Capstone Project-3

Cardiovascular Risk Prediction

By-Prabir Debnath

Points of Discussion

1. The Problem Statement
2. Concept of ML Model
3. Summary of the Experience Set
4. Exploration and Pre-processing of Data
5. Building and Evaluation of Model-1
6. Building and Evaluation of Model-2
7. Building and Evaluation of Model-3
8. Final Conclusion

The Problem Statement

We are provided with a labeled dataset on the details of patients with or without cardiovascular disease. The task is to explore and analyze the data and to build a classification model for 10 Years Coronary Heart Disease prediction.

Concept of ML Model

The performance of a machine learning model depends on three factors:

i. Quality of Data

(cleaner experiences for better learning)

ii. Quantity of Data

(more experiences for better learning)

iii. Quality of Model

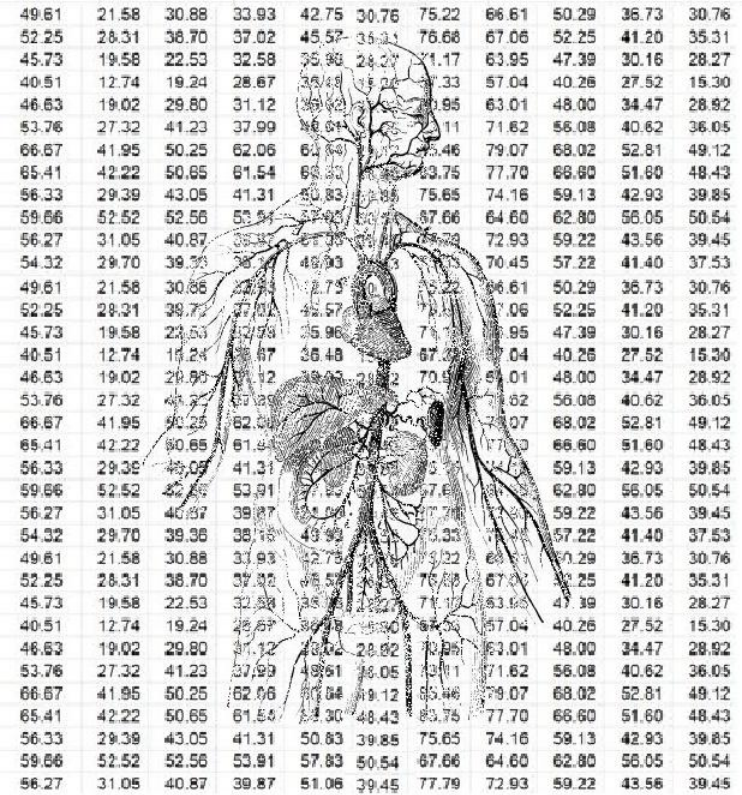
(right model and right hyperparameters for better learning)

Summary of the Experience Set

Here, the dataset has 3,390 rows, which means 3,390 experiences about patients with or without cardiovascular disease and

It has 17 columns, which means each experience is observed along 17 features or dimensions.

Let’s Decode the Experiences !



Exploration and Pre-processing of Data

We have done the exploration and pre-processing in seven steps to transform raw data into quality data for our ml model.

1. Connection with the Data
2. First Feelings of the Data
3. Deeper Understanding of the Data
4. Cleaning the Data
5. Treating Anomalies in the Data
6. Final Feature Selection from the Data
7. Preparation of Input and Output Data

Cleaning the Data

> We have handled all null values in 'cigsPerDay', 'totChol', 'BMI', 'heartRate', 'glucose', 'education', and 'BPMeds' columns with imputation. Thus there is no loss of data.

> We have encoded ‘sex’ column with two categories: F: 0, M: 1

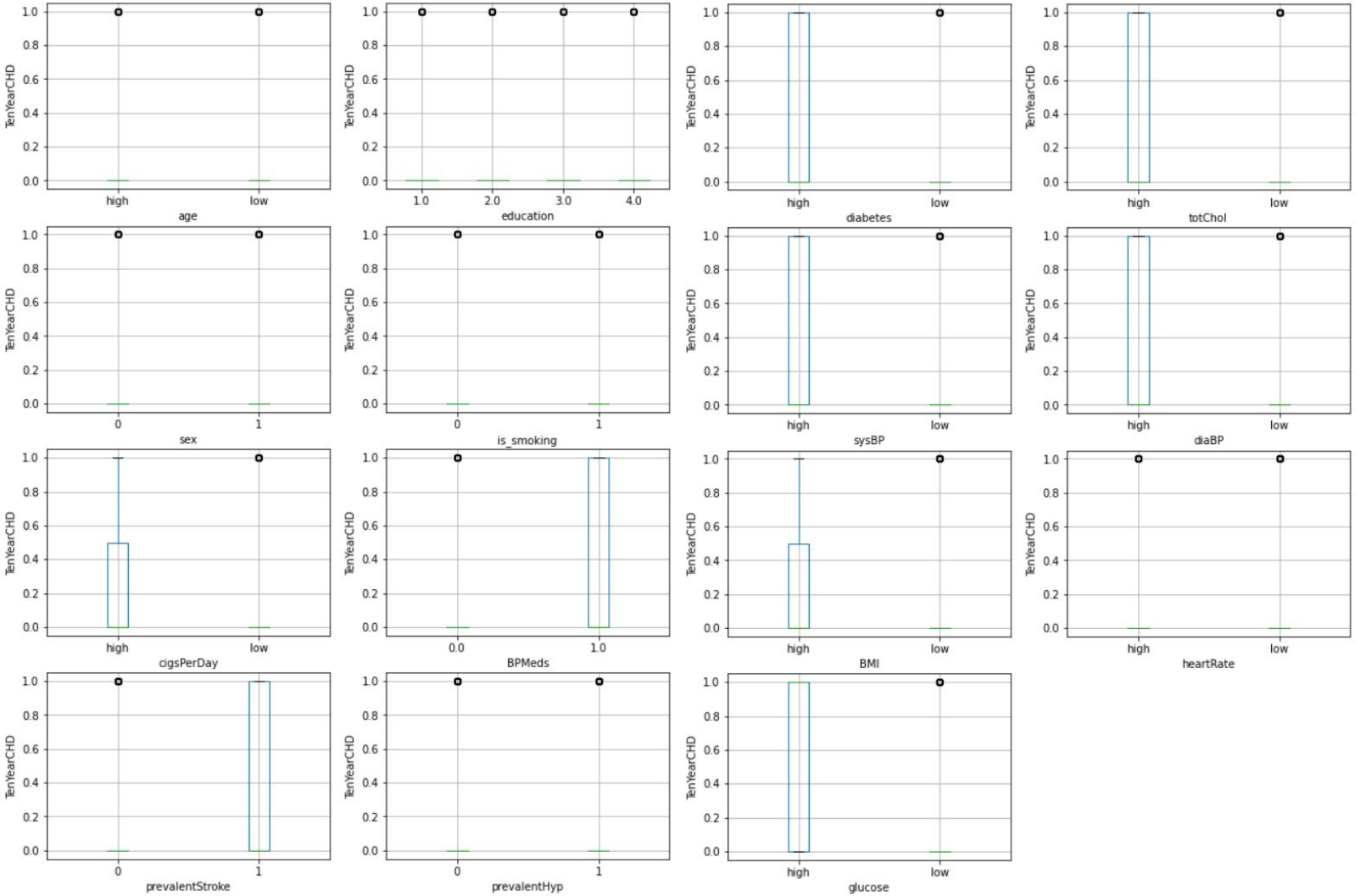
> We have encoded ‘is_smoking’ column with two categories: NO:0, YES:1

Treating Anomalies in the Data

In our dataset, for most of the features class 1 targets are outliers. Thus we need more experience with class 1 targets to bring a balance in prediction.

In simple words, with the available experience set, our model will be expert in predicting the features for which there will be no heart disease.

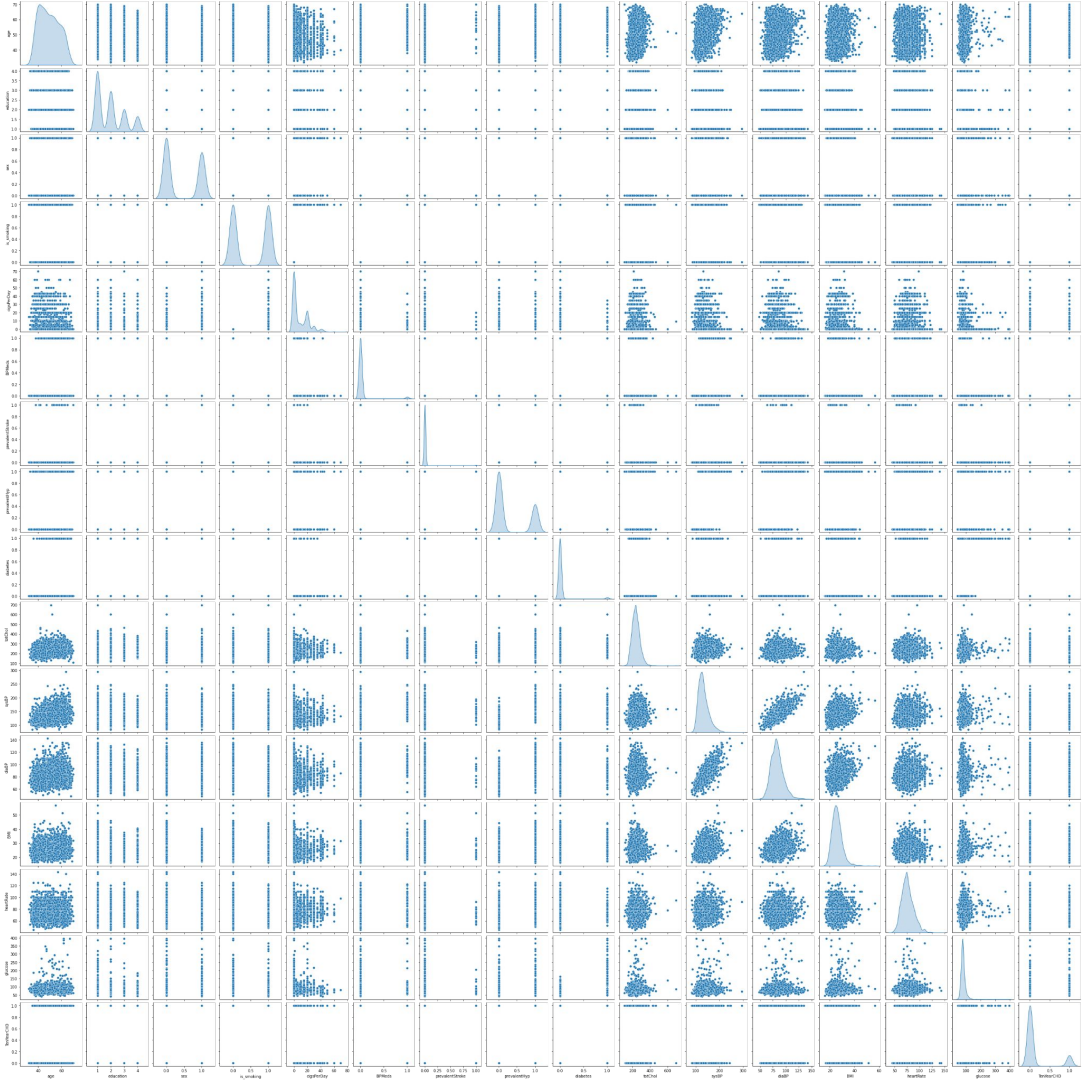
Let's check the boxplot!


TenYearCHD
1.0
0.8
0.6
0.4
0.2
0.0
high
low
age
1.0
2.0
3.0
4.0
education
diabetes
totChol
0
1
sex
is_smoking
sysBP
diaBP
cigsPerDay
0.0
BPMeds
BMI
heartRate
prevalentStroke
prevalentHyp
glucose

Overall Feature Understanding

Here, the distribution of 'cigsPerDay', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate' and 'glucose' are positively skewed. Thus we have done log transformation on these features to normalize their distribution.

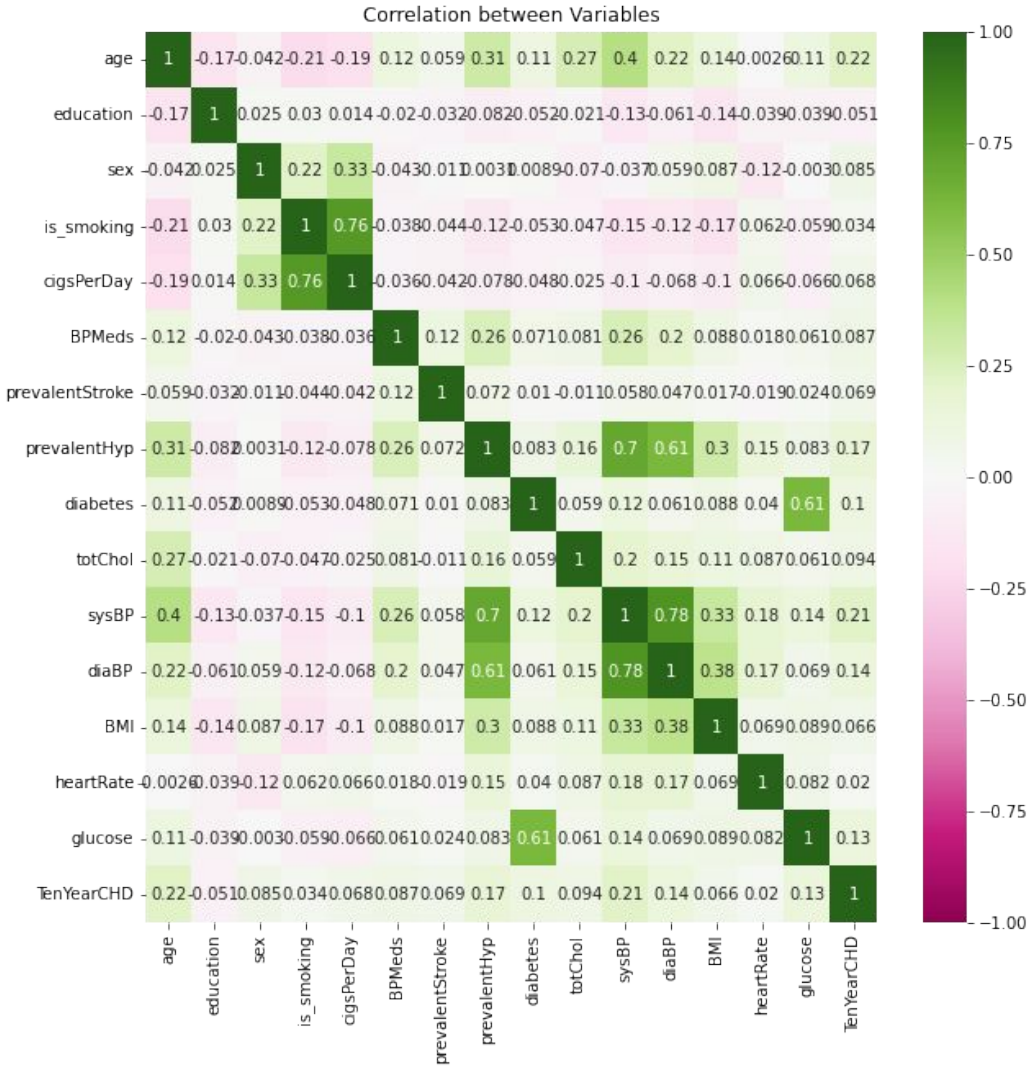
Let's check the pairplot!


age
education
sex
is_smoking
cigsPerDay
BPMeds
prevalentStroke
prevalentHyp
diabetes
totChol
sysBP
diaBP
BMI
heartRate
glucose
TenYearCHD

Looking for Truly Independent Features

We have removed ‘sysBP’, ‘diaBP’, ‘BMI’, ‘Age’, ‘heartRate’, ‘totChol’, ‘is_smoking’ and ‘Diabetes’ in sequence from our dataset to bring all the VIF values below 10. Thus all our input variables became truly independent.

Let's check the heatmap!



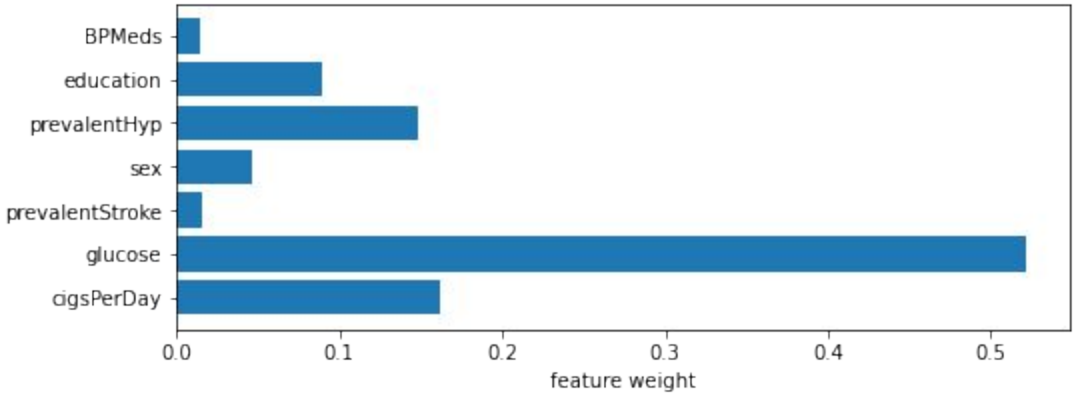
	variables	VIF
0	education	3.655103
1	sex	1.957236
2	cigsPerDay	1.733063
3	BPMeds	1.117896
4	prevalentStroke	1.024059
5	prevalentHyp	1.561774
6	glucose	4.497470

Let’s Train the Models !

Building and Evaluation of Model-1

Final Random Forest Model:

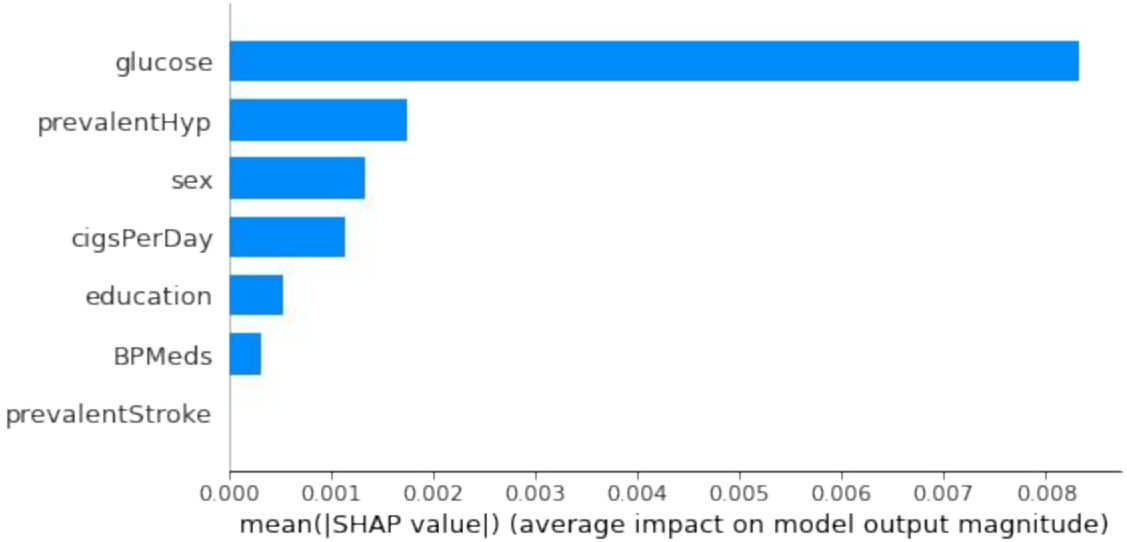
After cross validation and hyperparameter tuning, the best parameters are {'max_depth': 25, 'max_features': 'auto', 'min_samples_leaf': 5, 'min_samples_split': 15, 'n_estimators': 100} (test accuracy is 84% and variance in prediction is 2%)



Building and Evaluation of Model-2

Final KNN Model:

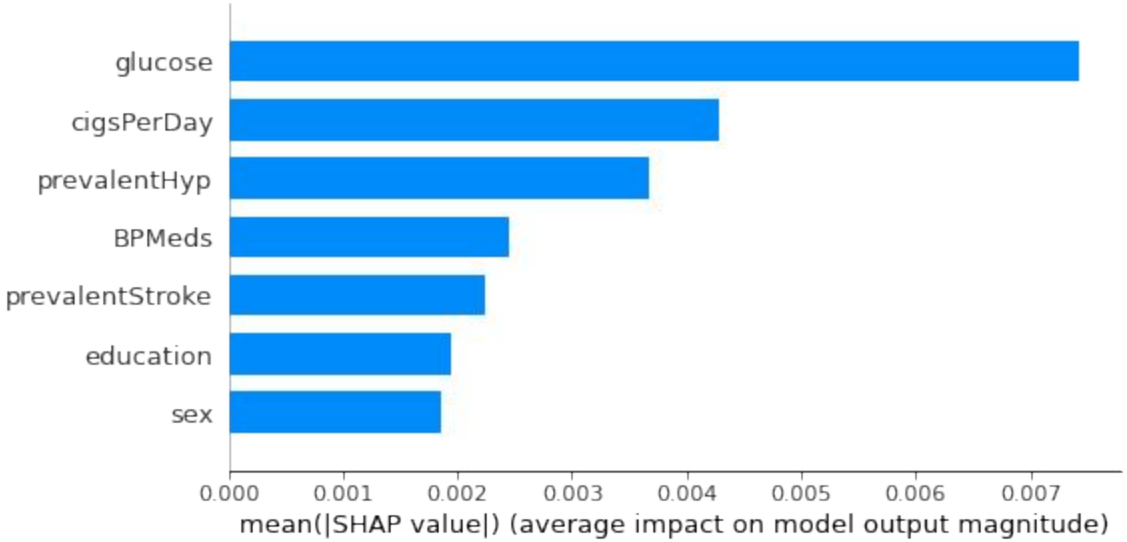
After cross validation and hyperparameter tuning, the best parameters are {'leaf_size': 30, 'n_neighbors': 19} (test accuracy is 84% and variance in prediction is 1%)



Building and Evaluation of Model-3

Final SVC Model:

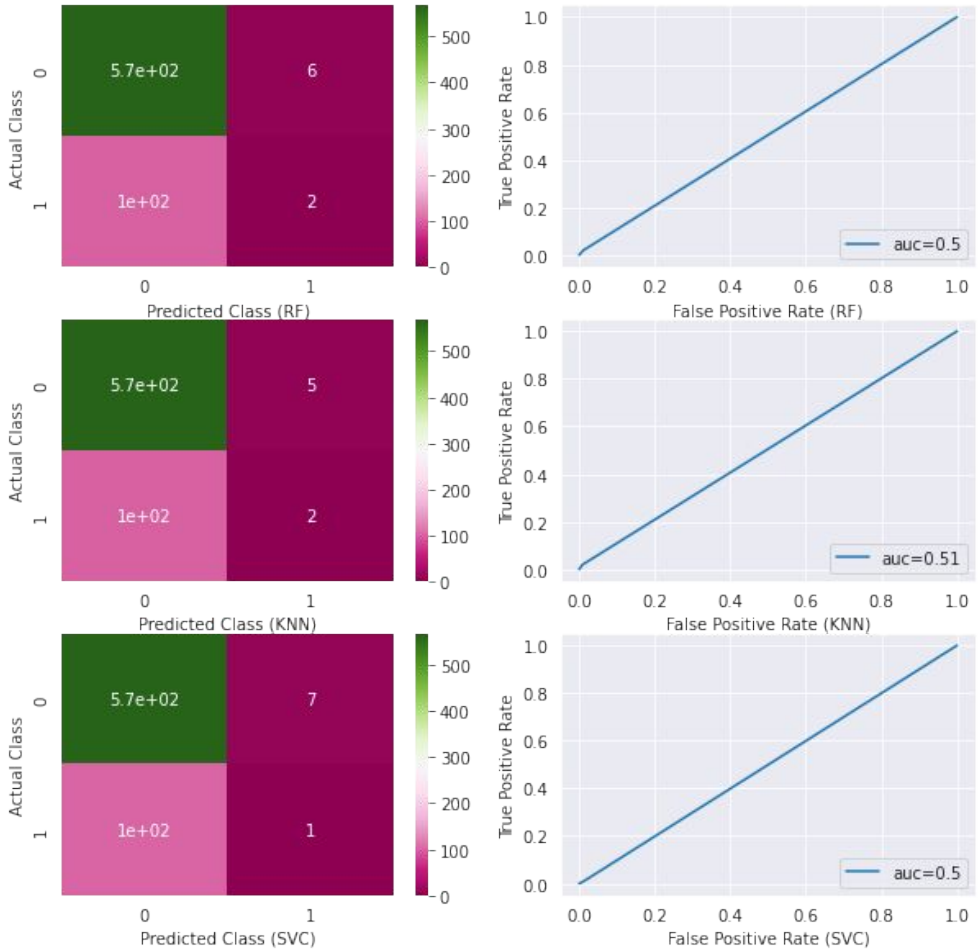
After cross validation and hyperparameter tuning, the best parameters are {'C': 6, 'gamma': 0.1} (test accuracy is 84% and variance in prediction is 2%)



RF- Train_Accuracy: 0.86, Test_Accuracy: 0.84, Test_F1: 0.47

KNN- Train_Accuracy: 0.85, Test_Accuracy: 0.84, Test_F1: 0.48

SVC- Train_Accuracy: 0.86, Test_Accuracy: 0.84, Test_F1: 0.46



Final Conclusion

On the basis of the performance study of our three models, we are selecting **KNN** classifier *(the best warrior)* for predicting 10 Years Coronary heart disease, as it has low variance in prediction, good f1_score and good ROC_AUC score among all three models

Thank you !