Cardiovascular Risk Prediction

Presented by- Madhavi Mali

ABSTRACT

Several machine learning (ML) algorithms have been increasingly utilized for cardiovascular disease prediction. Cardiovascular disease (CVD) is the leading cause of mortality worldwide. Accurately identifying subjects at high-risk of CVD may improve CVD outcomes. We aim to assess and summarize the overall predictive ability of ML algorithms in cardiovascular diseases. Even if Medicine and Computer Science seem

apparently intangible domains, they collaborate each other for few decades. One of the faces of this cooperation is Data

Mining, a relative new and multidisciplinary field capable to extract valuable information from large sets of data. Despite
this fact, in cardiology related studies it was rarely used. We assume that some data mining tools can be used as a substitute for some complex, expensive, uncomfortable, time consuming, and sometimes dangerous medical examinations. This paper aims to show that cardiovascular diseases may be
predicted by classical risk factors analyzed and processed in a "non-invasive" way.

PROBLEM STATEMENT

The World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression Data Preparation

SOURCE

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). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Variables Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

ATRIBUTE INFORMATION

Demographic:

- Sex: male or female(Nominal)
- Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- Education: no further information provided

Behavioral:

- Current Smoker: whether or not the patient is a current smoker (Nominal)
- 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.)

Information on medical history:

- BP Meds: whether or not 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)

Information on current medical condition:

- 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)

Target variable to predict:

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

INTRODUCTION

Cardiovascular disease (CVD), the leading cause of mortality in the world, has been an important public health concern globally, causing massive socioeconomic burdens on patients, families, and countries every year. Risk stratification can be used to identify high-risk subjects of having CVD through predictive models, and then interventions, such as lifestyle changes and initiation of statins use, specific to this target population can reduce the risk of developing CVD and promote the primary prevention of CVD. Several guidelines on the assessment and management of CVD recommended applying predictive models to identify the high-risk population and support clinical decision-making. Widely used predictive models, such as the Pooled Cohort Equations (PCE) and the Framingham CV risk equation (FRS)[6] have been externally validated in multiple populations, however, the results demonstrated that both of them were in moderate discrimination and poorly calibrated.

Machine learning (ML) algorithms have emerged as highly effective methods for prediction in cardiovascular research.[3,11,12] They can capture the complex interactions between predictors and nonlinear relationships between predictors and outcomes, producing better predictive performance than traditional statistical models. Studies suggested that random forest (RF),[13] support vector machine (SVM),[14,15] outperformed traditional models. However, results are still inconsistent, a recently published meta-analysis showed that ML-based predictive models do not perform better than logistic regression.

The Kazakh ethnic population live in the remote northwest of China, Xinjiang, and they have similar genetic backgrounds to Caucasians. Most of them live in mountainous pastures, and this population has a relatively high incidence rate of CVD due to their unique lifestyle, dietary habits, and genetic characteristics.[17] Therefore, it is crucial to identify high-risk subjects who may benefit from targeted interventions using CVD predictive models for the prevention of CVD.

Consequently, we sought to assess the potential value of several widely used ML algorithms in predicting future CVD events in this Kazakh Chinese population and explored which ML-based model generated the best predictive performance and most accurate prediction. Then we evaluated the clinical usefulness of the best model through decision curve analysis and determined whether it could be used to guide CVD prevention and support the clinical decision-making process.

BREAKDOWN OF DATASETS

In order to go ahead for data visualization upon key factors we need to go for certain extra steps before proceeding to the main segment. In this part we are going with the following steps:

1. Importing Analytical necessary library classes for future analysis.
2. Reading the csv data file from Google drive.
3. Setting figure size for future visualization.

4. Removing future warnings in seaborne plots.
5. Visualizing all the columns of the respective Data frame.
6. Viewing all data information
7. Dropping duplicates if any.
8. Checking the Unique values, null count and data type of each column.
9. Converting the data types to similar objects as the Analysis Demands.
10. Checking for outliers

EXAMINING NULL / MISSING VALUES

Some values in our dataset are null or missing. These values affect the accuracy and performance of the models that predict the outcome, so these need to be handled. While analyzing our dataset the first thing we will do is to examine the null or missing values in our dataset. This makes our result accurate. Missing values are more in Size & Rating columns as can be seen by plotting graphs. Hence several methods are used to remove these values.

from 3390 rows, 386 rows contain 423 null values, So we are dropping 386 data entries for better prediction of our model ,we have enough data to test train model after drop small amount of rows

So the shape of filter data is (3004, 15)

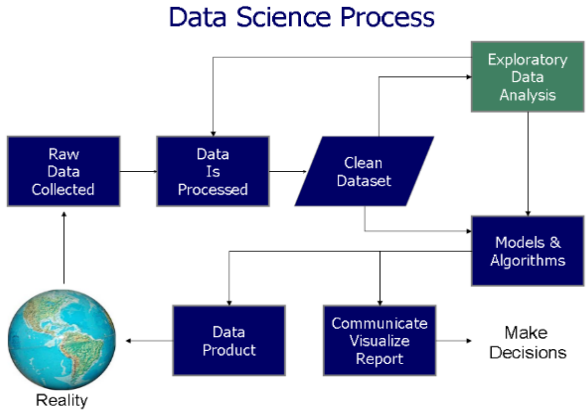
DATA CLEANING

Data cleaning is the foremost step in any data science project. Cleaner the data, better are the results. As the proverb goes by saying "More Data beats clever algorithm, but better data beats more Data" – Peter Norvig. To begin with our data cleaning, first we remove the duplicate values. Then we remove unnecessary characters in our dataset.

After doing so we find the unique values of each column and make the necessary changes in each column like converting datatypes, removing the null and 'nan' values. Lastly, we have done exploratory data analysis of our dataset.

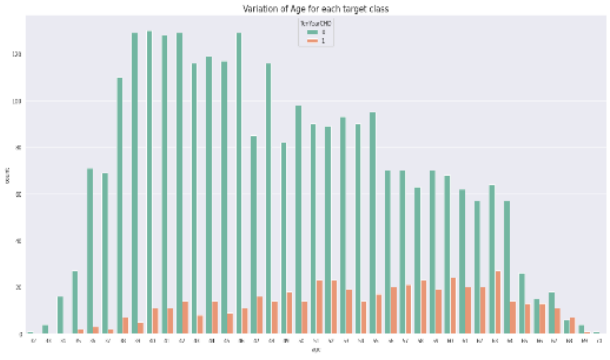
EDA

In statistics, exploratory data analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling and thereby contrasts traditional hypothesis testing. EDA is helped us figuring out various aspects and relationships among the target and the independent variables.



Observation 1:

First we Observed and plotted the graph for analysis of age for each target class .

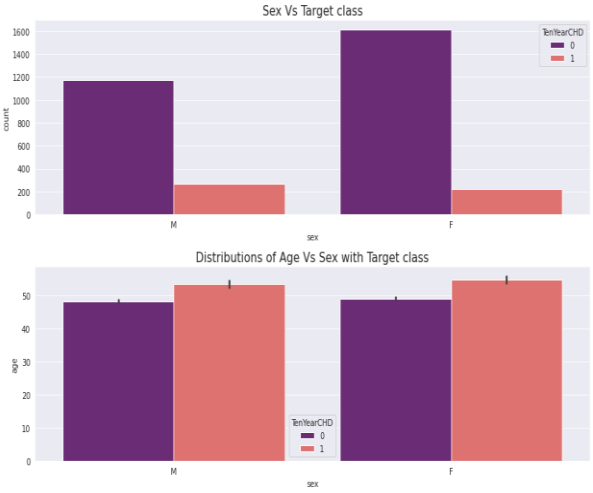


We observed the following

- Coronary heart disease(CHD) increases after age 51.
- Age group (34 < Age < 51) are at lower risk of cardiovascular disease

Observation 2

Next we plotted Analysis Of Age vs Sex with Target class :

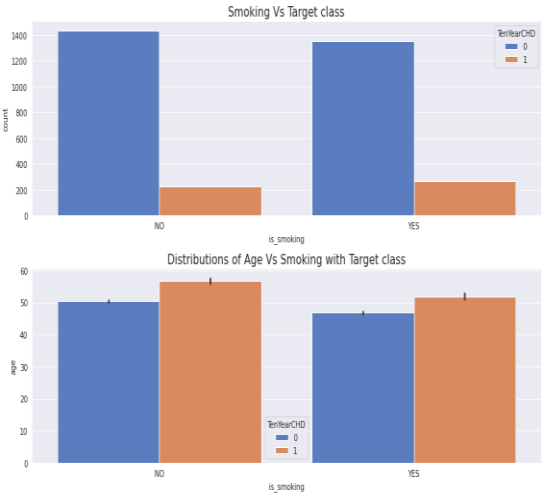


We observed the following:

- We can see from the countplot that no. of male heart patient is more than female.
- We can see from the barplot that male get early CHD as compared to female.

Observation 3:

Next we plotted Analysis Of Age vs Smoking with Target class

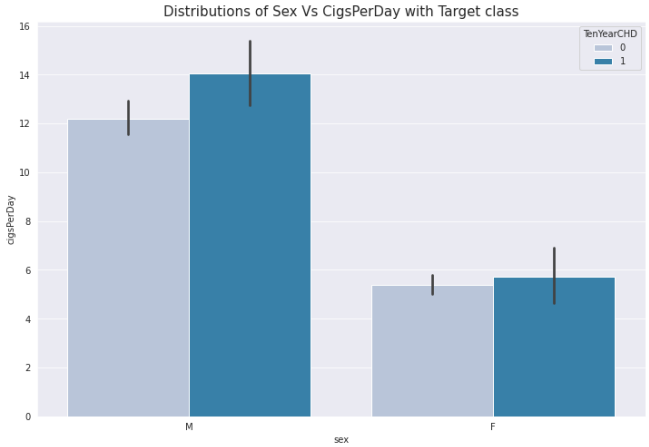


We observed the following:

- We can see from the countplot that no. of patient those who smoke more than as compared to those who won't.
- We can see from the barplot that those who smoke get early heart disease as compared to those who don't.

Observation 4:

Next we plotted Analysis Of Cigs per day vs Sex with Target class :

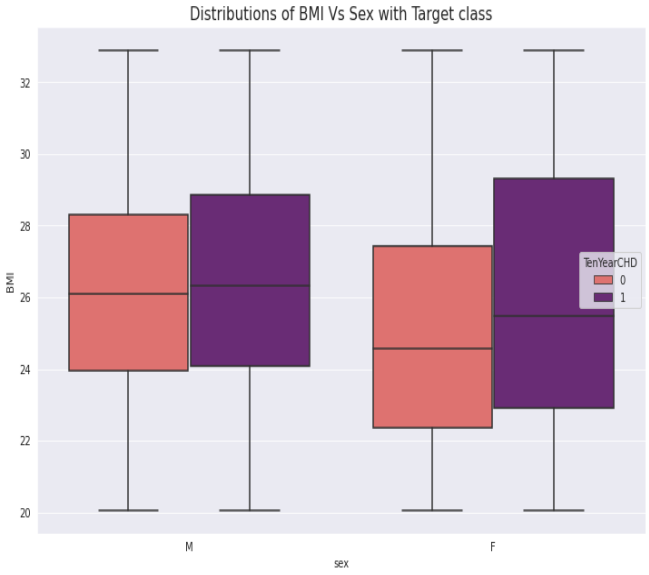


We noted the following :

- As we can see the barplot we can say that no. of cigsperday taken by male is more than female.
- So, male heart patient is more as compared to female.
- In case of male CHD = 1 when he take cigsperday > 12.1 and in case of female CHD = 1 when she take cigsperday > 4.8.

Observation 5:

Next we plotted graph of Analysis Of BMI vs Sex with Target class :



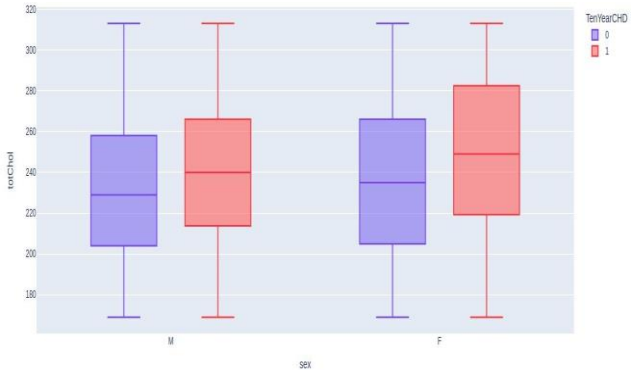
We observed the following:

- As we can see the boxplot we can say that female BMI is more than male BMI. that's leads to OVERWEIGHT.

- So, female CHD patient more than male CHD patient.
- If your BMI is:
    - below 18.5 – you're in the underweight range.
    - between 18.5 and 24.9 – you're in the healthy weight range
    - between 25 and 29.9 – you're in the overweight range
    - between 30 and 39.9 – you're in the obese range

Observation 6

Next we plotted graph of Analysis Of Cholesterol vs Sex with Target class :

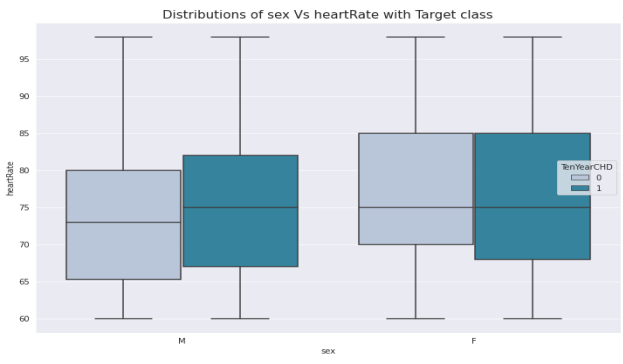


We observed the following :

- As we can see the boxplot we can say that female cholesterol is more than male cholesterol that's leads to OVERWEIGHT.
- So, In female heart disease is more due to cholesterol.

Observation 7

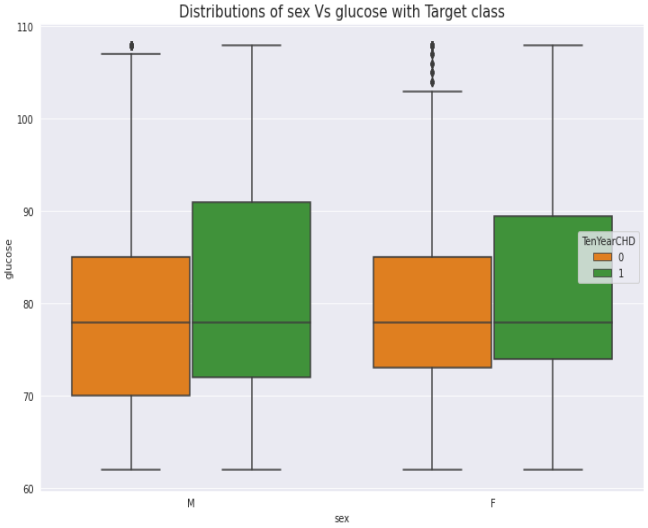
Next we plotted graph of Analysis Of Heart Rate vs Sex with Target class :



We can see the box plot we can say that for Female heart disease patients has more Heart Rate as compared to male heart disease patients.

Observation 8

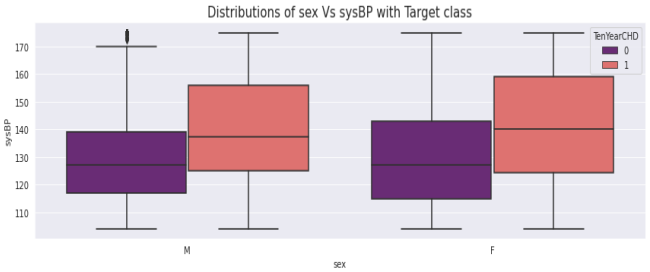
Then we plotted graph of Analysis Of Glucose vs Sex with Target class :

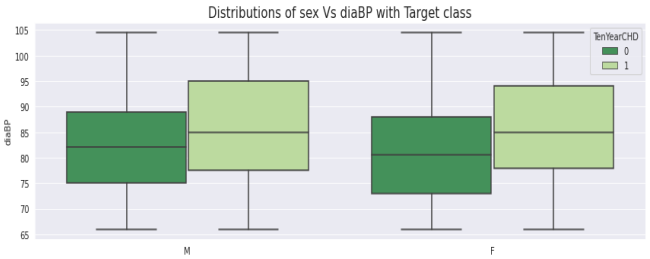


We can see from the box that for male heart disease patients has more glucose level as compared to female heart disease patients.

Observation 9

Analysis Of Systolic and Diastolic vs Sex with Target class :



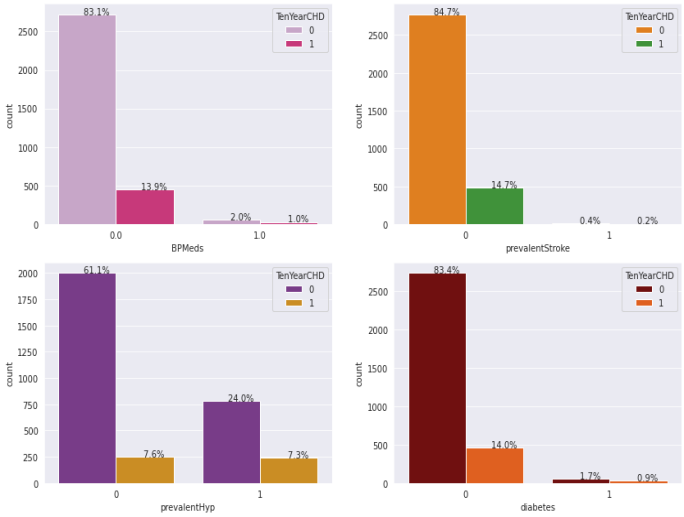


- We can see the box plot and say that for female heart disease patients has more

Systolic BP level as compared to male heart disease patients.

- Normal < 120 mmHg.

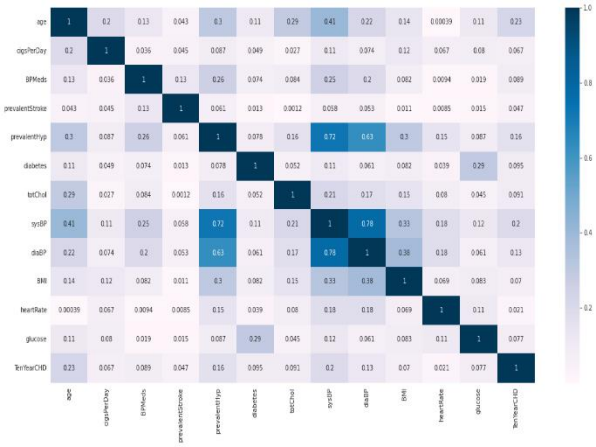
Observation 10



Correlation Matrix

Correlation is a statistical measure that expresses the strength of the relationship between two variables. Positive correlation occure when two variables move in same direction; as one increases so do the other. Negative correlation occurs when two variables move in opposite directions; as increases, the other decreases.

Correlation can be used to test hypotheses about cause effect relationships between variables. Correlation is often used in the real world to predict trends.



We observed that

- sysBP is moderately correlated with prevalenthyp, i.e. prevalent hypertension.
- diaBP and sysBP
- are somewhat moderately correlated.
- glucose level are also moderately correlated to whether patient is diabetic

Label Encoding

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

sex	is_smoking
M	NO
F	YES
M	YES
F	YES
F	NO

- We have two categorical columns i.e sex and is_smoking.

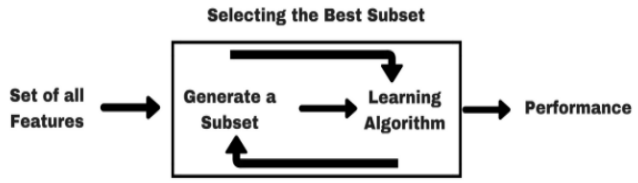
sex	is_smoking
1	0
0	1
1	1
0	1
0	0

- After applying label encoding we converted into 0's and 1's.

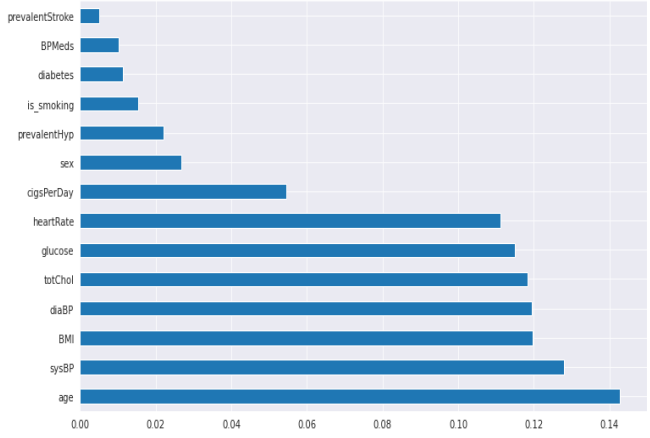
Feature Selection:

While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant. If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model. Hence it is very important to identify and select the most

appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning.

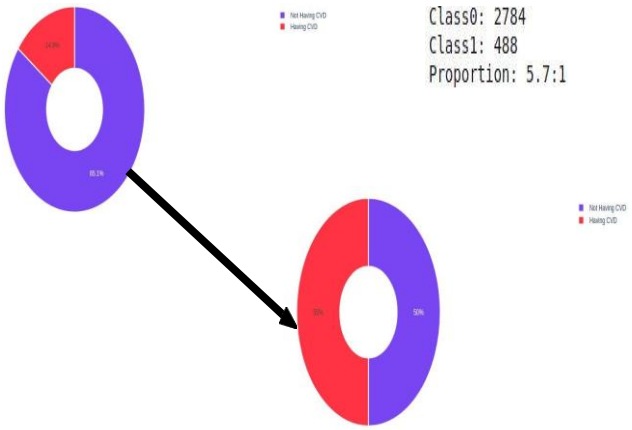


Feature selection is one of the important concepts of machine learning, which highly impacts the performance of the model. As machine learning works on the concept of "Garbage In Garbage Out", so we always need to input the most appropriate and relevant dataset to the model in order to get a better result.



- For feature selection we used ExtraTreeClassifiers.
- We found that every feature is important.

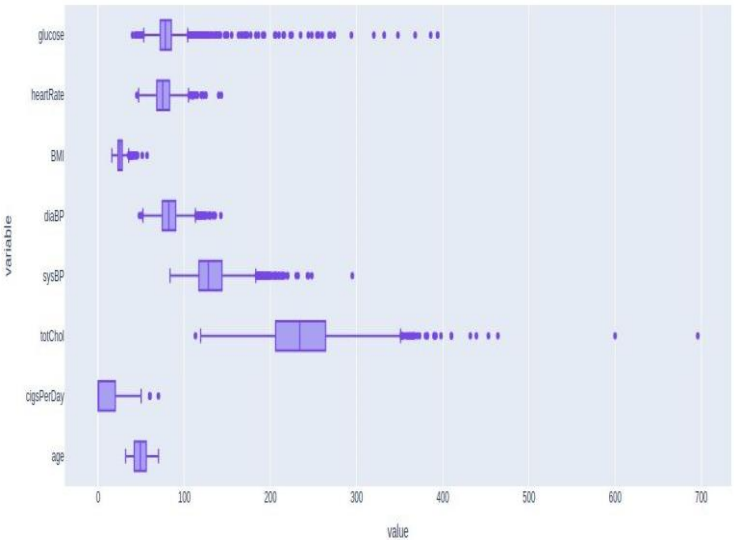
Handling Imbalanced Data:



SMOTE (Synthetic Minority Oversampling Technique) works by randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbours.

OUTLIER

Outliers is a data point in the dataset that differs significantly from the other data or observation. The thing to remember that, not all outliers are the same. Some have a strong influence, some not at all. Some are valid and important data values. Some are simply errors or noise. Many parametric statistics like mean, correlations, and every statistic based on these is sensitive to outliers. since the assumptions of standard statistical procedures or models, such as linear regression and ANOVA also based on the parametric statistic, outliers can mess up your



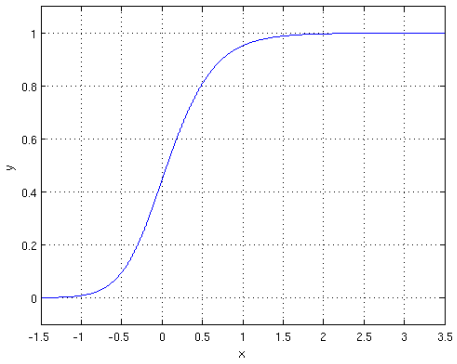
Model Building

I. Logistic Regression

Logistic Regression is actually a classification algorithm that was given the name regression due to the fact that the mathematical formulation is very similar to linear regression.

The function used in Logistic Regression is sigmoid function or the logistic function given by:

$$f(x)= 1/1+e^{(-x)}$$

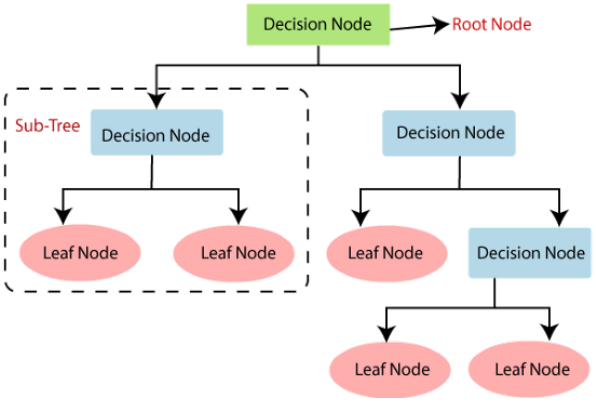


The optimization algorithm used is: Maximum Log Likelihood. We mostly take log likelihood in Logistic:

$$\ln L(\mathbf{y},\beta)=\ln\prod_{i=1}^{n}f_i(y_i)=\sum_{i=1}^{n}\left[y_i\ln\left(\frac{\pi_i}{1-\pi_i}\right)\right]+\sum_{i=1}^{n}\ln(1-\pi_i)$$

II. Decision Tree Classifier

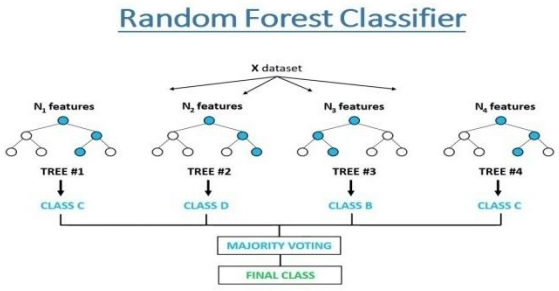
The decision tree creates classification or regression models as a tree structure.



It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed. The final tree is a tree with the decision nodes and leaf nodes. A decision node has at least two branches. The leaf nodes show a classification or decision. We can't accomplish more split on leaf nodes-The uppermost decision node in a tree that relates to the best predictor called the root node. Decision trees can deal with both categorical and numerical data.

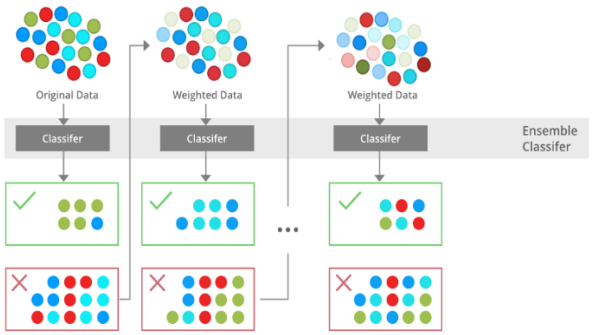
III. Random Forest Classifier

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.



IV. XGB Classifier

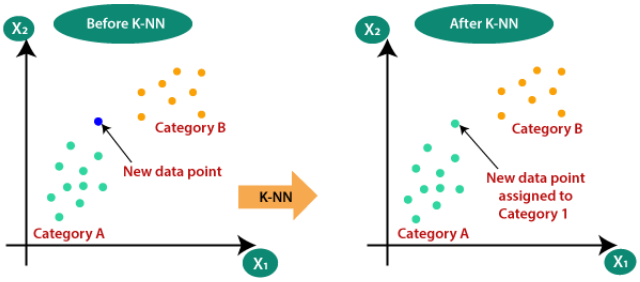
XGBoost: XgBoost (Extreme Gradient Boosting) library of Python was introduced at the University of Washington by scholars. It is a module of Python written in C++, which helps ML model algorithms by the training for Gradient Boosting.



In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and the variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

V. K-Neighbors Classifier

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.



It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

VI. Support Vector Machine Classifier

SVM is used mostly when the data cannot be linearly separated by logistic regression and the data has noise. This can be done by separating the data with a hyperplane at a higher order dimension.

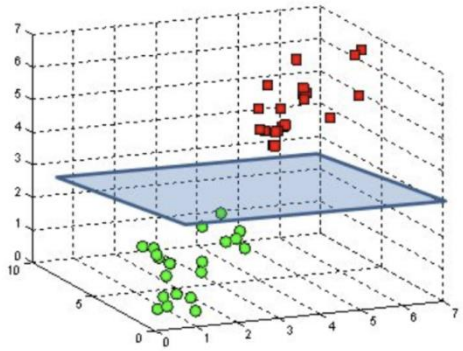
In SVM we use the optimization algorithm as:

$$\min_{\xi,w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t} \quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i,$$

$$\xi_i \geq 0;\ i = 1, \ldots, m.$$

where C is a cost parameter and ξ_i's are slack variables.



We use hinge loss to deal with the noise when the data isn't linearly separable.

Kernel functions can be used to map data to higher dimensions when there is inherent non linearity.

HYPER PARAMETER TUNING

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We have performed hyperparamter tuning to find the best suiting model and improve the accuracy.

Evaluating Model

		Model	Precision	Recall	F1-Score	Accuracy	ROC_AUC
Training set	0	Logistic regression	0.6757	0.6939	0.6847	0.6803	0.7488
	1	DecisionTree Classifier	0.6903	0.8092	0.7450	0.7229	0.8020
	2	RandomForest Classifier	0.7440	0.8048	0.7732	0.7638	0.8648
	3	XGB Classifier	0.8411	0.8218	0.8313	0.8332	0.9249
	4	KNeighbors Classifier	0.8056	0.9466	0.8704	0.8590	0.9535
	5	SVC Classifier	0.7510	0.7944	0.7721	0.7654	0.8502
Testing set	0	Logistic regression	0.6708	0.6817	0.6762	0.6741	0.7368
	1	DecisionTree Classifier	0.6781	0.7806	0.7258	0.7056	0.7766
	2	RandomForest Classifier	0.7174	0.7626	0.7393	0.7316	0.8188
	3	XGB Classifier	0.8305	0.7842	0.8067	0.8124	0.8946
	4	KNeighbors Classifier	0.7206	0.9137	0.8057	0.7801	0.8616
	5	SVC Classifier	0.7242	0.7698	0.7463	0.7388	0.8233

Conclusion:

- In the given dataset we observe that Coronary heart disease increases from age 51 to 67 then decreases.
- We draw the countplot and observe that no. of male heart patients is more than female and also notice that male get early age heart diseases as compared to females.
- We observe no. of heart patients who smoke more than as compared to those who won't and also notice that those who smoke get early heart disease as compared to those who won't.
- We draw the bar plot and observe that no. of cigsperday taken by male is more than female. So, male heart patients is more as compared to females.
- We draw the boxplot and observe that female BMI(The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2) is more than male BMI. that's leads to OVERWEIGHT and So, female CHD patients is more than male CHD patients.
- We draw the boxplot and observe that female Cholesterol is more than male Cholesterol. that's leads to OVERWEIGHT and So, in that case also female CHD patients is more than male CHD patients.
- We Observe that Female heart disease patients has more Heart Rate as compared to male heart disease patients.
- We also observe that male heart disease patients has more glucose level as compared to female heart disease patients.
- In the Models Evaluation Table(Testing set) our auc-roc score is more 0.80 except Logistic regression and Decision Tree.So we can say that our model predicted the classes in a good manner.
- XGBClassifier are performing well which has the best Recall,Precision,F1-Score and Accuracy Score.