Capstone Project

Cardiovascular risk prediction

By-
Kanika Singh
Tanjul Gohar

Content:

- Problem Statement
- Introduction
- Data Dictionary
- Data Description
- Steps
- Data Cleaning
- EDA
    - Categorical Variables
    - Continuous Variables
- Correlation
- Model Performances
- Model Comparison
- Conclusion

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).
- The goal of our project is to come up with a ML model that correctly predicts 10-year risk of a patient having coronary heart disease (CHD).
- The very important metric that we want to focus on is the Recall metric since we want to minimize false negatives i.e. person with 10-year CHD risk should be flagged positive by the model.

Introduction

- Coronary heart disease is caused due to accumulation of plaque in major heart blood vessels leading to blockage of oxygen-rich blood to heart.
- It is the most common type of heart disease, killing about 300 K people in US alone every year.
- Changes in lifestyle, lack of exercise, increased stress and various other reasons have made people more vulnerable to heart diseases.
- These diseases sometimes lack symptoms and can cause sudden death without any indication.
- By understanding the main reason behind such heart diseases, we can reduce potential heart diseases and ensure a healthier life.
- This project is mainly aimed at predicting, 10 year risk of Coronary Heart Disease (CHD) given a set of variables.

Data Dictionary

The meanings of the various columns are as follows-

Demographic:

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

Behavioral:

1. **is_smoking**: whether or not the patient is a current smoker ("YES" or "NO")
2. **Cigs Per Day**: the number of cigarettes that the person smoked on average in one day.(continuous as one can have any number of cigarettes, even half a cigarette.)

Medical(history):

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

Medical(current):

1. **Tot Chol:** total cholesterol level (Continuous)
2. **Sys BP:** systolic blood pressure (Continuous)
3. **Dia BP:** diastolic blood pressure (Continuous)
4. **BMI:** Body Mass Index (Continuous)
5. **Heart Rate:** heart rate (Continuous)
6. **Glucose:** glucose level (Continuous)

The Data

- The dataset is from an ongoing cardiovascular study on the residents of the town of Framingham, Massachusetts. It has 3390 rows and 17 columns.

- The attributes are divided into various sections such as demographic, behavioral, past medical records and current medical records.

- Some of the variables is categorical in nature whereas other variables are continuous in nature.

- The dependent variable in this dataset is the Ten Year CHD column, which contains binary values.

Steps

- The project has mainly been divided into 5 major steps, each contributing significantly in achieving the goal of predictions.

The 7 steps are as follows :-

1. Data Cleaning
2. Exploratory Data Analysis (EDA)
3. Data Transformation
4. Model Building and Evaluation
5. Hyperparameter Tuning
6. Metrics Comparison
7. Conclusion

Data Cleaning

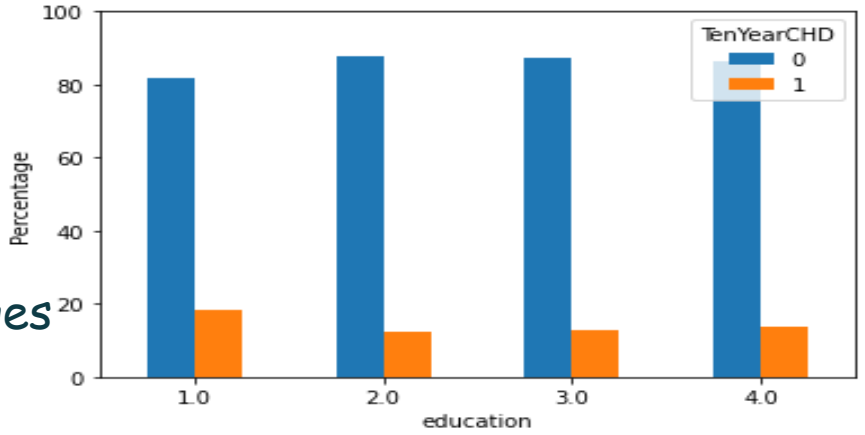
- This the first section of the project. After importing necessary libraries and the data itself, it is a must to clean the data.
- Null values in the columns were filled with the mode and median for categorical and continuous variables respectively. Median was chosen due to outliers.
- Columns of smoking and cigarettes per day did not match in some cases, and these were corrected.
- Outliers in this dataset do exits, but after analysis it was found that removal of outliers would lead to a high deduction of cases with the risk of CHD. Naturally, people with extreme values are more prone to heart diseases. Hence, removal of outliers was not viable.

EDA

Categorical Variables-

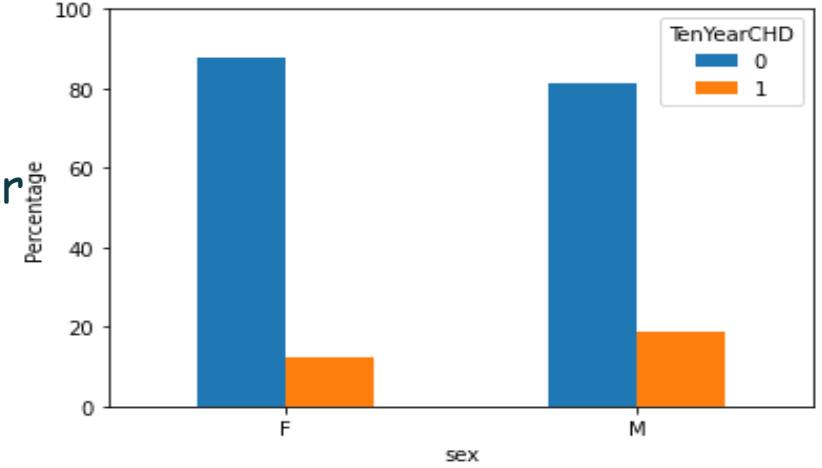
Education:

The education rate was quite same
in all attributes with respect to percentages



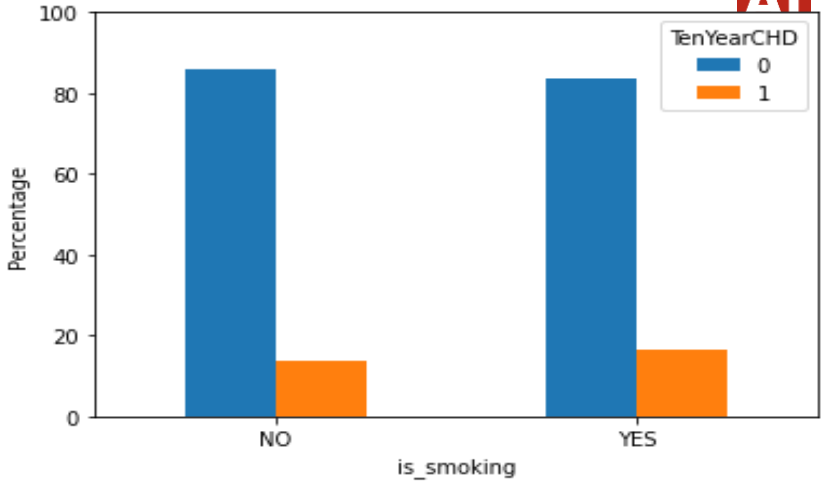
Sex:

From this graph we conclude that Higher fraction of males are prone to cardiovascular diseases.



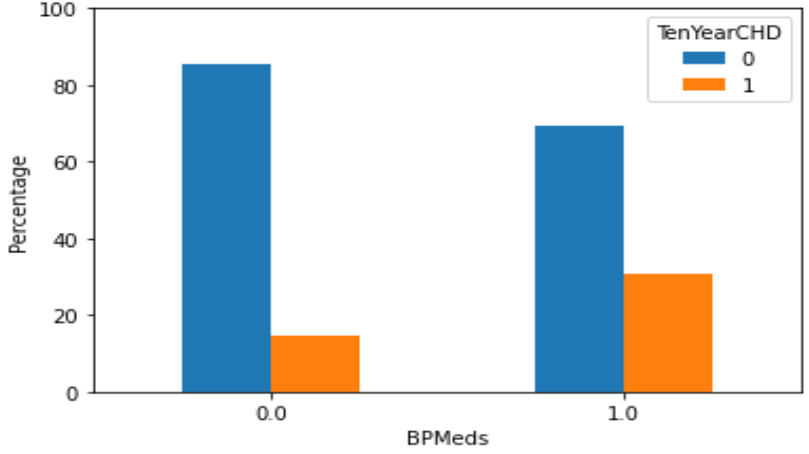
Smoking:

- Statistically smoking has no impact on 10-year risk of CHD.
- checking the dependency, the categorical variable (smoking)was not dependent.



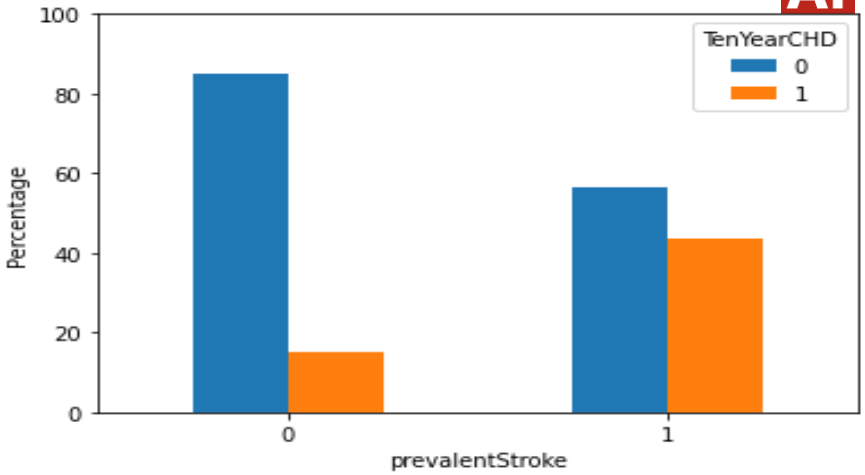
BPMeds:

- People who take blood pressure meds have a higher chance of having CHD.
- Checking the dependency, the categorical variable (Blood Pressure) was dependent.



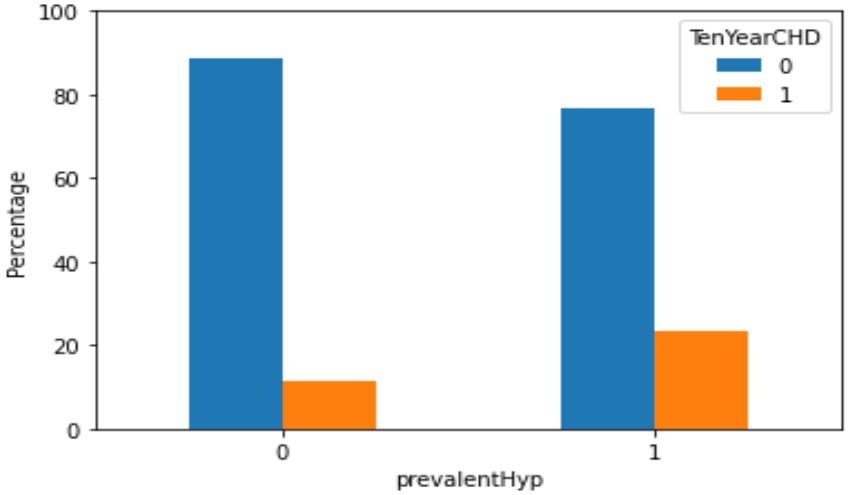
Prevalent stroke:

- Higher percentage of patient with prevalent stroke symptoms have a 10-year risk of CHD.
- Upon checking the dependency, stroke was dependent.



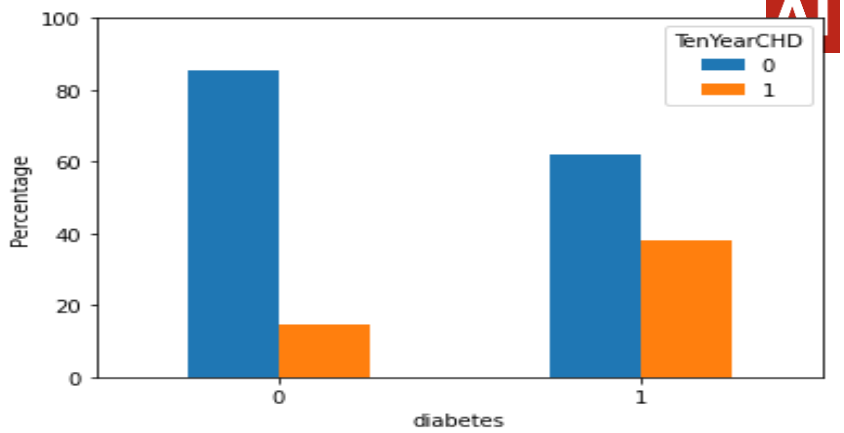
Prevalent Hypertension:

- Hypertensive patient was higher at risk of CHD.
- Hypertensive was also a dependent variable



Diabetes:

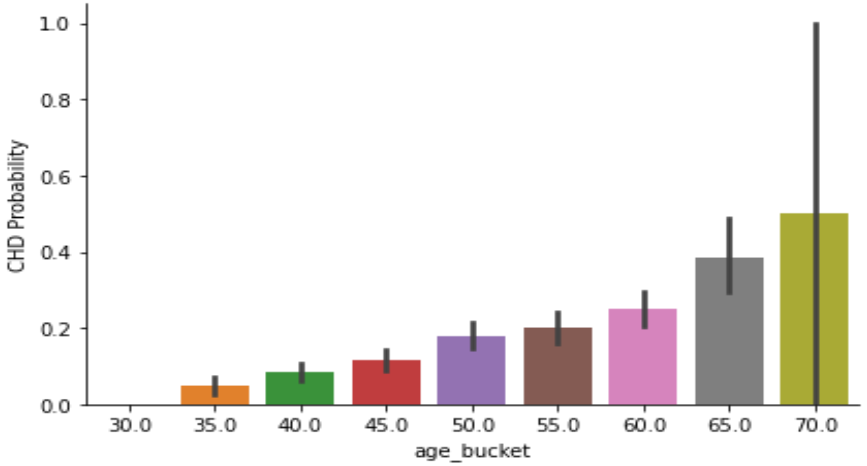
Diabetes patient tend to have a higher 10-year risk of CHD.



Continuous Variables –

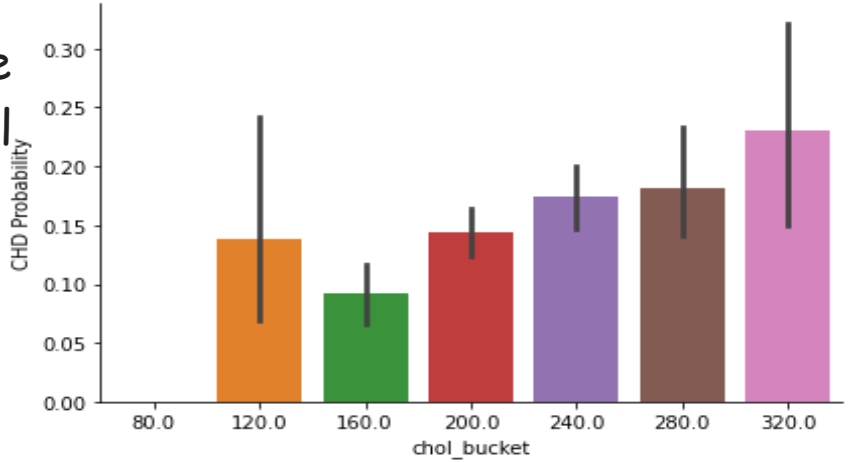
Age:

By bucketing the age features from this graph, we conclude that the positive cases CHD seems to be more prevalent in older people.



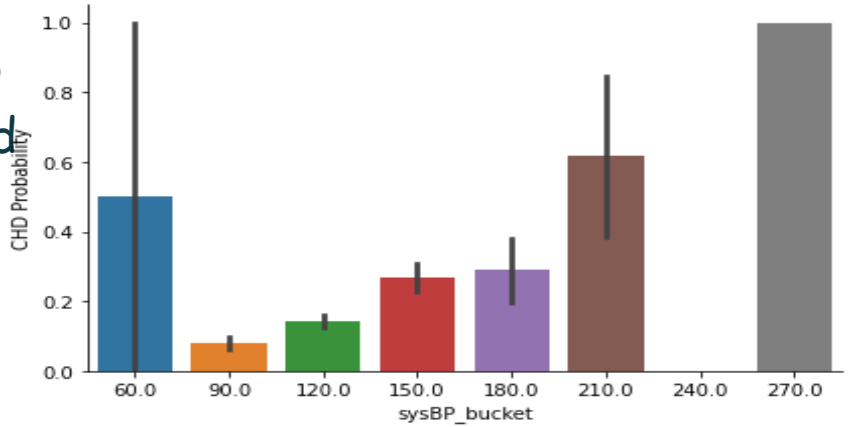
Total cholesterol level:

the median is slightly higher for the positive cases, which means people whose cholesterol level high were higher at risk of CHD.



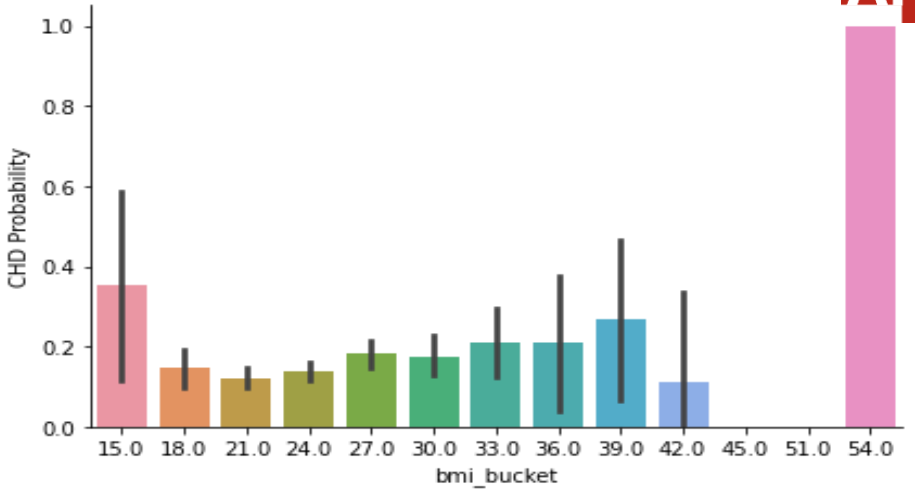
Sys and Dia Blood pressure:

We see a slightly positive inclination of CHD risk towards high systolic and diastolic blood pressure.



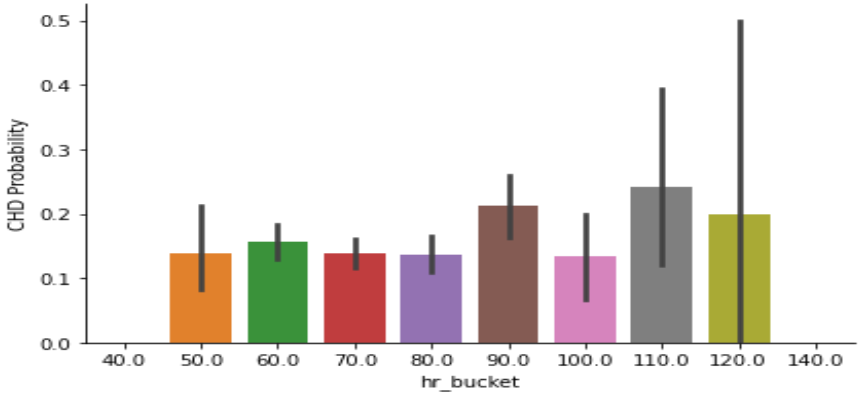
Body Mass Index:

From the BMI bucketing, we see those who have high BMI were at a risk of CHD.

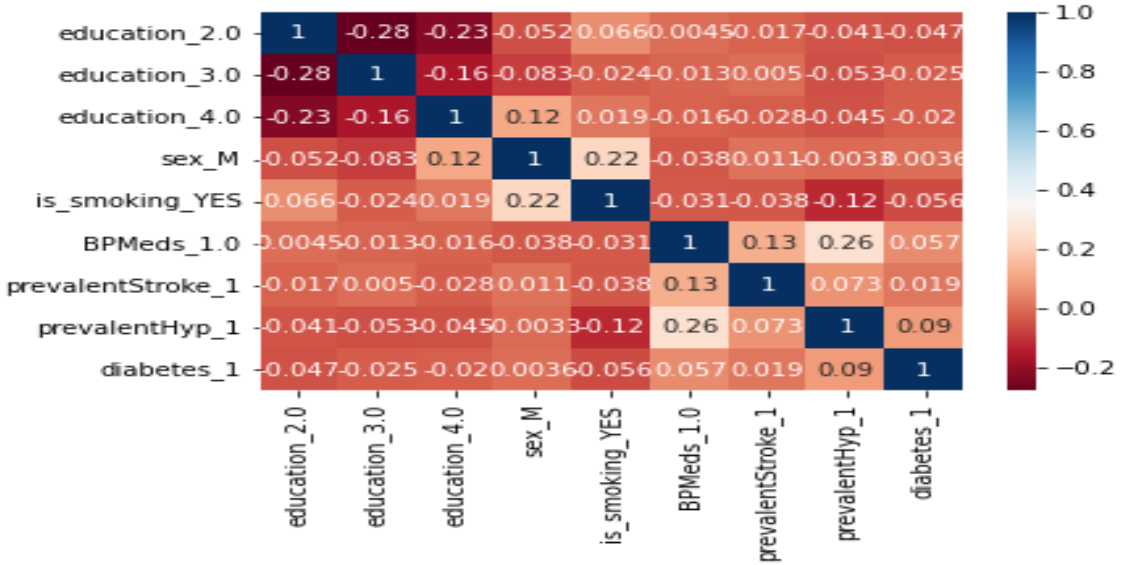


Heart Rate:

By bucketing the heart rate feature we conclude that ten-year risk of CHD was more common in between the 90 to 110 rates.



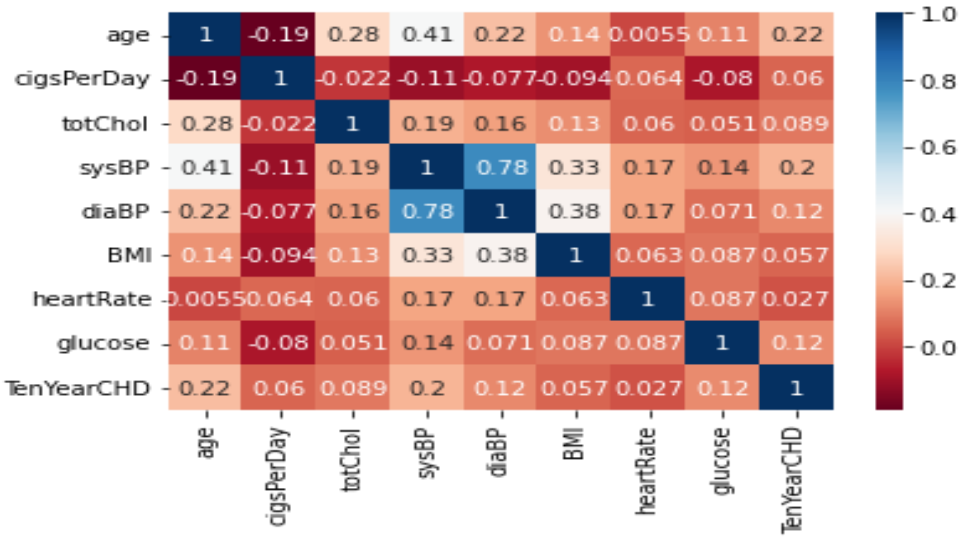
Correlation of all categorical variables



Series	variables	VIF
1	education_2.0	0.0067901
2	education_3.0	0.4802802
3	education_4.0	0.2136053
4	sex_M	0.1984034
5	is_smoking_YES	0.0864625
6	BPMeds_1.0	1.0259727
7	prevalentStroke_1	1.0259727
8	prevalentHyp_1	0.1277658
9	diabetes_1	1.030739

Based on VIF score we can clearly remove Education feature.

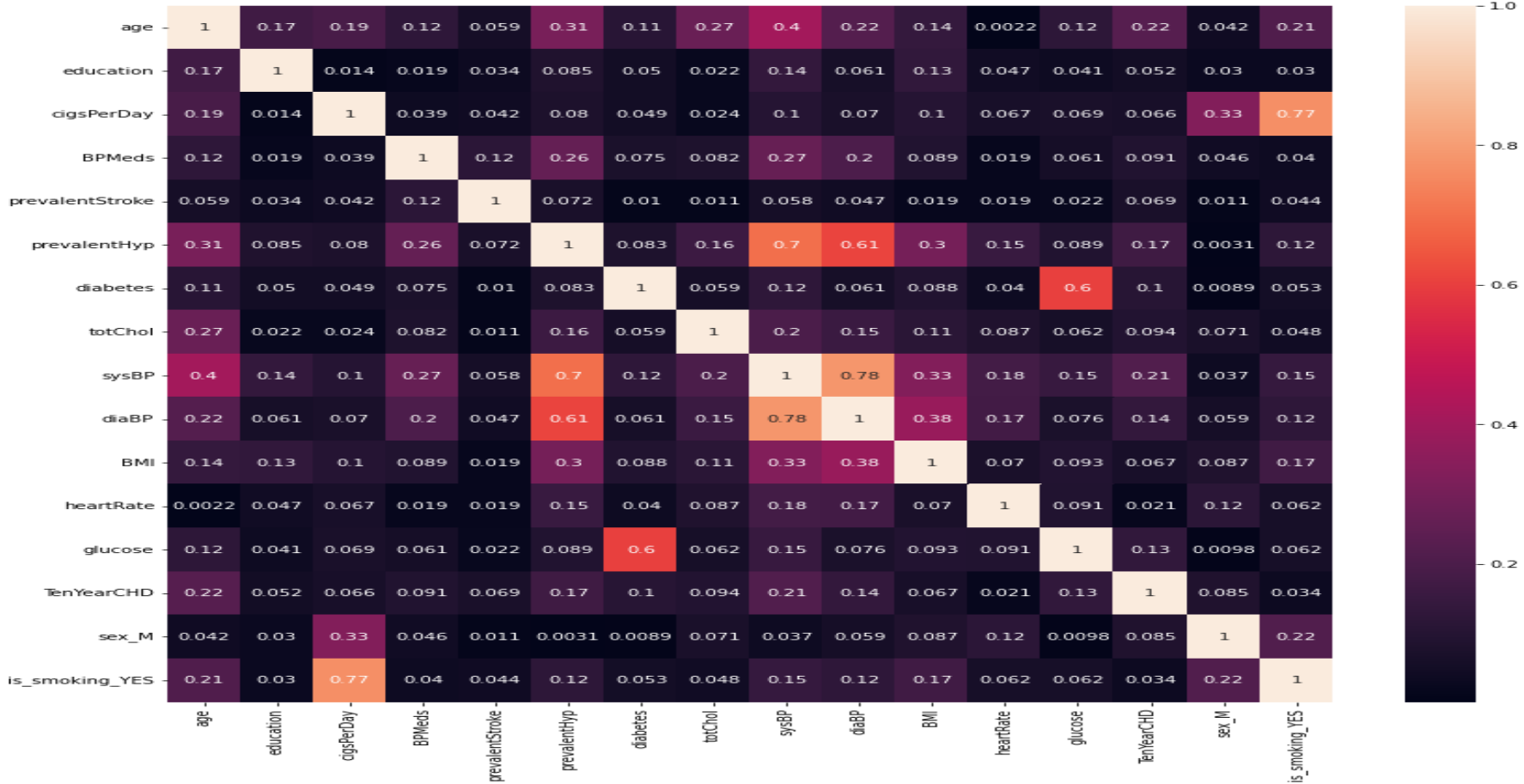
Correlation of all continuous variables



Based on correlation metrics and VIF score
We can safely remove diastolic BP feature

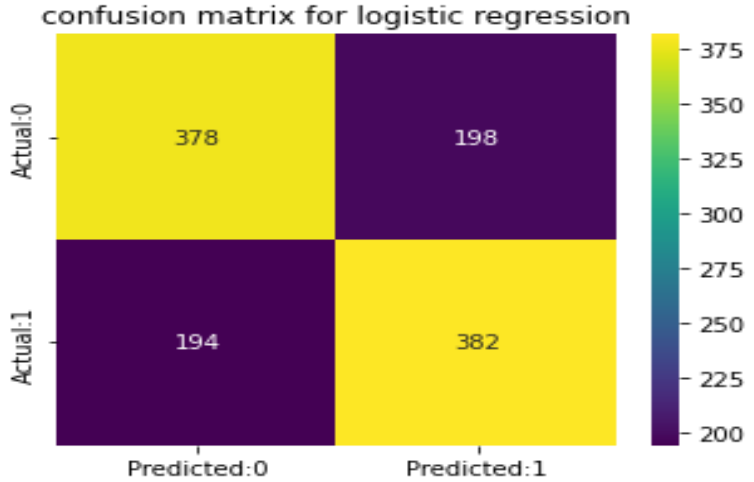
Series	variables	VIF
1	age	41.049805
2	cigsPerDay	1.621425
3	totChol	34.108171
4	sysBP	49.471280
5	BMI	39.070687
6	heartRate	33.297511
7	glucose	13.291682

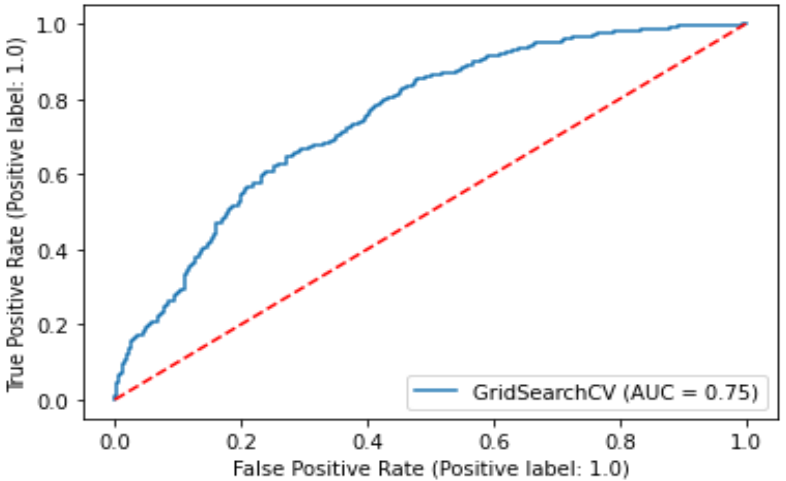
Correlation of all the variables



Model Performance

Logistic Regression:





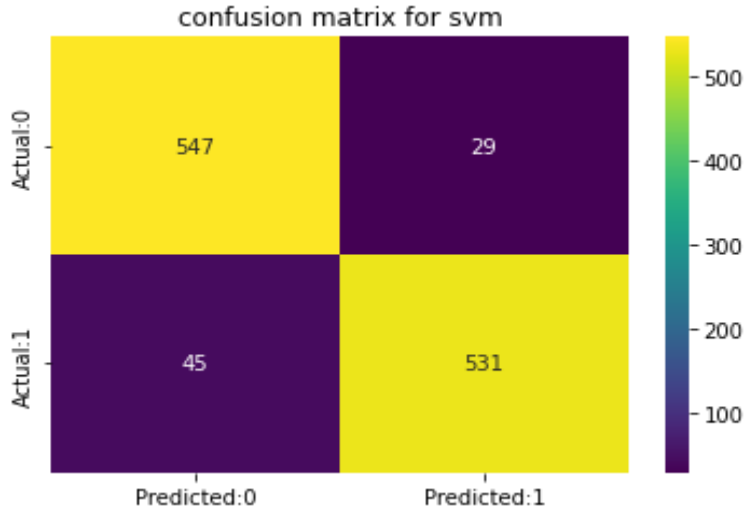
Accuracy: 0.65972222222222

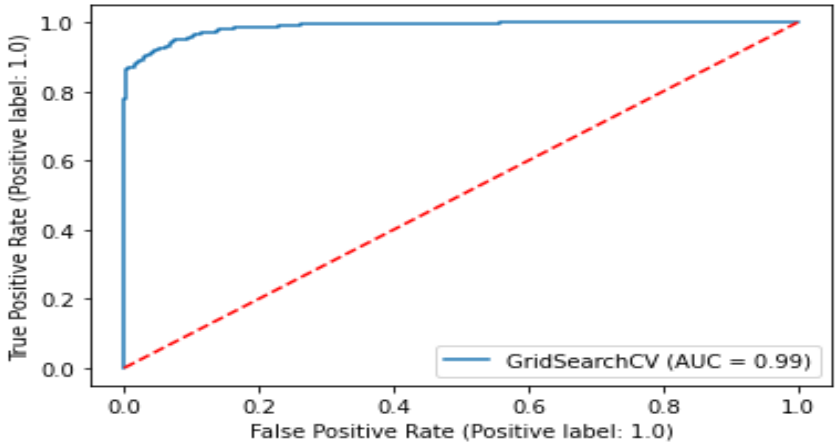
Precision: 0.663194444267

Recall: 0.658620908937605397

F1Score: 0.6608992138579983

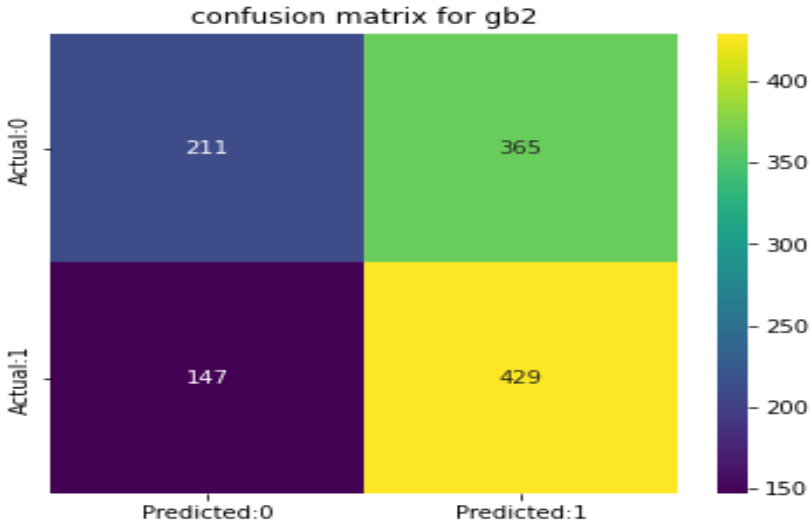
Support Vector Machine (SVM)

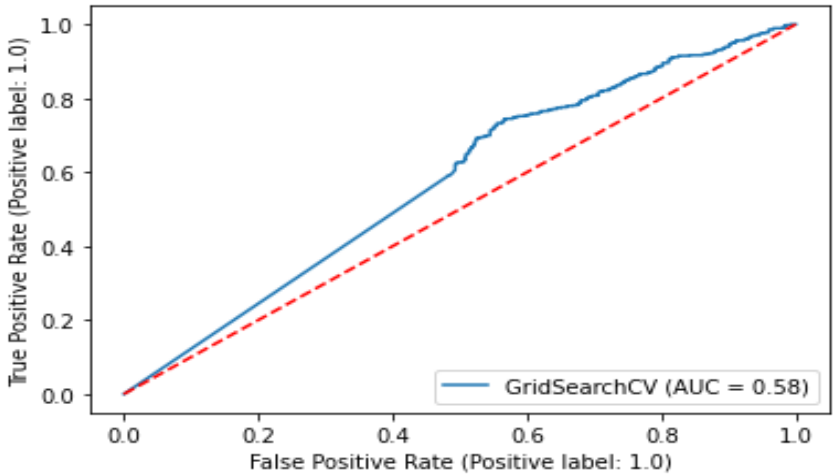




- Accuracy: 0.9357638888888888
- Precision: 0.9364932966820988
- Recall: 0.9357638888888888
- F1 Score: 0.9357873245960077

Naïve Bayes Classifier:





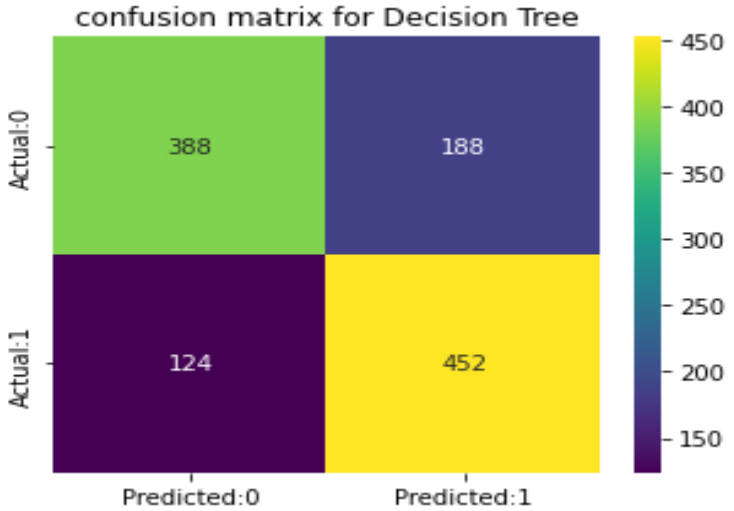
Accuracy: 0.555552777777778

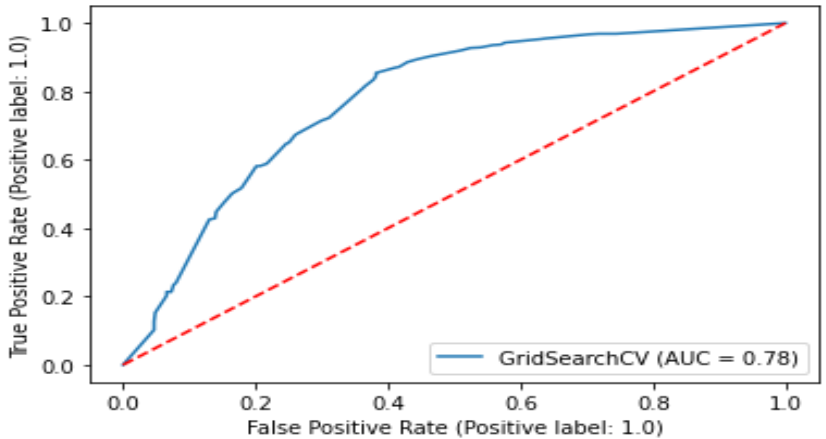
Precision: 0.62717615116705247

Recall: 0.5512152777777778

F1 Score: 0.5720609634549716 4

Decision Tree:





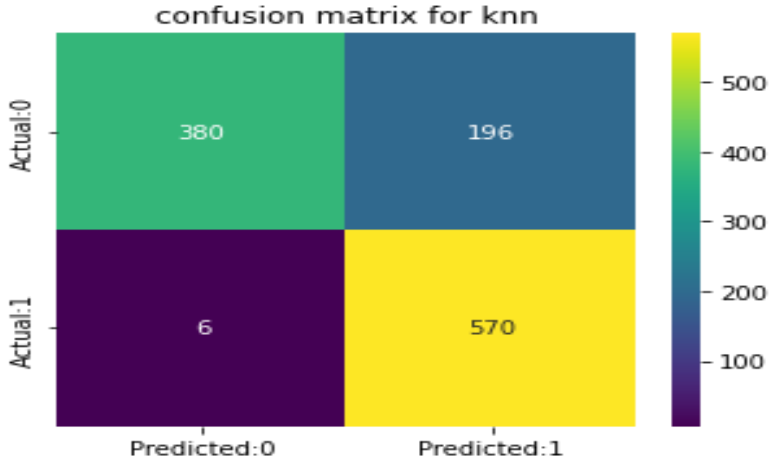
Accuracy: 0.729166333333334

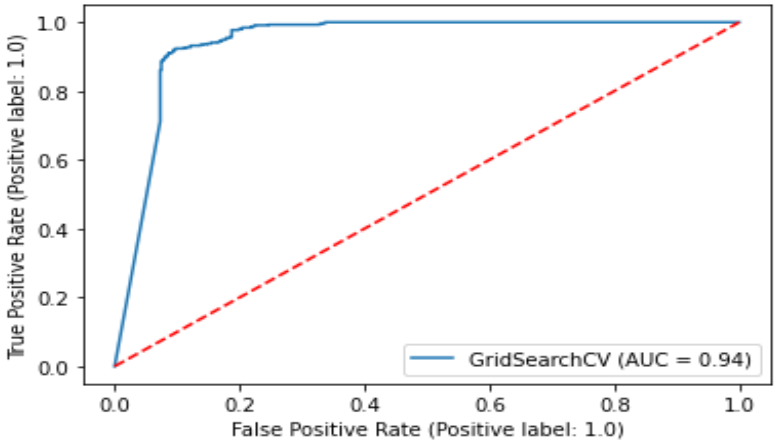
Precision: 0.7353387422839505

Recall: 0.7291658333333334

F1 Score: 0.7300541561891342

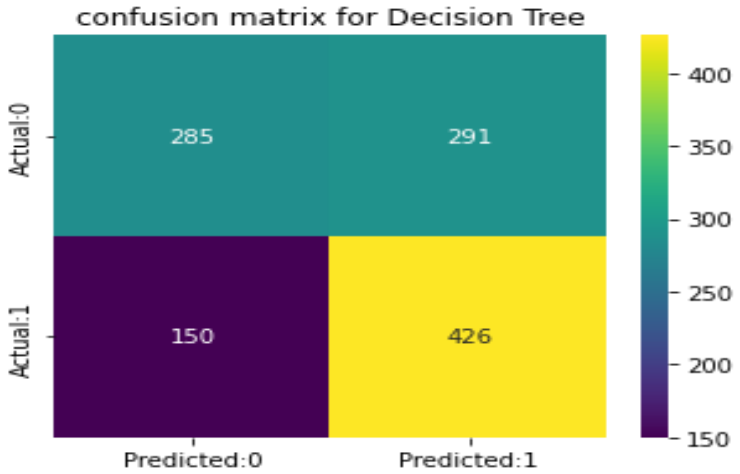
KNN:

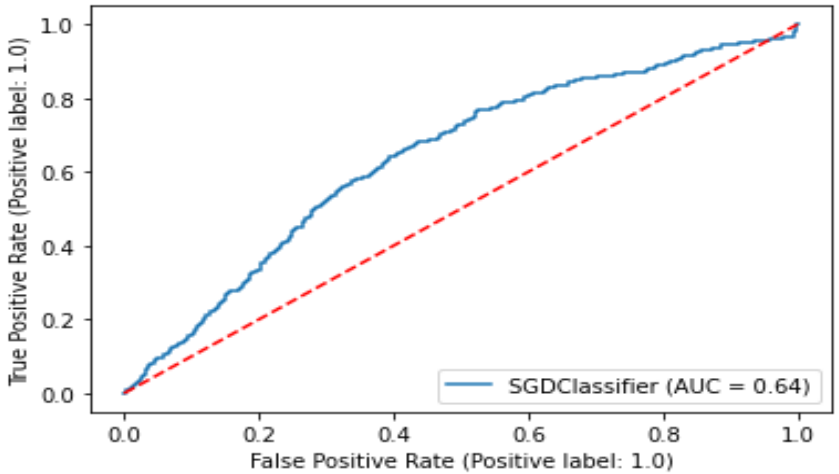




Accuracy : 0.8246527777777778
Precision : 0.9895833333333334
Recall : 0.7441253263707572
F1 Score : 0.8494783904619971

Stochastic Gradient Descent





Accuracy : 0.6171875
Precision : 0.7395833333333334
Recall : 0.5941422594142259
F1 Score : 0.6589327146171694

Model Comparison

Index	Accuracy	F1 score	Precision	Recall
Support vector machine	0.9357638888888888	0.9357762825047699	0.9361496913580246	0.9357638888888888
SGDClassifier	0.6171875	0.6589327146171694	0.7395833333333334	0.5941422594142259
NaiveByes	0.5555555555555556	0.5720623607398956	0.627176167052469	0.5555555555555556
Logistic regression	0.6597222222222222	0.6608996539792388	0.6631944444444444	0.6586206896551724
K-nearest neighbours	0.8246527777777778	0.8494783904619971	0.9895833333333334	0.7441253263707572
Decision trees	0.7291666666666666	0.7300051599587205	0.7353395061728395	0.7291666666666666

Support vector machine gives highest Accuracy, Recall, Precision and AUC score.

Conclusion

- 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 models' sensitivity.
- With more data(especially that of the minority class) better models can be built.