TECHNICHAL DOCUMENT

BY:

TANJUL GOHAR

KANIKA SINGH

Cardiovascular Risk Prediction

Abstract: Identifying people at risk of cardiovascular diseases (CVD) is a complicated task in medical field. Risk prediction models currently recommended by clinical guidelines are typically based on a limited number of predictors with sub-optimal performance across all patient groups. The proposed work predicts the probabilities of heart condition and classifies patient's risk level by implementing different data processing techniques like Naive Bayes, Decision Tree, Logistic Regression, SDG Classifier, Support Vector Machine, KNN classifier.

1. INTRODUCTION

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world.

Coronary Heart Disease (CHD) is the most common type of heart disease, killing over 370,000 people annually.

Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.

2. DATA SET DESCRIPTION

The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The data set provides the patients' information. It includes over 3390 records and 17 attributes. Each attribute is a potential risk factor.

Attributes:

1. **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)

2. **Education**: no further information provided

3.**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.)

4. **Information on medical history**:

- BP Meds: whether or not the patient was on blood pressure medication (Nominal)

5. **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: blood glucose level (Continuous)

Target variable to predict:

10-year risk of developing coronary heart disease (CHD) — (binary: “1”, means “There is a risk”, “0” means “There is no risk”)

3. Steps in Data Pre-processing in Machine Learning

1. Acquire the dataset

Acquiring the dataset is the first step in data pre-processing in machine learning. To build and develop Machine Learning models.

2. IMPORTANT LIBRARY

- **pandas**: pandas provide high-performance data structures and operations for manipulating numerical tables and time series.
- **NumPy**: NumPy provides scientific computing capabilities such as a powerful N-dimensional array object, linear algebra, and random number capabilities.
- **sklearn**: scikit-learn provides tools for data mining and data analysis.
- **imblearn**: imbalanced-learn provides a number of re-sampling techniques commonly used in datasets showing strong between-class imbalance.
- **plotly**: Plot.ly is a graphing library which can produce interactive graphs.

3. **Import the dataset**

After importing the libraries, the next step was loading data into the data frame by the use of Pandas library. Now read the data from a CSV file into a Pandas Data Frame. We can see that the value from the data set were comma-separated.

4. **Data Cleaning**

1. **Identifying and handling the missing values**

The dataset has null values in some columns which is ‘education’, ‘BPMeds’, ‘totChol’, ‘BMI’, ‘Heartrate’, ‘glucose’ column. It was handled by KNN Imputer.

knnimputer replaces NaNs in the input data with the corresponding value from the nearest-neighbour column.

2. Encoding the categorical data

Few columns ‘sex’, ‘is_smoking’ have categorical values which is converted into numerical values by One Hot Encoding.

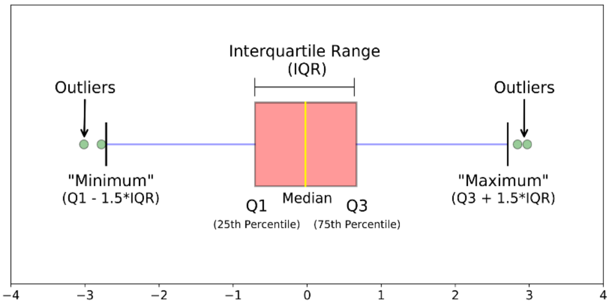
3. Drop irrelevant columns and duplicate

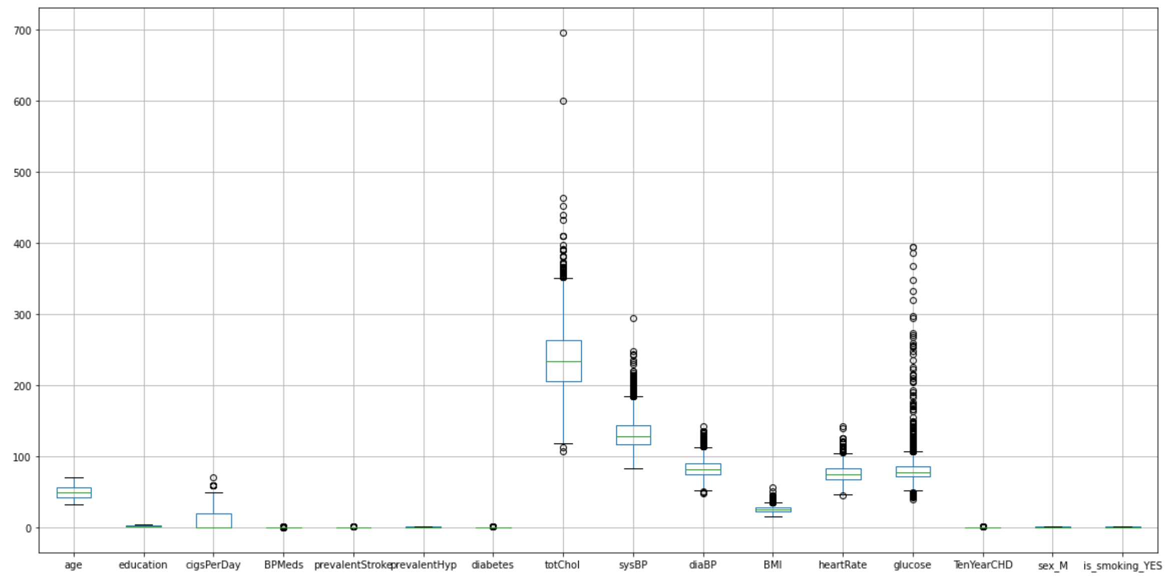
We drop the id columns because it has no correlation with heart disease. There were no duplicate values.

4. finding outliers

An outlier is a piece of data that is an abnormal distance from other points. We used boxplot for finding outliers.

- The box plot is based on a summary of five numbers: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. The lower and upper ranges are the boundaries of the data distribution. Any data points that show above or below the ranges can be considered outliers or anomalies.





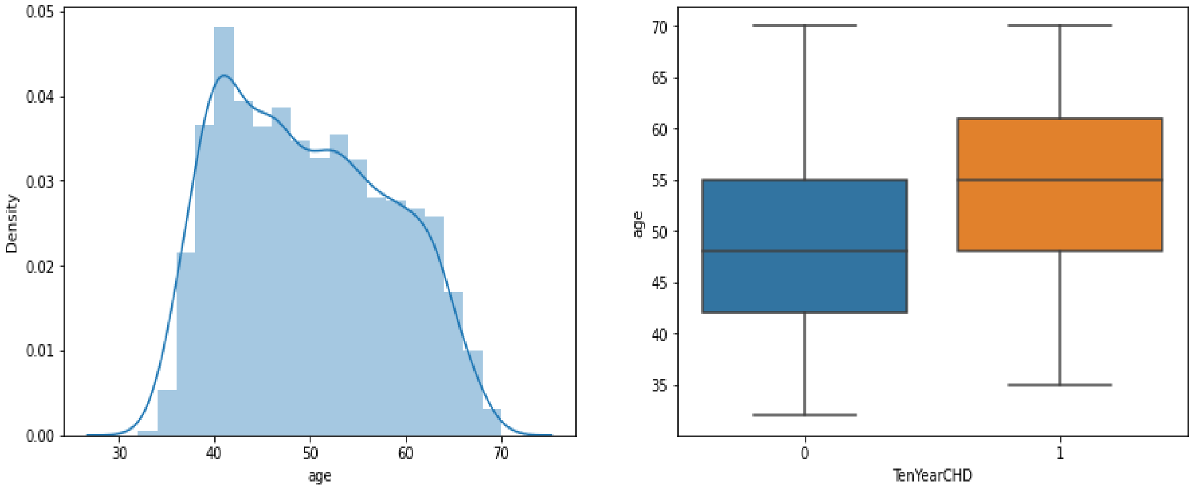
We have not removed outliers because it was important. There could be a possibility that a person might have a high risk of heart disease and by the removing it, it was a chance to delete some crucial information which would further affect the machine learning model.

1. **Exploratory Data Analysis**:

- The first step was to check the distribution different attributes and this was best visualized by histograms.
- I wanted to get extract insights and summarize the data. Data was extracted by the data visualization and plot a graph which is histogram and heatmap. I checked for were the distributions of the different attributes, correlations of the attributes with each other and the target variable
- The first step was to create a variable for categorical and continuous columns. Then make a graph with categorical and continuous with dependent variable (TenYearCHD).

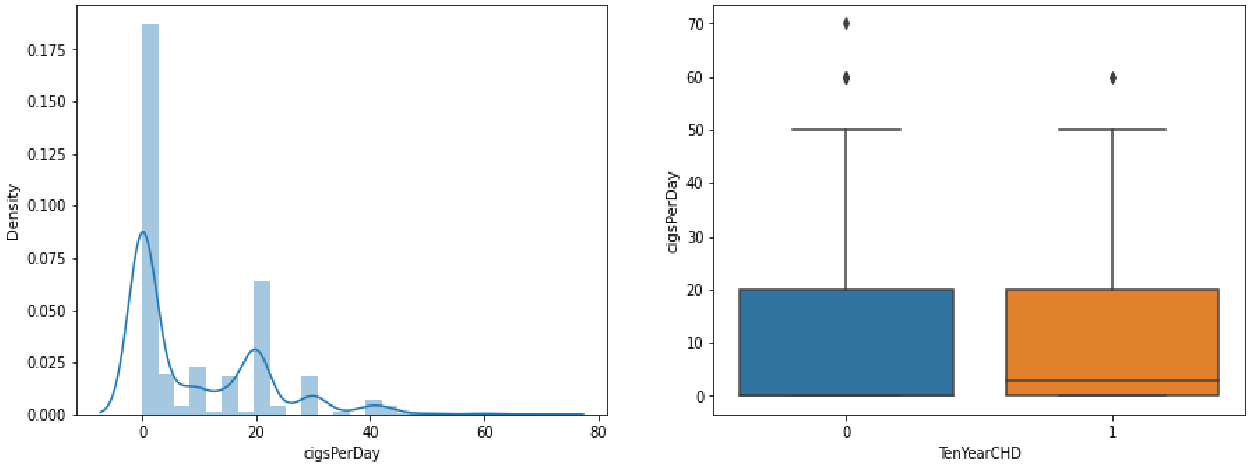
- ➢ **Plot histogram with continuous variable and plot boxplot of target variable: -**

1. Age Vs TenYearCHD



- Above histogram is showing that people age group from 30 years old to 70 years old and boxplot shows that there are no outliers.
- I checked distribution of the ages of the people who had CHD and the number of the sick generally increased with age.
- But the people who are at more risk of CHD are in the age group 50 years to 65 years old.

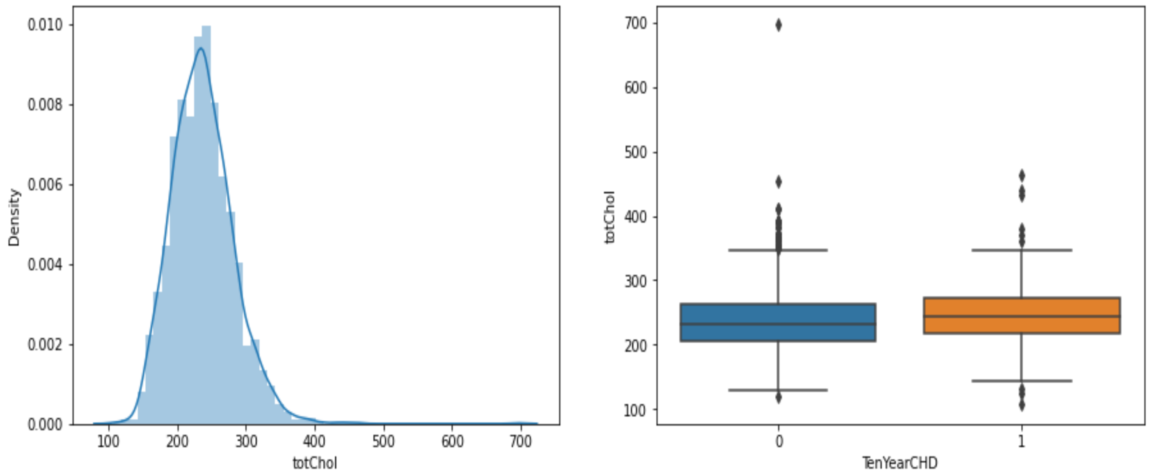
2. cigsPerDay Vs TenYearCHD



- CigsPerDay has highly uneven distribution. Most of the people in our dataset do not smoke and after that there are people who smoke 20 cigarettes a day.

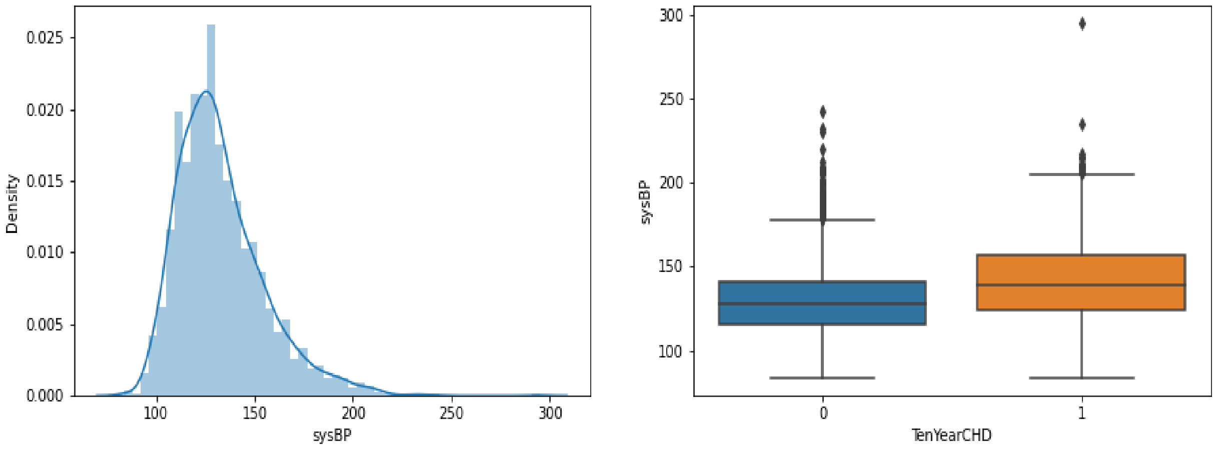
- The graph ranges from 0 cigarettes to 20 cigarettes a day which is very high quantity.
- People who smoked 0 cigarettes per day has a disease in other hand people who had 20 cigarettes per day has also a disease.
- CigsPerDay column is not that much relevant with our target feature.

3. totChol Vs TenYearCHD



- The normal total cholesterol in adult human is under 200mg/Dl.
- People who was at risk of CHD have total have cholesterol ranging in between 100-400.
- Some people were high risk of CHD which have cholesterol range 200-300. Approximately one fourth people have a disease.
- totChol is dependent to CHD.

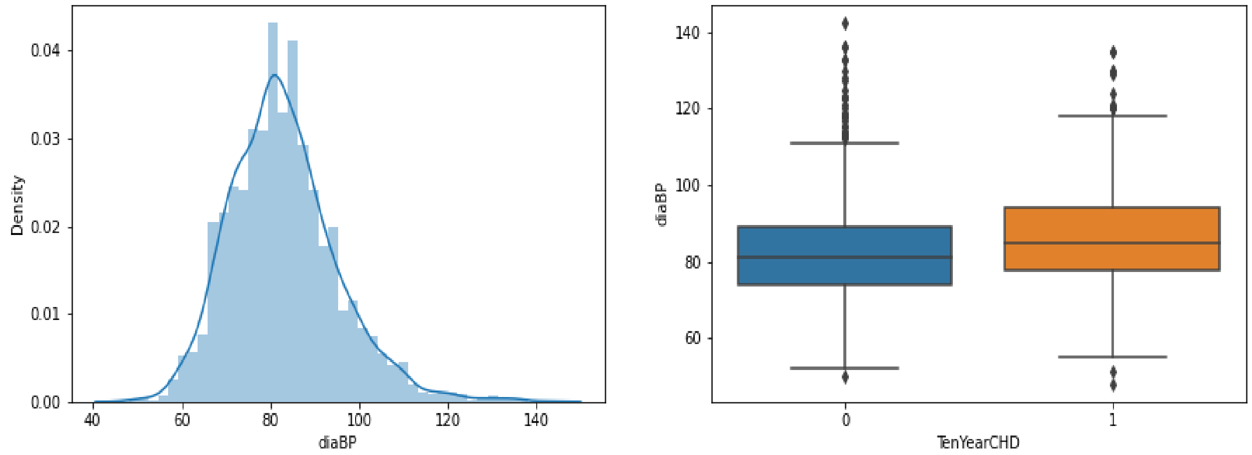
4. SysBP Vs TenYearCHD



- Systolic BP less than 120 is considered normal.

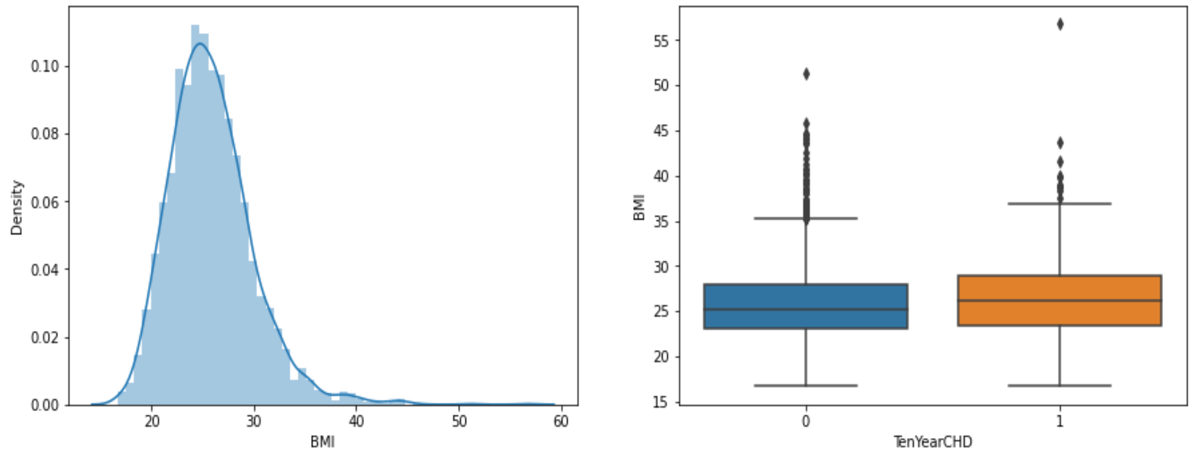
- we have cases where people are high risk of CHD there the readings of systolic BP between 120-130 mmHg.

5. diaBP Vs TenYearCHD



- Diastolic BP less than 80 is considered normal.
- Although there is no evidence to assume whether BP readings (systolic and diabolic) are contributing to the risk of CHD or not.

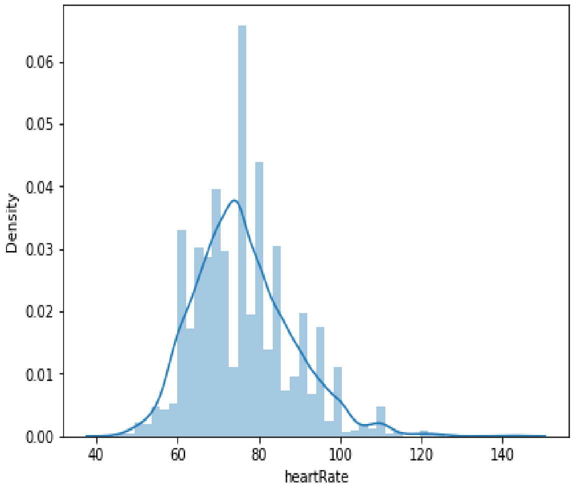
6. BMI Vs TenYearCHD

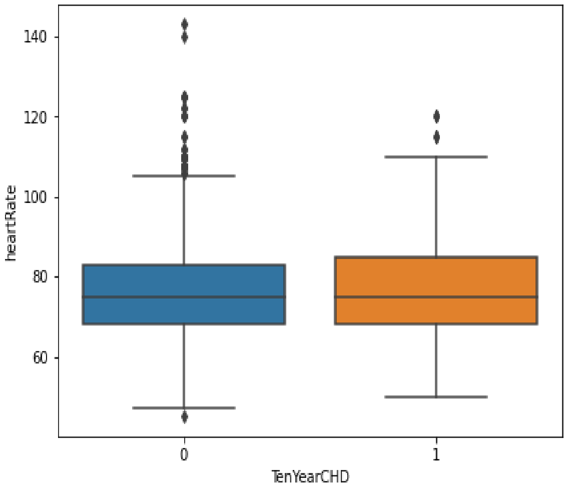


- BMI in our dataset ranges in between 15 to almost 60.
- People with BMI in the range 18.5 to 24.9 are considered healthy, 25.0 to 29.9 as overweight and after 30.0 are classified as obese.
- People with risk of coronary heart disease are spread quite evenly.

- So, there must be other factors other than BMI that are contributing to the potential risk of coronary heart disease. We have cases where people are in the category of obese but still not at risk of CHD and a lot of people in the category of healthy but still at the risk of CHD.

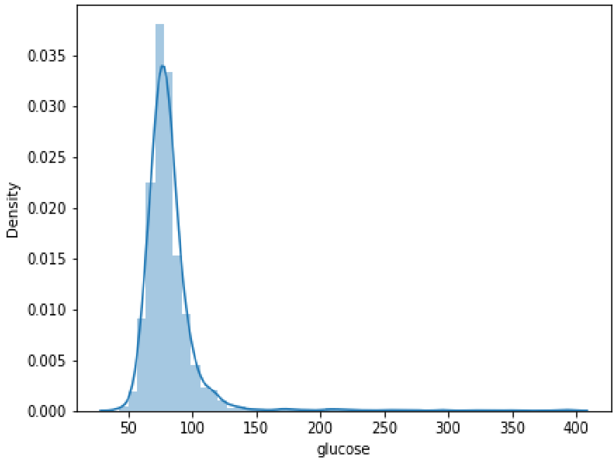
7. heartRate Vs TenYearCHD

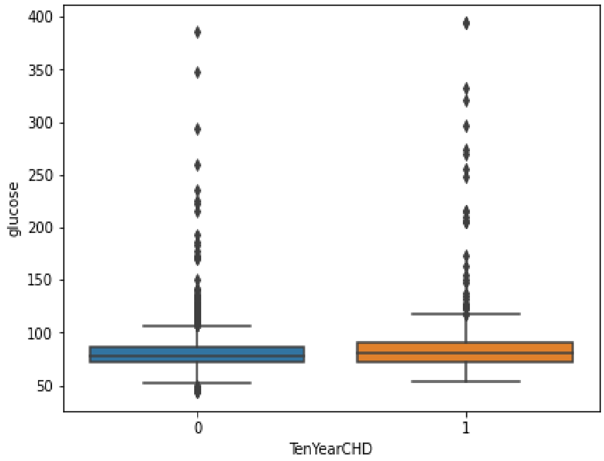




- Resting healthy heart rate for a normal human body is between 60 bpm to 100 bpm but in our dataset, it ranges between 38 bpm to 155 bpm.
- In patients with known coronary heart disease, elevated heart rate reduces.
- Surprisingly in our dataset no conclusion can be made to distinguish between the people who are at risk of CHD or not at risk as for both categories of people the heart rate varies in a similar way.

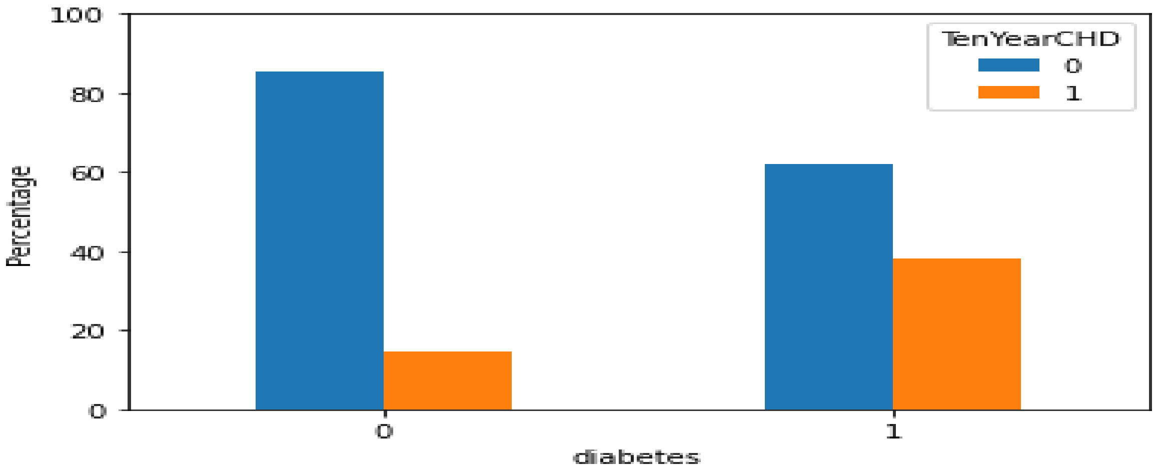
8. glucose Vs TenYearCHD





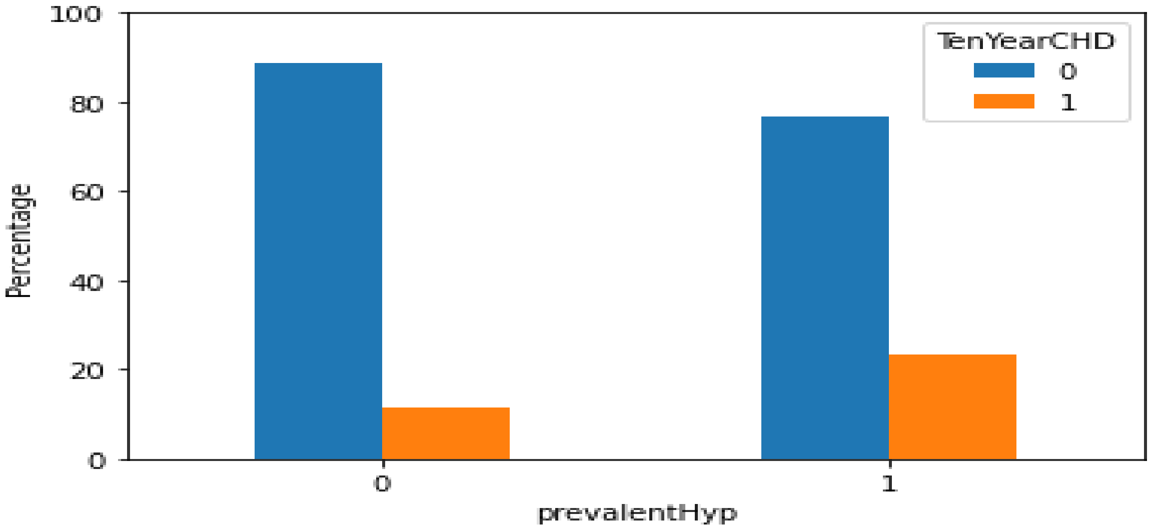
- ## **Histogram plot between categorical and target variable**

1. diabetes Vs TenYearCHD



- Most of the people in our dataset are not diabetic and very few are in number who have diabetes.
- 3303 counts of non-diabetic and diabetic people count are 87 but diabetic people have high CHD.

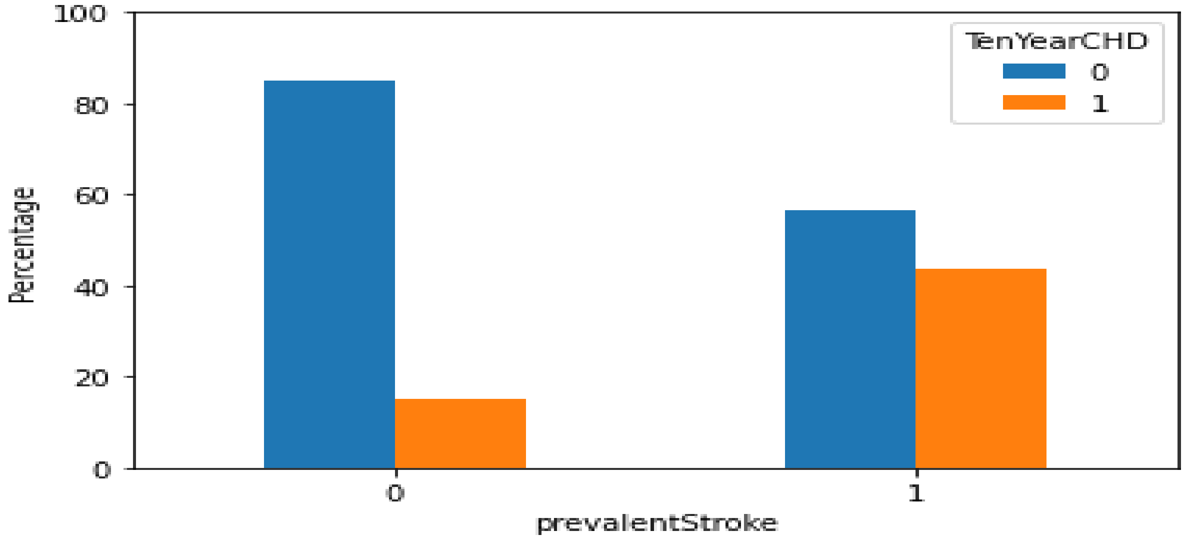
2. prevalentHyp Vs TenYearCHD



- The excess strain and resulting damage from hypertension cause the coronary arteries serving the heart to slowly.
- We can see that most of the people who have hypertension are at risk of coronary heart disease. Hypertension is one of the contributing factors to the risk of CHD.

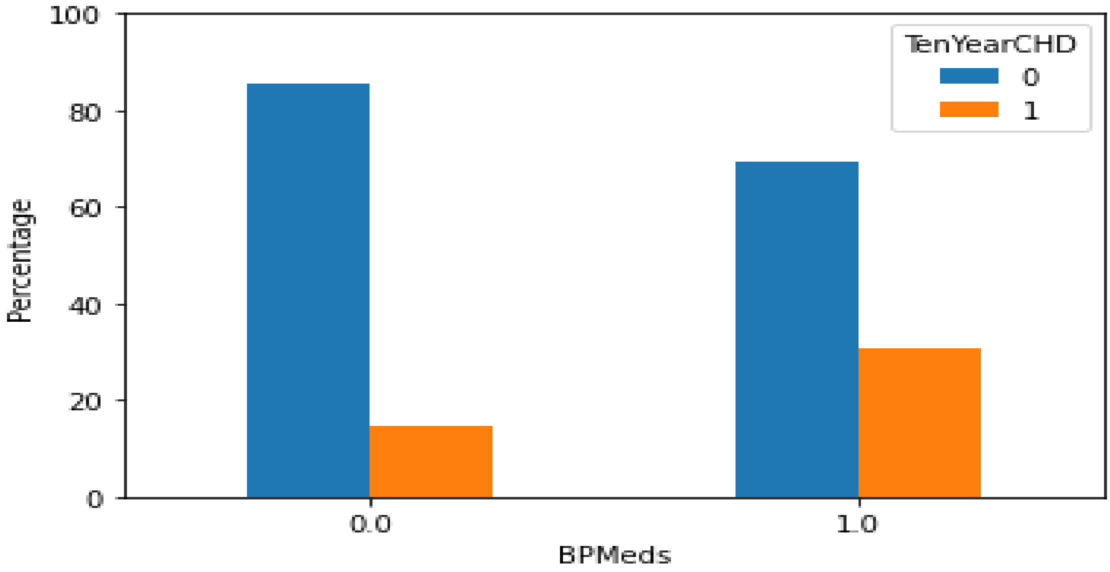
- There are cases where hypertension is not there but the risk of CHD is still present.

3. prevalentStroke Vs TenYearCHD



- Cases with positive prevalent stroke are very negligible (22) in our dataset. It would be immature to make any hard assumptions from this variable.
- 3368 People who didn't have prevalent stroke have cases where the risk of CHD is positive.

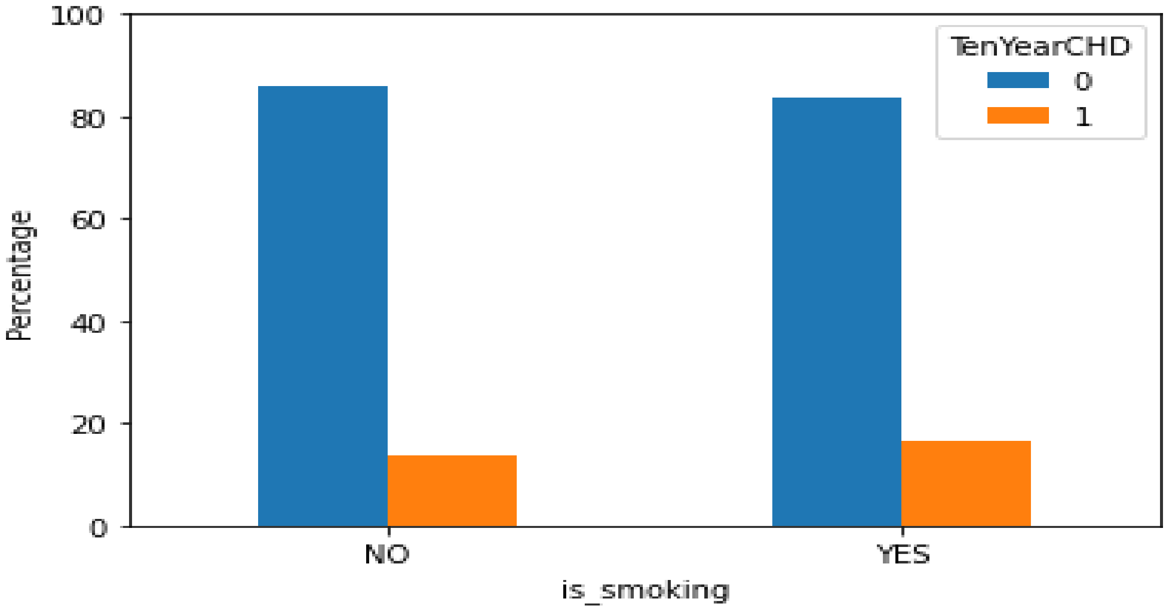
4. BPMeds Vs TenYearCHD



- People who have not on BP medication represented by 0 and it was present in large number 3283 but in this few people have suffered from cardiac disease.

- Quite clearly visible that very a smaller number of people are on BP medication.
- it is similar to prevalent Stoke.

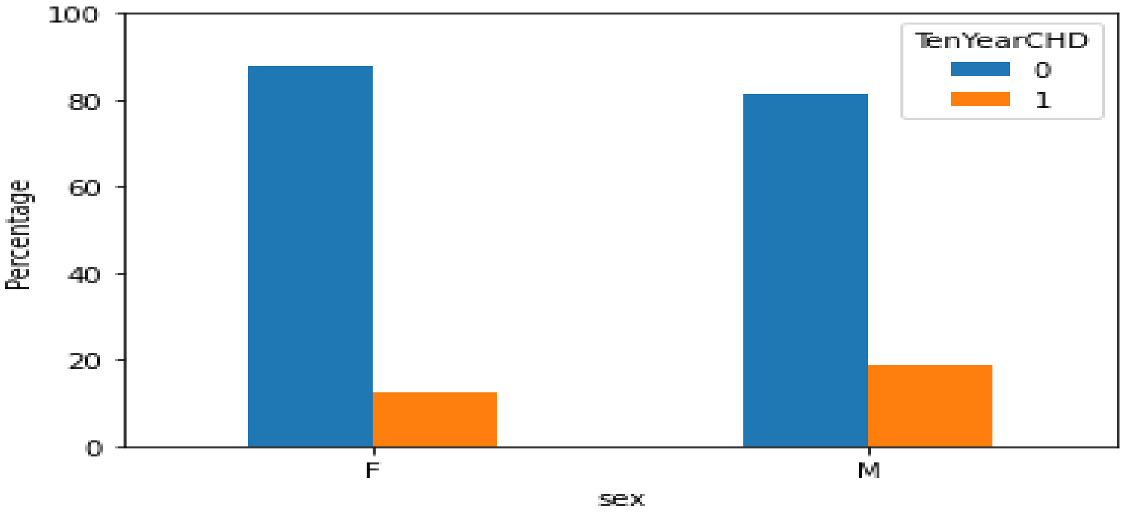
5. is_smoking Vs TenYearCHD



- Here, people who does not smoke represented by NO which is 1703 and YES represented smoker persons which is 1687.

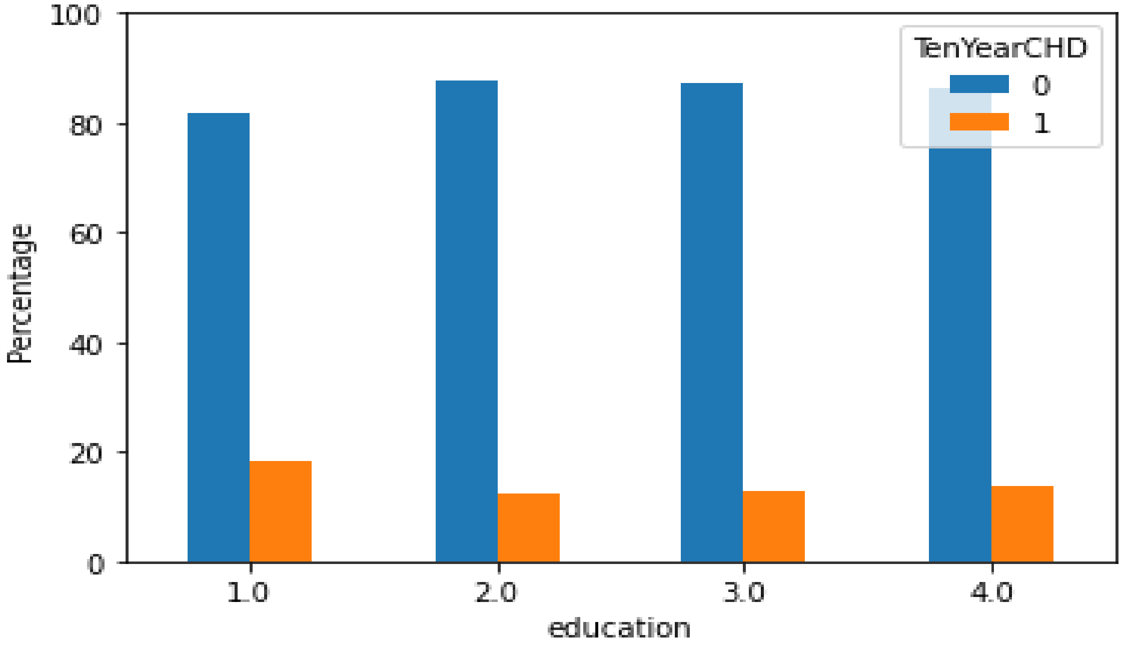
- Above graph shown that CHD is equal in both smoker and non-smoker. It is not affected on CHD.

6. SEX Vs TenYearCHD



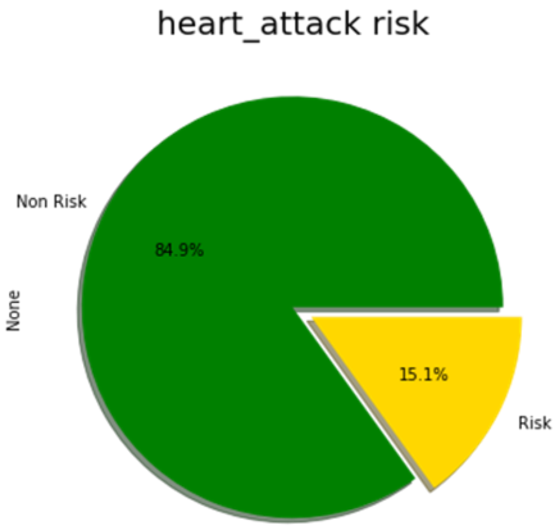
- Our dataset has close to 1467 males and 1923 females.
- The proportion of TenYearCHD is more inclined towards males even though the females are in majority in our dataset.

7. education Vs TenYearCHD



- Educated people have low risk of CHD it means educated people are aware about that and take care.

➢ Check dependent variable distribution



This is quite an imbalanced data as the number of people without the disease greatly outnumber the people who have the disease. there were 2879 respondents without CHD and 511 patients with CHD. To address this problem. I balanced the data set using the Synthetic Minority Oversampling Technique (SMOTE).

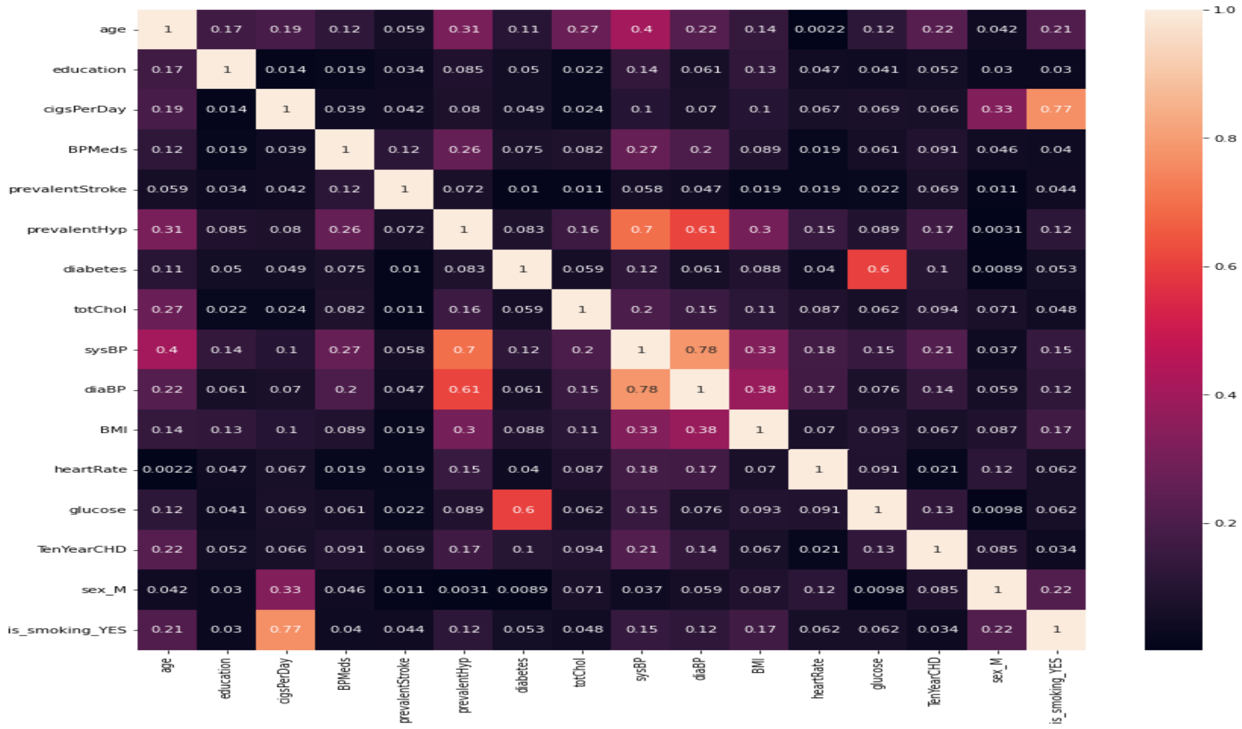
The final step was to check the correlation of the different features with the target variable and with each other as this would not only give a good estimate of the strength of the features as predictors of coronary heart disease.

It was difficult to make conclusions but based on what is observed but these are the conclusions that could be drawn:

- Slightly more males are suffering from CHD than females.
- The percentage of people who have CHD is almost equal between smokers and non-smokers.
- The percentage of people who have CHD is higher among the diabetic, and those with prevalent hypertension as compared to those who don’t have similar morbidities.
- A larger percentage of the people who have CHD are on blood pressure medication.

Another interesting trend I checked for was the distribution of the ages of the people who had CHD and the number of the sick generally increased with age with the peak being at 63 years old.
From the matrix, there are no features with a correlation of more than 0.5 with the 10-year risk of developing CHD and this shows that the features are poor predictors. However, the features with the highest correlation are age, prevalent hypertension and systolic blood pressure.
Also, there are a couple of features that are highly correlated with one another and it makes no sense using both of them to building a machine learning model. These include: blood glucose and diabetes (obviously); systolic and diastolic blood pressures; cigarette smoking and the number of cigarettes smoked per day.

CORRELATION MATRIX



From the above correlation plot we can conclude that,

● There are no features with more than 0.5 correlation with the Ten-year risk of developing CHD and this shows that the features are poor predictors. However, the features with the highest correlations are age, prevalent hypertension (prevalentHyp) and systolic blood pressure (sysBP).

● Also, there are a couple of features that are highly correlated with each other and it makes no sense to use both of them in building a machine learning model. These includes:

● Blood glucose and diabetes;

● systolic and diastolic blood pressures;

● cigarette smoking and the number of cigarettes smoked per day. Therefore, we need to carry out feature selection to pick the best feature.

MODEL DEVELOPMENT AND COMPARISON:

I used seven classification models, i.e., **Logistic Regression**, **K-Nearest Neighbours**, **Decision Trees** and **Support Vector Machine, SGD Classifier, XGBoost**. After training each model and tuning their hyper-parameters using grid search and random search cv, I evaluated and compared their performance using the following metrics:

1. **The accuracy score:** which is the ratio of the number of correct predictions to the total number of input samples. It measures the tendency of an algorithm to classify data correctly.

2. **The F1 Score**: Which is defined as the weighted harmonic mean of the test's precision and recall. By using both precision and recall it gives a more realistic measure of a test's performance.

3. **Precision** – it also called the positive predictive value, is the proportion of positive results that truly are positive.

4. **Recall** – it also called sensitivity, is the ability of a test to correctly identify positive results to get the true positive rate.

5. **The Area under the ROC Curve (AUC):** Area Under Curve (AUC) score represents the degree or measure of separability. A model with higher AUC is better at predicting True Positives and True Negatives. AUC score measures the total area underneath the ROC curve.

It is not advised to train a classifier on an imbalanced data set as it may be biased towards one class thus achieve high accuracy but another have poor sensitivity. In our case, the number of negative cases (2879) greatly exceeds the number of positive cases (511). Solve this problem by **Synthetic Minority Oversampling Technique (SMOTE)**. This is how it works:

SMOTE - This procedure can be used to create as many synthetic examples for the minority class as are required. It suggests first using random undersampling to trim the number of examples in the majority class, then use SMOTE to oversample the minority class to balance the class distribution.

After using this technique, the resultant data set was much more balanced with 2879 negative cases and 2871 positive cases.

After balancing the data set, I scaled the features to speed up the training of the classifiers and then split the data into a training and test set at a ratio of 0.8 to 0.2 respectively.

Standardization - Standardization is scaling technique used to remove outliers by making mean 0 and standard deviation 1. It is a good practice to fit the scaler on the training data and then use it to transform the testing data. This would avoid any data leakage during the model testing process.

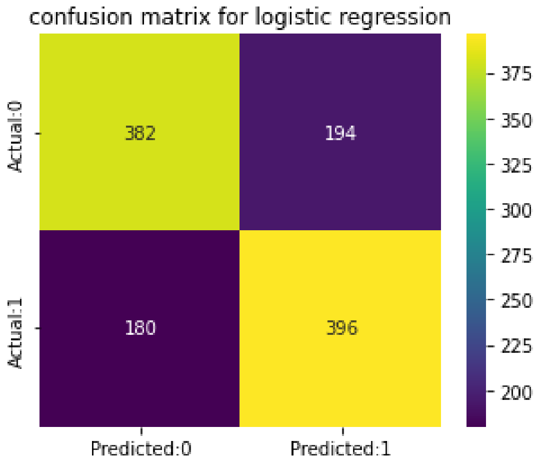
Principle component analysis (PCA) - Principle component analysis (PCA) is a technique for reducing the dimensionality of features, which have most variance but at the same time minimizing information loss.

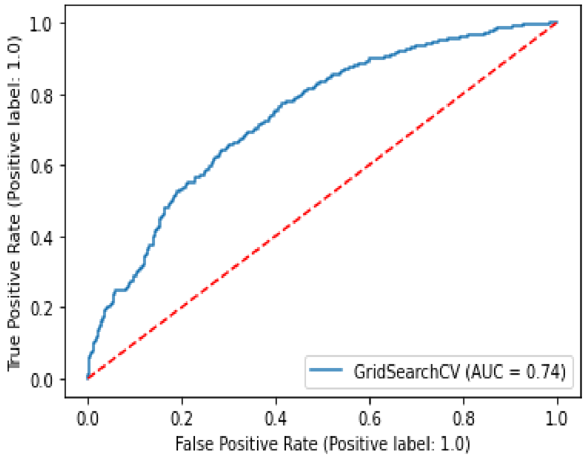
MODELLING

I used all the classification algorithm because the target variable was present in binary form.

AIM – we want less type 2 error or False Negative values and high Recall value.

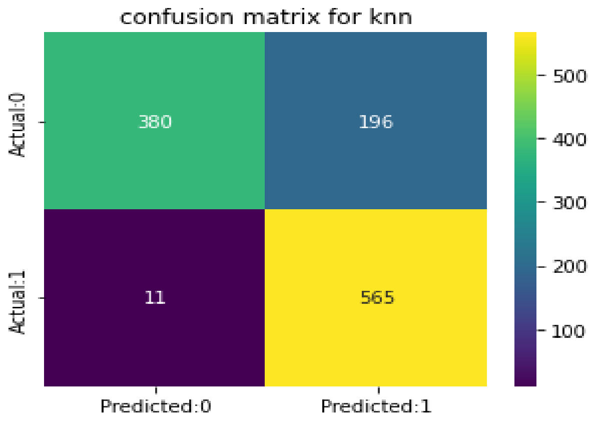
1. Logistic regression: Logistic regression is a supervised learning classification algorithm and it used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes either 1 or 0. a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

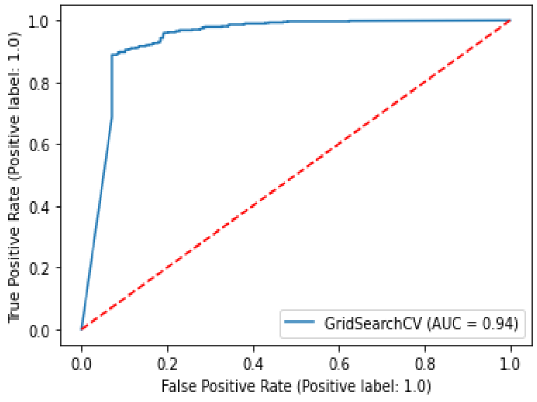




Insights – It has high type 2 error means false negative values are more and AUC score was 0.74. As you can see in the image above, the classifier becomes better as the ROC curve moves away from the red dashed line but it was not near to 1. We want less type 2 error.

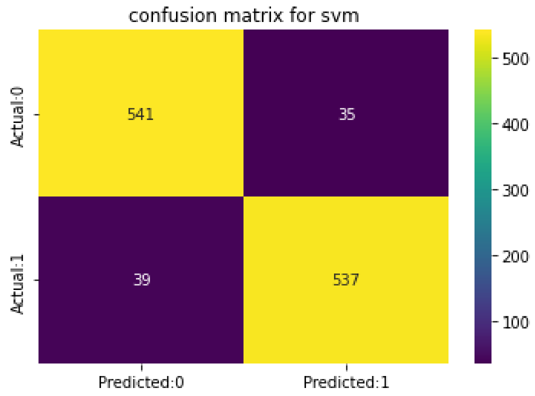
2. K-nearest Neighbor: K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification (most commonly) and regression. It is a versatile algorithm also used for imputing missing values and resampling datasets.

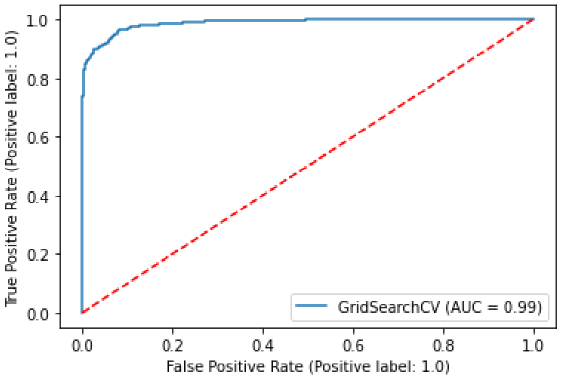




Insights - – In knn type 2 error means false negative values are very less and AUC score was 0.94 which was best. As you can see in the image above, the classifier becomes better as the ROC curve moves away from the red dashed line.

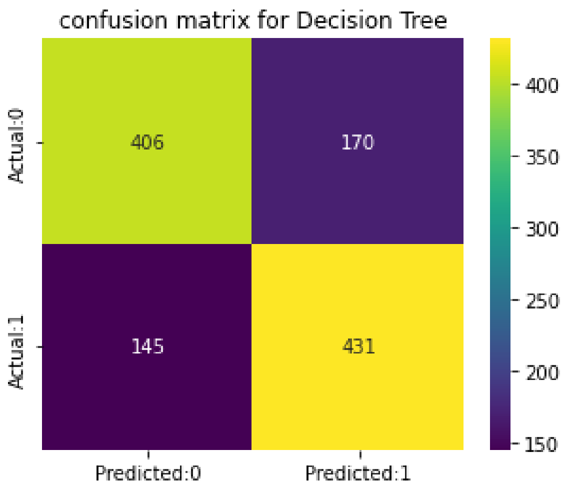
3. Support vector machine: Which is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data, the algorithm outputs an optimal hyperplane which categorizes new examples based on which side they lie in relation to it. In a two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lies on either side.

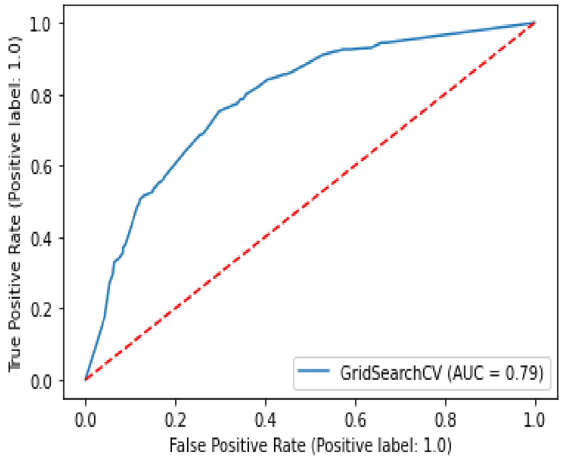




Insights - SVM has 39 false negative values and also a AUC score was 0.99 which much better than knn. ROC curve is reach to 1 means true postive is also high. SVM performs best from previous model.

4. Decision tree: Decision tree algorithm belongs to the family of supervised learning algorithms and it used to solve clarification and regression problem. The goal of using a decision tree is to create a training model that can use to predict the class or value of the target variable by learning training data.

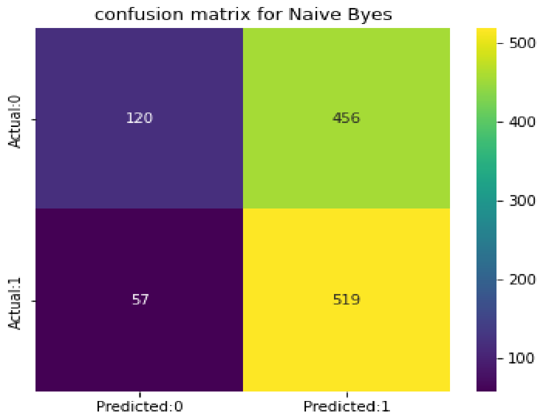


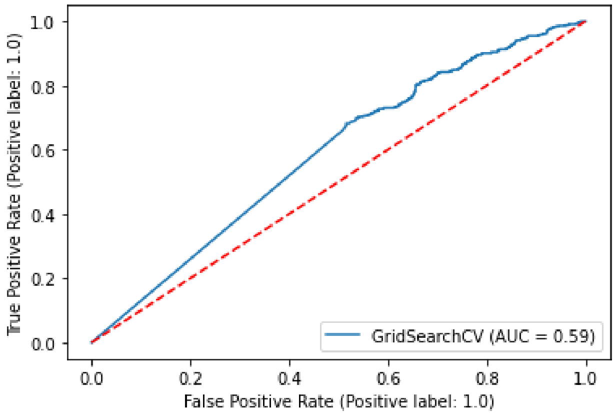


Insights – Decision Tree has lots of false negative values and AUC score was less 0.79.

6. Naive Bayes classifiers: Naive Bayes models are a group of extremely fast and simple classification algorithms that are often suitable for very high-dimensional datasets. Because they are so fast and have so few tunable parameters.

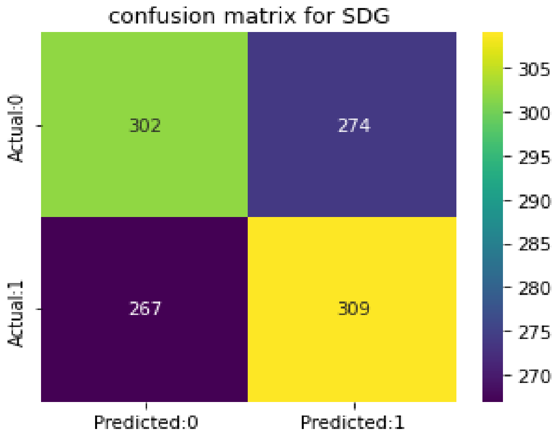
Naive Bayes classifiers are built on Bayesian classification methods. These rely on Bayes's theorem, which is an equation describing the relationship of conditional probabilities of statistical quantities.

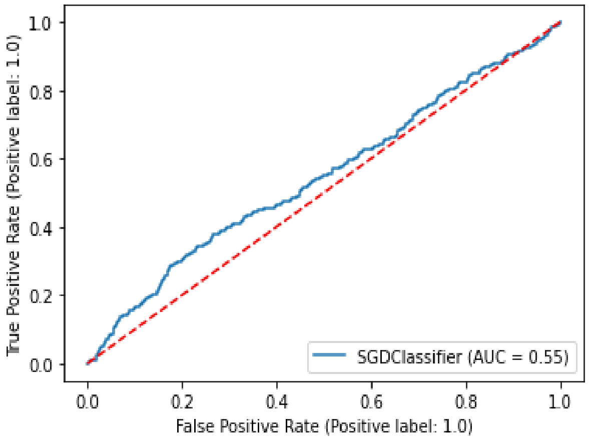




Insights- Naïve Bayes model given 57 false negative values low AUC score that was 0.59.

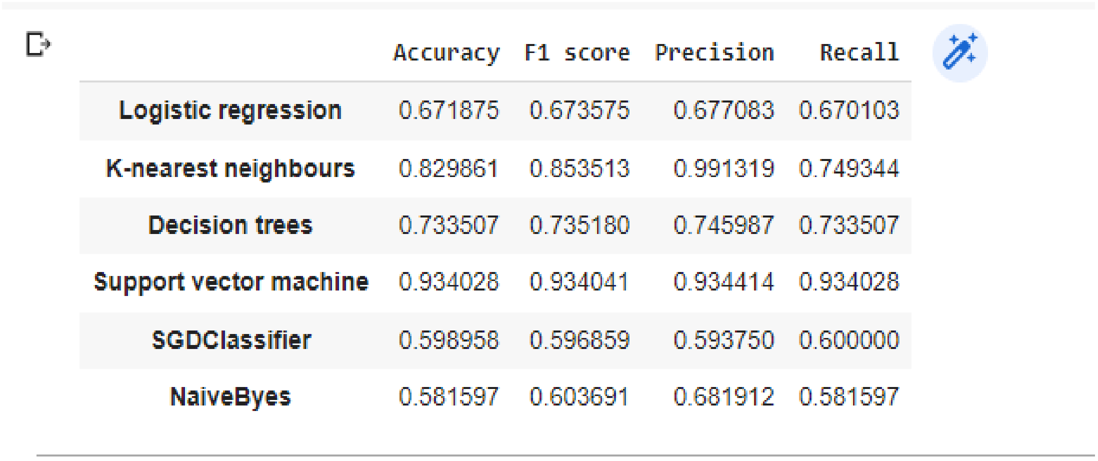
7. SGD Classifier – SGD Classifier is a linear classifier (SVM, logistic regression) optimized by the SGD. These are two different concepts. While SGD is optimization method, Logistic Regression or linear Support Vector Machine is a machine learning algorithm/model.





Insights- SGD classifier has 267 false negative values and 0.55 AUC score. SGD was performed worst on this dataset.

MATRIX COMPARISON



	Accuracy	F1 score	Precision	Recall
Logistic regression	0.671875	0.673575	0.677083	0.670103
K-nearest neighbours	0.829861	0.853513	0.991319	0.749344
Decision trees	0.733507	0.735180	0.745987	0.733507
Support vector machine	0.934028	0.934041	0.934414	0.934028
SGDClassifier	0.598958	0.596859	0.593750	0.600000
NaiveByes	0.581597	0.603691	0.681912	0.581597

Here are the results:

The support vector machine was the best performing model across all metrics. Its best parameters were a radial kernel. Its high AUC and Recall also show that the model has a low false negative rate and is thus sensitive to predict if one has a high risk of developing CHD, i.e., getting a heart attack within 10 years.

4.CONCLUSIONS

This model can then be used as a simple screening tool and all that we need to do is to input ones: age, BMI, systolic and diastolic blood pressures, heart rate and blood glucose levels after which the model can be run and it outputs a prediction.

- The number of people who have Cardiovascular heart disease is almost equal between smokers and non-smokers.

- The top features in predicting the ten-year risk of developing Cardiovascular Heart Disease are 'age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose'.

- The Support vector machine with the radial kernel is the best performing model in terms of accuracy and the F1 score and Its high AUC-score shows that it has a high true positive rate.
- Balancing the dataset by using the smote technique helped in improving the model’s sensitivity with more data (especially that of the minority class) better models can be built.