**Cardiovascular Risk Prediction**

**Pratik M Gumble**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract**

Cardiovascular disease is a major health burden worldwide in the 21st century. The information which is gathered by data analysis of hospitals is utilizing by applying different blends of calculations and algorithms for the early-stage prediction of Cardiovascular ailments. Machine Learning is one of the slanting innovations utilized in numerous circles far and wide including the medicinal services application for predicting illnesses. In this project, we compared the accuracy and other parameters of machine learning algorithms that could be used for predictive analysis of heart diseases and predicting the overall risks. The proposed experiment is based on a combination of standard machine learning algorithms such as Logistic Regression, Random Forest, K-Nearest, XG Boost classifier & SVM. Likewise, we attempt to generate predictive model that identify the patient whether they are having a Cardiovascular disease or on the other hand not.

***Keywords: machine learning, random forest, logistic regression, XG Boost, KNNclassifer, SVM***

1. **Problem Statement**

The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

1. **Introduction**

Cardiovascular disease is presently the leading problem of death worldwide. An expected 3.8 million men and 3.4 million women die each year from cardiovascular disease. Generally, the major causes of heart disease are diabetes, obesity, unhealthy diet, overweight, excessive alcohol use, and physical inactivity. The biological factors affected by these causes, along with age, thalassemia, chest pain, and preexisting conditions, also contribute significantly. Several datasets have been proposed to comprehensively train a machine learning model based on the several features and parameters identified by experts in heart disease prediction or heart disease detection

To begin with, the work we are using different machine learning algorithms. In this project, the classification of machine learning techniques and algorithms are used to generate best predictive model. We will use a dataset to understand how to build different classification models in python from scratch. The models that will be introduced in this project are:

1. Logistic Regression
2. Random Forest
3. XG Boost
4. K-Nearest Neighbor
5. SVM

After we build the models using training data, we will test the accuracy of the model with test data and determine the appropriate model for this dataset.

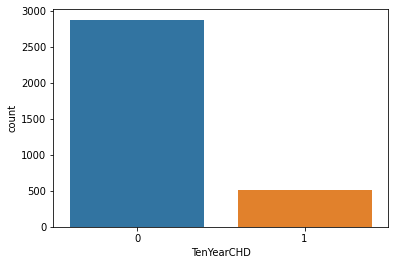
1. **Dataset Analysis**

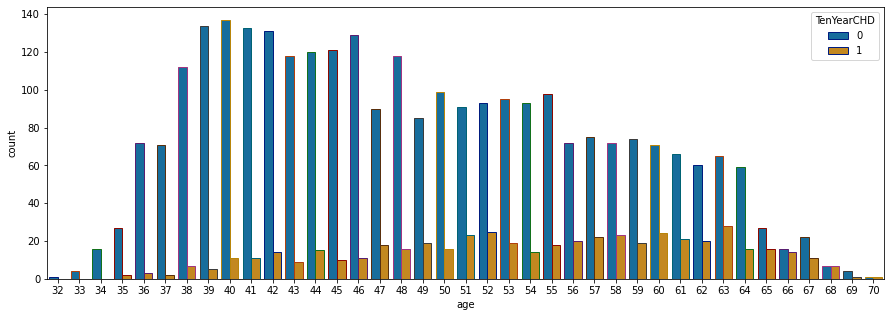
The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The dataset provides the patients’ information. It includes over 4,000 records and 15 attributes.

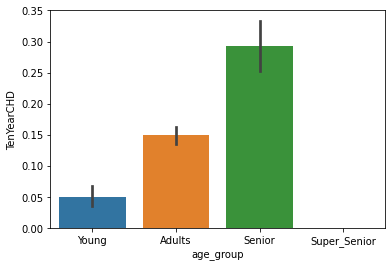
1. **Steps involved:**
2. **Exploratory Data Analysis**

After loading the dataset, we performed this method by comparing our target variable that is **10-year risk of coronary heart disease** with other independent variables. Performing an Exploratory Data Analysis allows data scientists to detect errors, debunk assumptions, and much more to ultimately select an appropriate predictive model.

This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

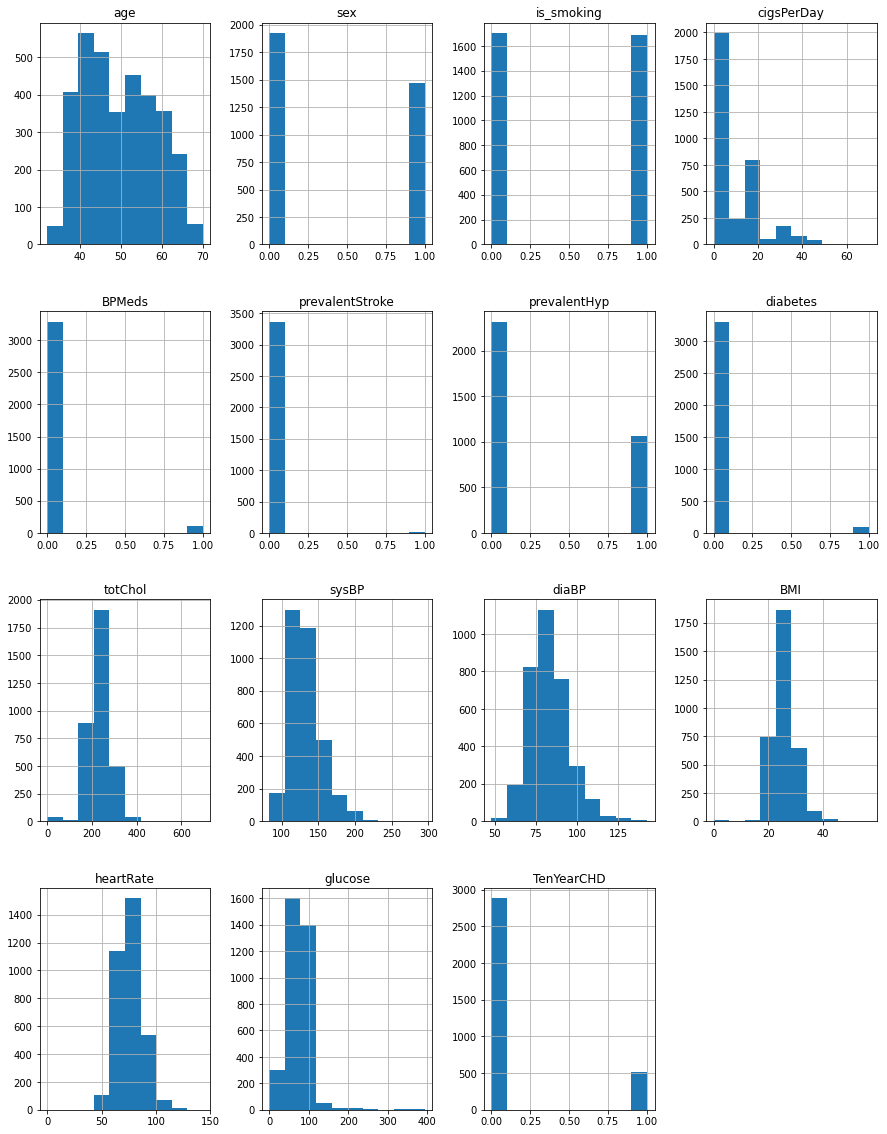


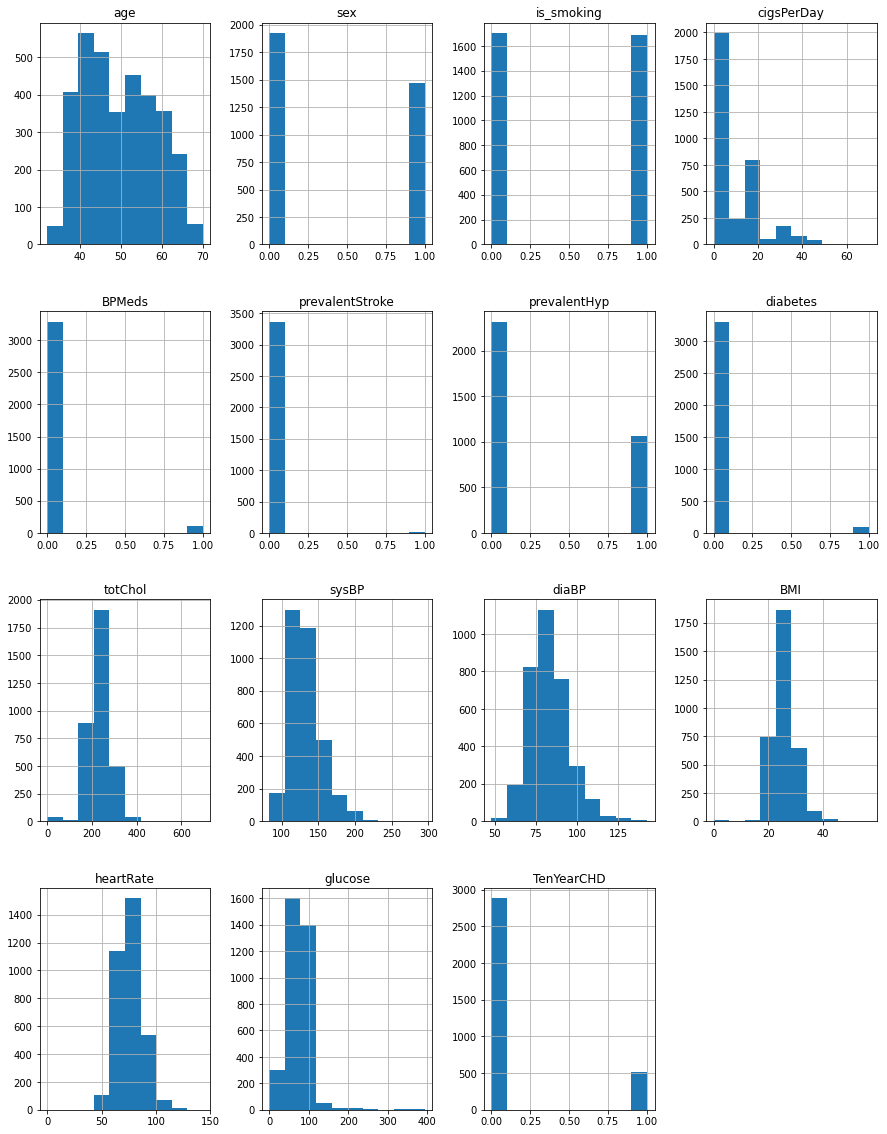




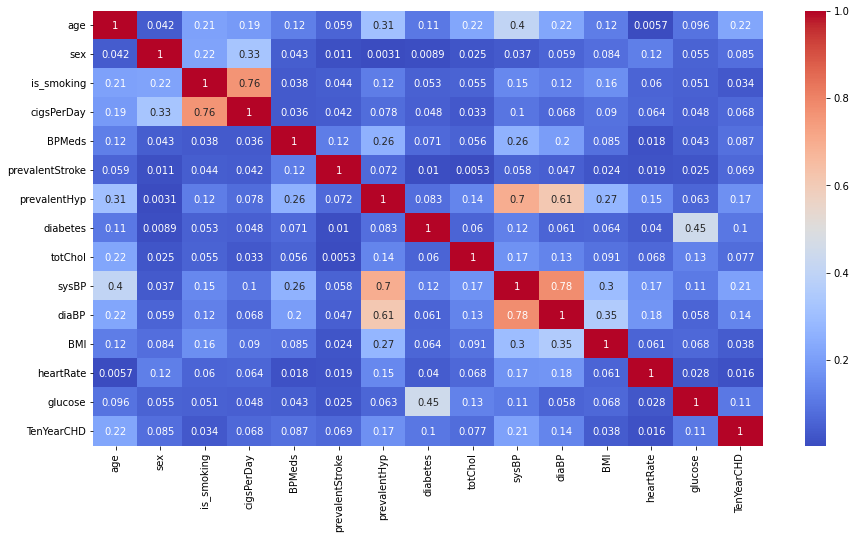
From the graph and statistical analysis it is clear that most of the people in the dataset have not 10-year risk of coronary heart disease CHD(2879 peoples) and 511 peoples have 10-year risk of coronary heart disease CHD.

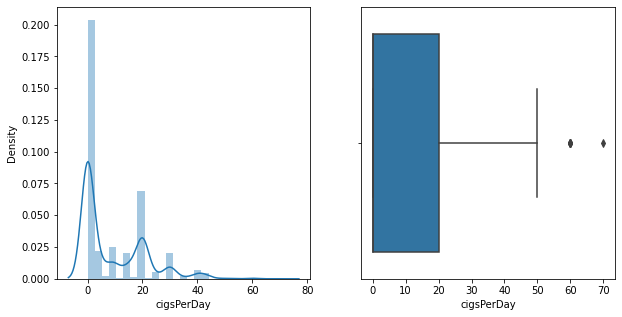
A histogram is a graph showing frequency distributions. It is a graph showing the number of observations within each given interval. With Histograms we can see the shape of each feature and provides the count of number of observations in each bin.

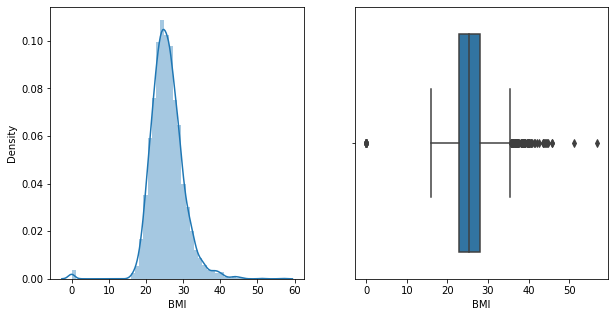
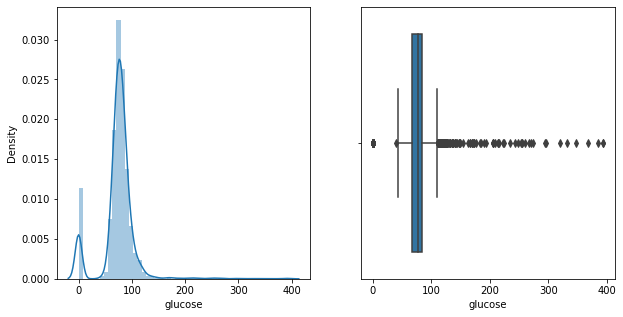


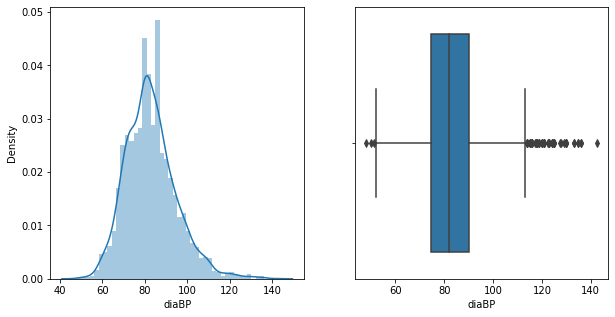


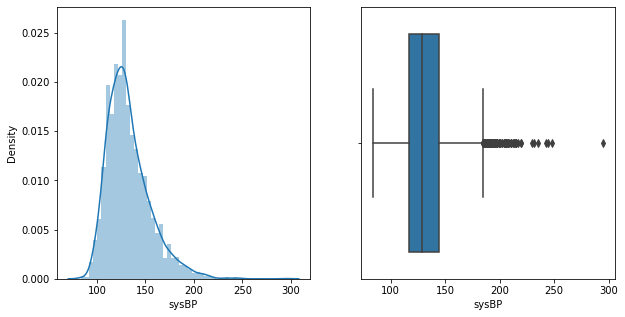
We can see there is a positive correlation between age, glucose, diaBP, SysBP, Diabetes, pravalentHyp with TenYearCHD . This makes sense since, all these features increase a greater chance of having heart disease.

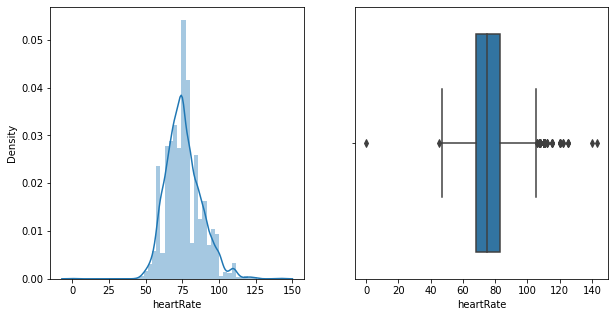


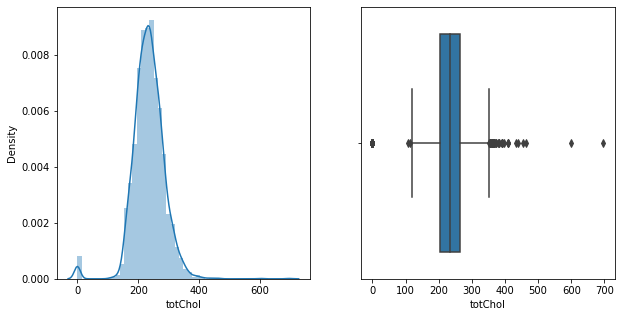










Boxplot is used to find out outliers present in dataset. From the box plots, outliers are present in cigsPerDay, totchol, BMI, Heart rate, glucose, sysBP, diaBP.

The Outliers are removed using two methods

1. Interquartile Range
2. Z-score
3. **Outliers Removal**
4. **Interquartile Range (IQR)**

After removing outliers using IQR, the data contains (2439, 15)

1. **Z-Score**

After using Z-score to detect and remove outliers, the number of records in the dataset is (3046, 15)**.**

As the number of records available is higher after Z-score, we will proceed with z score.

1. **Imbalanced Dataset**

Sampling technique comes to save us and deal with imbalanced data.

There are two sampling techniques available to handle the imbalanced data:

**Under Sampling**

**Over Sampling**

Here we choose Oversampling technique as Unlike under-sampling, this method leads to no information loss.

Now to overcome data imbalance we apply Oversampling.

In oversampling we use SMOTE.

It stands for Synthetic Minority Oversampling Technique.

1. **Algorithms:**
2. **Logistic Regression**

This type of statistical model (also known as logit model) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure.

1. **Random Forest**

The random forest algorithm provides flexibility and robustness for classification tasks using tabular data, which few other standard models can. Given its simplicity and versatility, the random forest classifier is widely used for fraud detection, loan risk prediction, and predicting heart diseases. With the ensemble learning theorem, the random forest classifier combines results from several decision trees and optimizes training. It aims to utilize different subsets and find the best combinations to increase the dataset’s predictive accuracy. The first step is building, optimizing, mixing, and matching several decision trees. Next, it uses these trees for prediction and ensembles their results to yield the final output prediction.

1. **XG Boost Classifier**

[XGBoost](https://xgboost.ai/), which stands for Extreme Gradient Boosting, is a scalable, distributed [gradient-boosted](https://en.wikipedia.org/wiki/Gradient_boosting) decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

It’s vital to an understanding of XGBoost to first grasp the machine learning concepts and algorithms that XGBoost builds upon: supervised machine learning, decision trees, ensemble learning, and [gradient boosting](https://developer.nvidia.com/blog/gradient-boosting-decision-trees-xgboost-cuda/).

1. **K-Nearest Neighbors**

As the name says, a k neighbor’s classifier takes a data point and finds k other data points nearest to it in the vector space. In a supervised fashion, KNN creates clusters of the data samples having the same target value. Whenever a new value needs to be classified, it uses a distance metric to assign it to one of the classes. For heart disease detection, there are only two classes that KNN needs to build. Thus, it is pretty robust and efficient for this task. Euclidean distance is one of the popular distance metrics used by KNN, but there are many more available. However, the metric choice also impacts the classifier's speed For larger datasets, KNN is already relatively slower than its contemporaries.

1. **Support Vector Machine**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

### **Model performance:**

1. **Logistic Regression:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test accuracy | Test precision | Test recall | Test f1 score | Test ROC-AUC |
| 68% | 65% | 71% | 68% | 68% |

1. **Random Forest**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test accuracy | Test precision | Test recall | Test f1 score | Test ROC-AUC |
| 90% | 87% | 93% | 90% | 90% |

1. **XG Boost**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test accuracy | Test precision | Test recall | Test f1 score | Test ROC-AUC |
| 80% | 77% | 83% | 80% | 80% |

1. **KNN**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test accuracy | Test precision | Test recall | Test f1 score | Test ROC-AUC |
| 77% | 69% | 93% | 79% | 78% |

1. **SVM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test accuracy | Test precision | Test recall | Test f1 score | Test ROC-AUC |
| 74% | 70% | 80% | 75% | 75% |

1. **Let’s collect all our best models!**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | Test accuracy | Test precision | Test recall | Test f1 score | Test ROC-AUC |
| Logistic Regression | 68% | 65% | 71% | 68% | 68% |
| Random Forest | 90% | 87% | 93% | 90% | 90% |
| XG Boost | 80% | 77% | 83% | 80% | 80% |
| KNN | 77% | 69% | 93% | 79% | 78% |
| SVM | 74% | 70% | 80% | 75% | 75% |

It is quite evident from the results Random forest is the best model that can be used for cardiovascular risk prediction dataset since all the performance metrics (accuracy, precision, recall and roc-auc score) show a higher value for the random forest model !

1. **Predictive Model**

Predictive model help us to understand possible future occurrences by analyzing our existence model. Predictive models make assumptions based on what has happened in the past and what is happening now. If incoming, new data shows changes in what is happening now, the impact on the likely future outcome must be recalculated, too.

In our predictive model we can see that Random forest classifier predict well as patient having 10-year risk of coronary heart disease or not.

1. **Feature Importance**

Feature (variable) importance indicates how much each feature contributes to the model prediction. Basically, it determines the degree of usefulness of a specific variable for a current model and prediction.

By analyzing variable importance scores, we would be able to find out irrelevant features and exclude them. Reducing the number of not meaningful variables in the model may speed up the model or even improve its performance.

In almost all classifier CigsPerDay, age, sysBP, diaBP, tatchol, heartrate, BMI, glucose are the important feature to determine person having 10 year risk for heart diastase or not. Out of that CigsPerDay, age and BP are the most important feature which decided the person having chance of 10 year heart diseases or not.

**Conclusion:**

The main aim of this project is to compare the accuracy and other parameters like precision, recall ,f1 score, auc roc of all the classification algorithms to evaluate the risk of 10-year CHD using 14 features. After implementing four classification models and comparing their accuracy and other scores, we can conclude that for this dataset Random forest Classifier is the appropriate model to be used. Also out of all features CigsPerDay, age and BP are the most important feature which decided the person having chance of 10 year heart diseases or not.

**References:**

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya