting the dataset into training and testing dataset:** In train test split we take x as dependent variables and y take as independent variable then train the mode
- **Smote technique is used to deal with biased data**

**Model building using Logistic regression: Checking for assumptions of logistic regression:**

- Target variable is binary class is satisfied
- Zero or minimal multicollinearity we removed highly correlated independent variables using vif
- Outliers are treated with capping using median value.

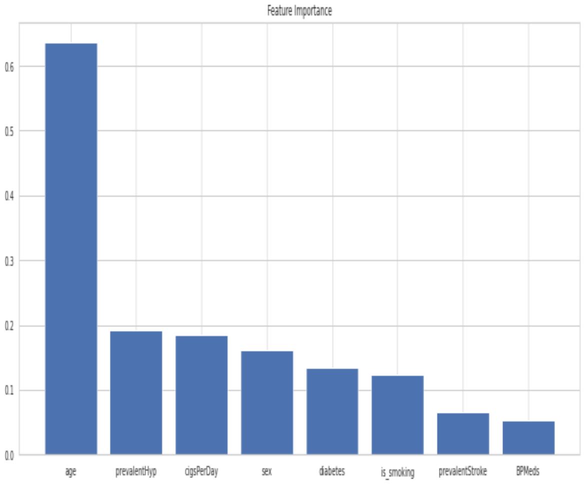
We have used recall as evaluation metrics because in medical field True positive rate need to be high while dealing with risk of having heart disease prediction. Even though person who does not have risk of having heart disease is predicted as has risk to having disease doctors, or patient will become more conscious about his health and take precaution to maintain the health. For that reason, false positive does not impact badly. But a person who does have risk of having heart disease, has been predicted as a no risk of heaving heart disease will have adverse effect i.e, he may get careless about maintaining healthy lifestyle and end is life by heaving heart disease. So, recall should be high in this case.

For logistic regression obtained recall value

Using logistic regression, we get an recall score of train set **68.01**%

Using logistic regression, we get an recall score of test set **64.44**%

Feature importance obtained from logistic regression is given as follows:



## 2.Decision Tree:

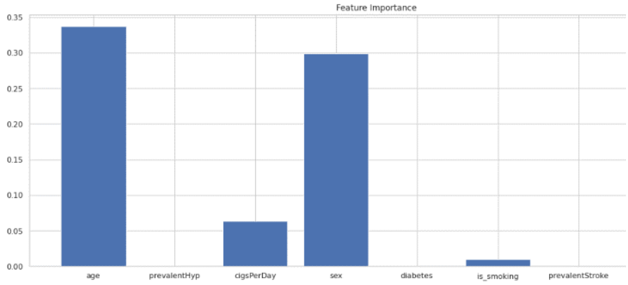
All the features included except id and education feature.

Recall score is as follows:

Train set: **78.39%**

Test set**: 63.72**%

Feature importance obtained using decision tree as follows:



## 3.**Random Forest:**

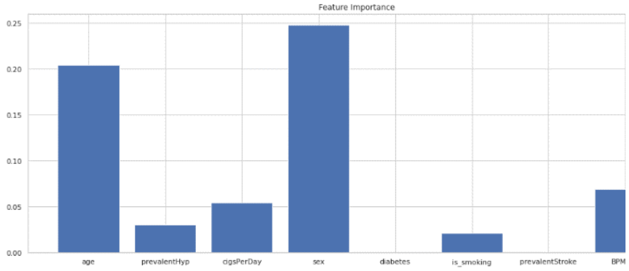
All the features included except id and education feature.

Recall score is as follows:

Train set: **72.69%**

Test set**: 56.64%**

Feature importance obtained using Random Forest as follows:



## 4.**XGBoost:**

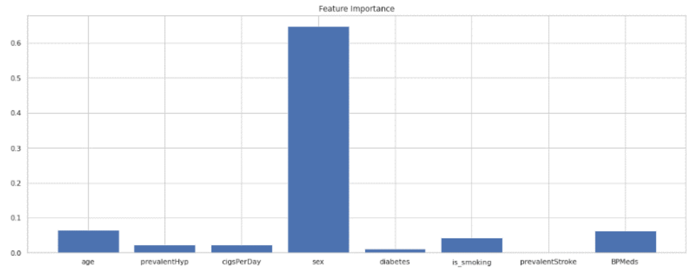
All the features included except id and education feature.

Recall score is as follows:

Train set: 94.47**%**

Test set:36.28%

Feature importance obtained using XGBoost as follows:



## 5.Support Vector Machine:

All the features included except id and education feature.

Recall score is as follows:

Train set: **94.47%**

Test set:**94.69%**

Final comparison table:

| models | Recall train set in % | Recall test set in % |
|---|---|---|
| Logistic regression | **68.01** | **64.44** |
| Decision tree | **78.39** | **63.72** |
| Random Forest | **72.69** | **56.64** |
| XGBoost | **94.47** | **36.28** |
| Support vector machine | **94.47** | **94.69** |

## Conclusion:

Coronary heart disease is the leading cause of death worldwide, although its occurrence is unevenly distributed. It is one of the most common causes of death in North America and Europe. Through our analysis we come know that various aspects may lead to have coronary heart disease age, gender, cigarette smoking, a high level of cholesterol in the blood (hypercholesterolemia), and high blood pressure (hypertension) and obesity.

Crucial part of work is to building a machine learning algorithm, we tried with many algorithms even though data is imbalanced, we used smote technique to balance the class of target variable. Build the best model with support vector machine having recall rate 94.47 for training and 94.69 for test set.

By early prediction of coronary heart disease will create an awareness in people to maintain the healthy life style. In order to reduce the risk of getting coronary heart disease

Reference:

1. https://www.britannica.com/science/cardiovascular-disease
2. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds