Capstone Project-3

Cardiovascular Risk Prediction

Name: Saurabh Yadav

Content

1. Problem Statement
2. Data Summary
3. Analysis of Data
4. Null value Imputation/ Data Cleaning
5. Data Preprocessing
6. Feature Engineering/ Selection
7. Model Training
8. Evaluation Metrics
9. Challenges
10. Conclusion

Problem Statement:

1. The objective of the project is to come up with the machine learning model to predict whether a patient has 10-year risk of developing coronary heart disease (CHD) using the residents of the town of Framingham, Massachusetts dataset.

Data Summary:

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.

Attributes:

Demographic:

- Sex: male or female("M" or "F")(Nominal)
- Age: Age of the patient(Continuous)

Behavioral:

- is_smoking: Whether or not the patient is a current smoker ("YES" or "NO") (Nominal)
- cigsPerDay: the number of cigarettes that the person smoked on average in one day.(Continuous)

Contd..

Medical(history):

- BPmeds: whether or not the patient was on blood pressure medication (Nominal)
- prevalentStroke: whether or not the patient had previously had a stroke (Nominal)
- prevalentHyp: whether or not the patient was hypertensive (Nominal)
- diabetes: whether or not the patient had diabetes (Nominal)

Medical(current)

- totChol: total cholesterol level (Continuous)
- sysBP: systolic blood pressure (Continuous)
- diaBP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- heartRate: heart rate (Continuous)
- glucose: glucose level (Continuous)

Predict variable (desired target):

- **TenYearCHD:**10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) –Discrete variable

Contd..

Medical(history):

- BPmeds: whether or not the patient was on blood pressure medication (Nominal)
- prevalentStroke: whether or not the patient had previously had a stroke (Nominal)
- prevalentHyp: whether or not the patient was hypertensive (Nominal)
- diabetes: whether or not the patient had diabetes (Nominal)

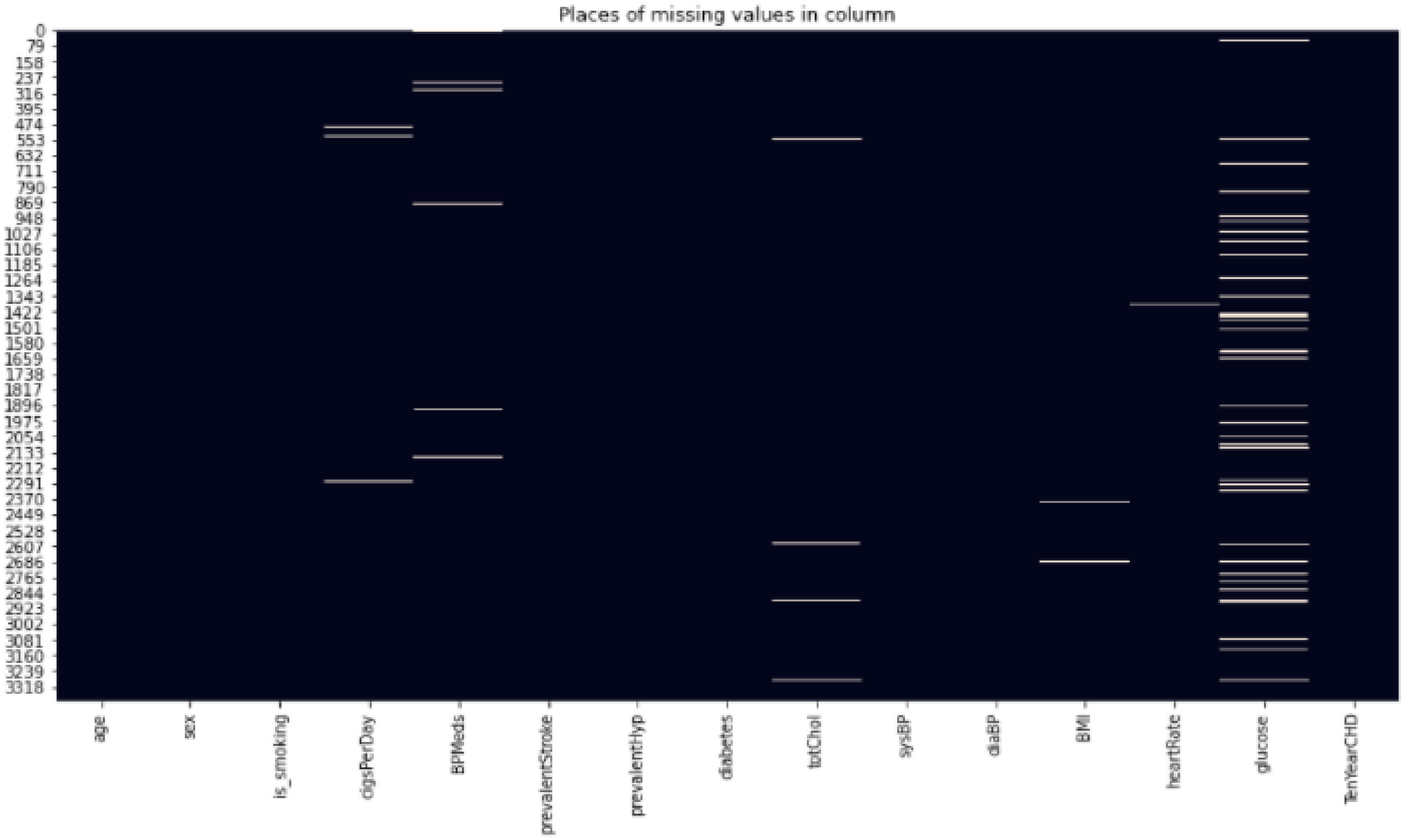
Medical(current)

- totChol: total cholesterol level (Continuous)
- sysBP: systolic blood pressure (Continuous)
- diaBP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- heartRate: heart rate (Continuous)
- glucose: glucose level (Continuous)

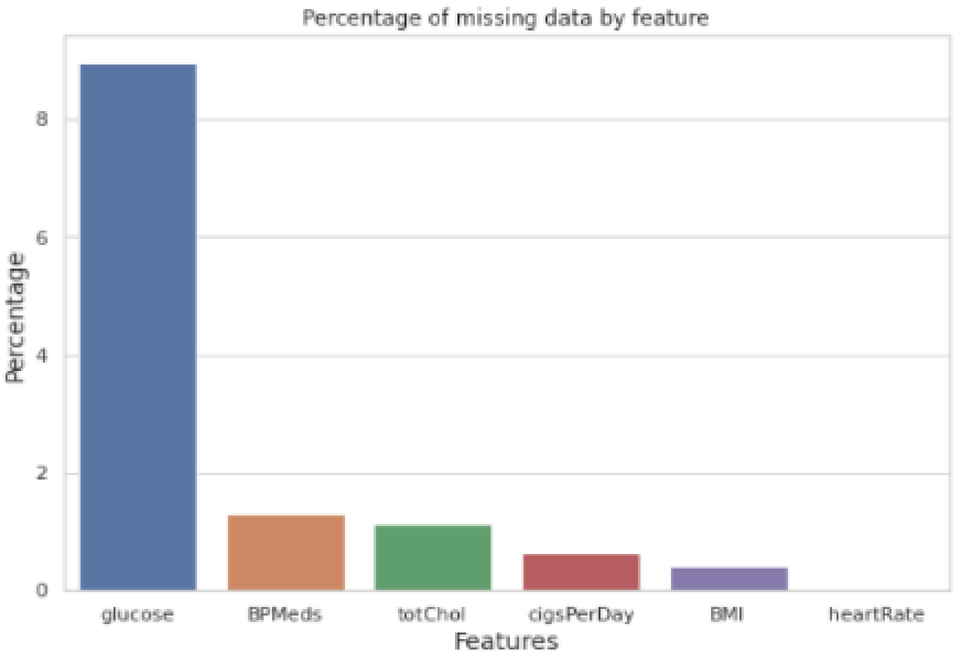
Predict variable (desired target):

- **TenYearCHD:**10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”) –Discrete variable

Where are the missing values in our data?



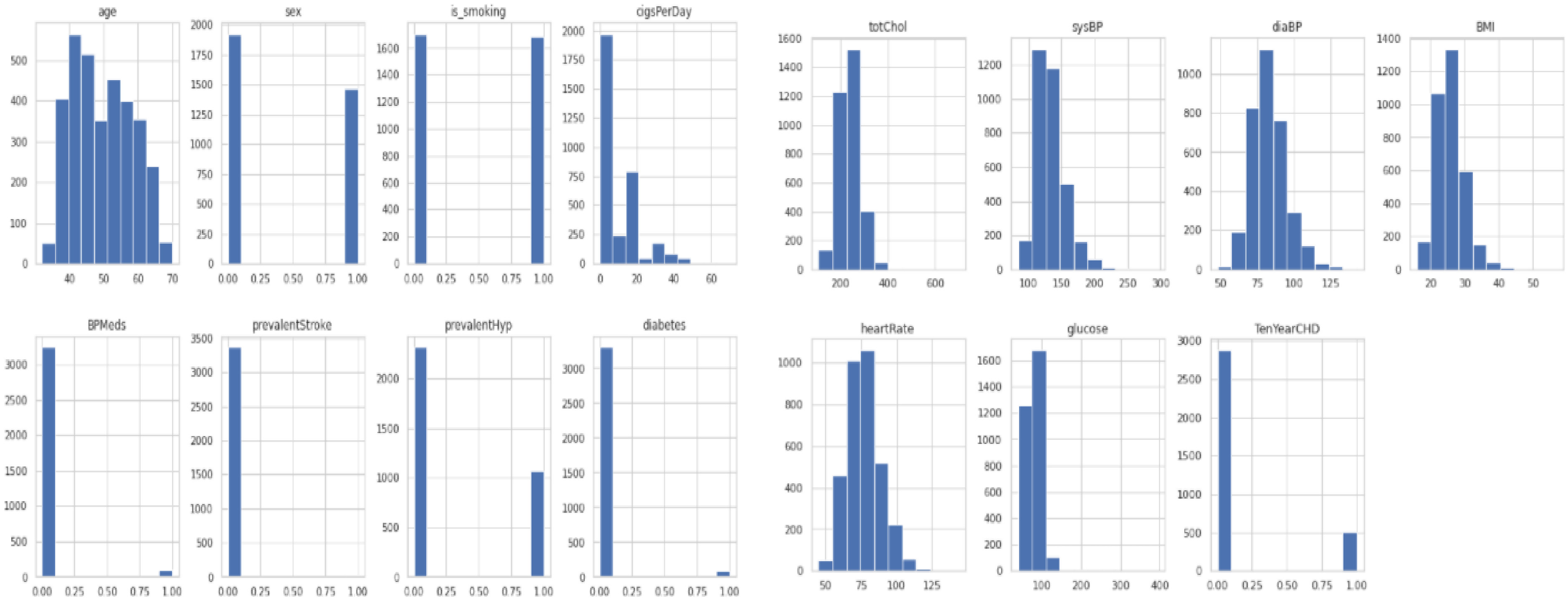
Missing value treatment :



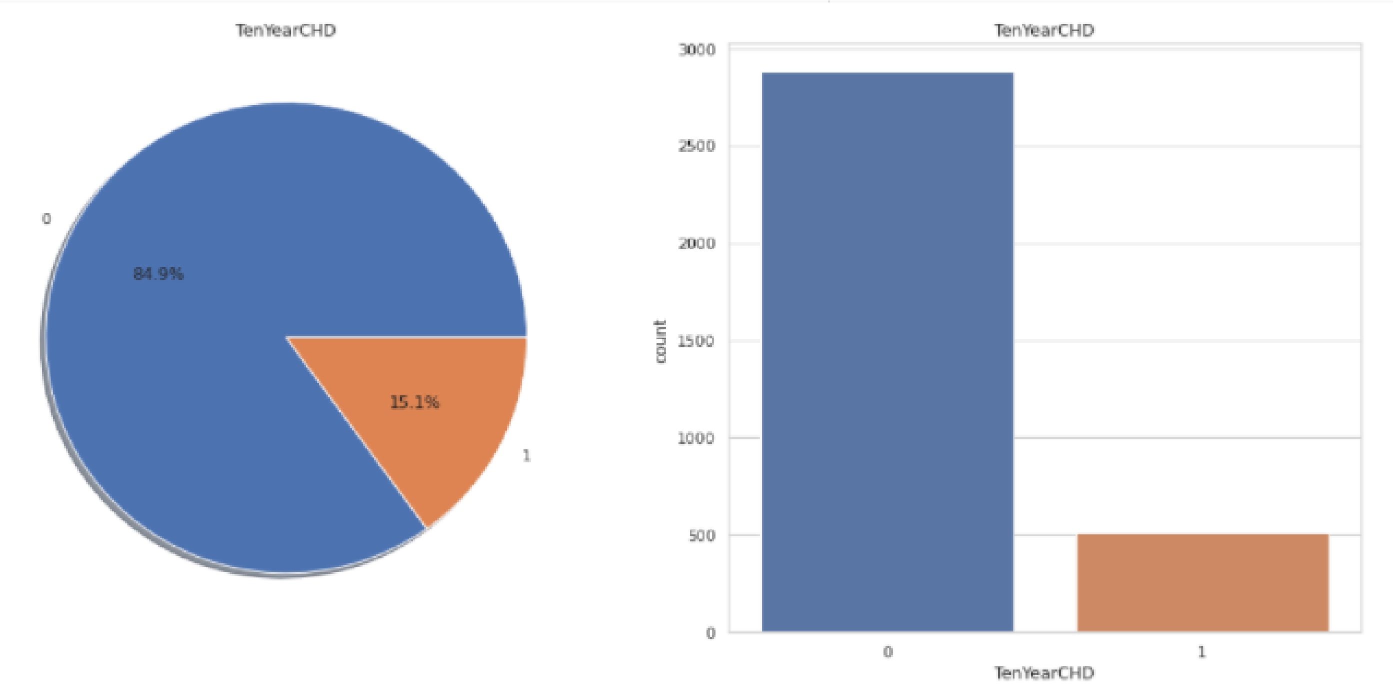
At **8.97%**, the blood glucose entry has the highest percentage of missing data. The other features have very few missing entries.

Since the missing entries account for only **11%** of the total data so, we can exclude these entries without losing most of the data.

Distribution of the features

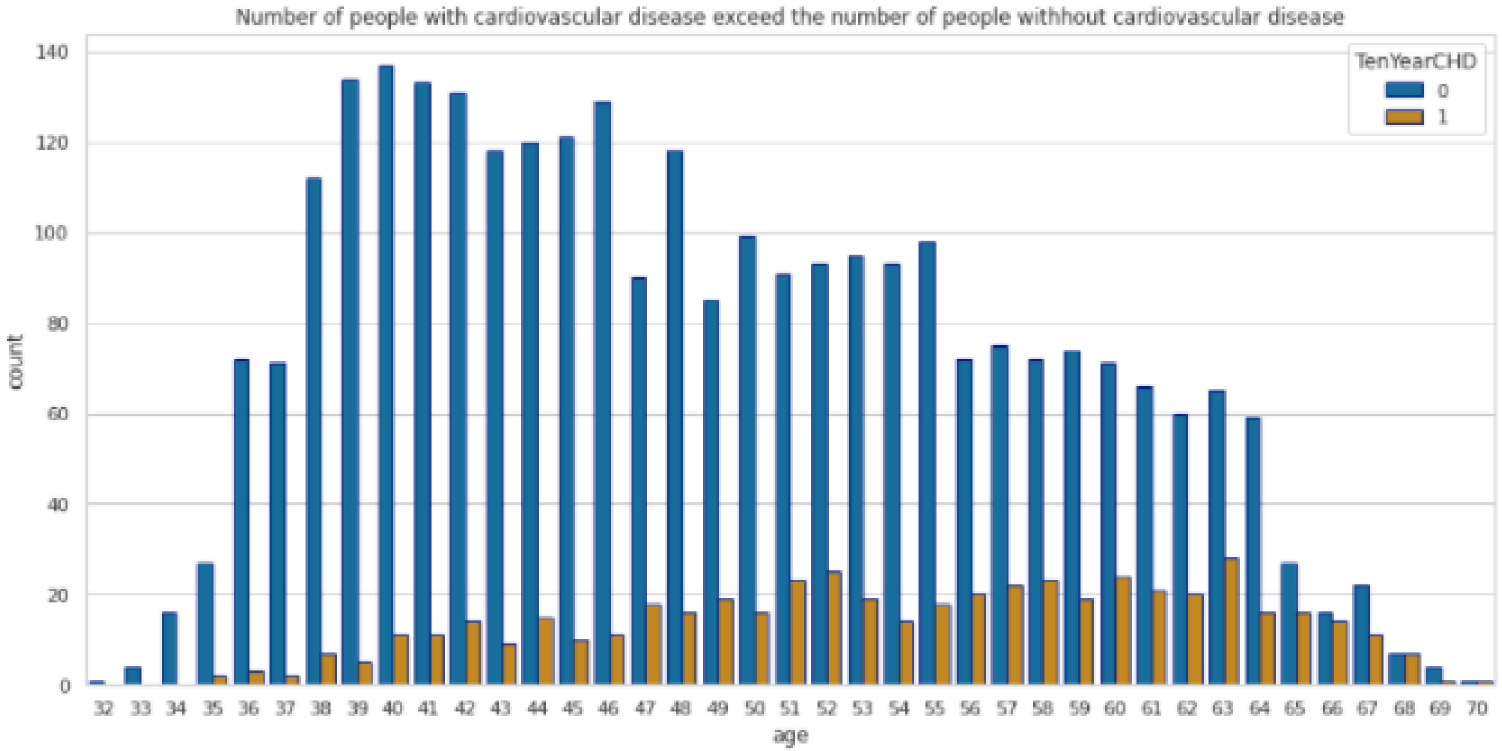


Analysis of features

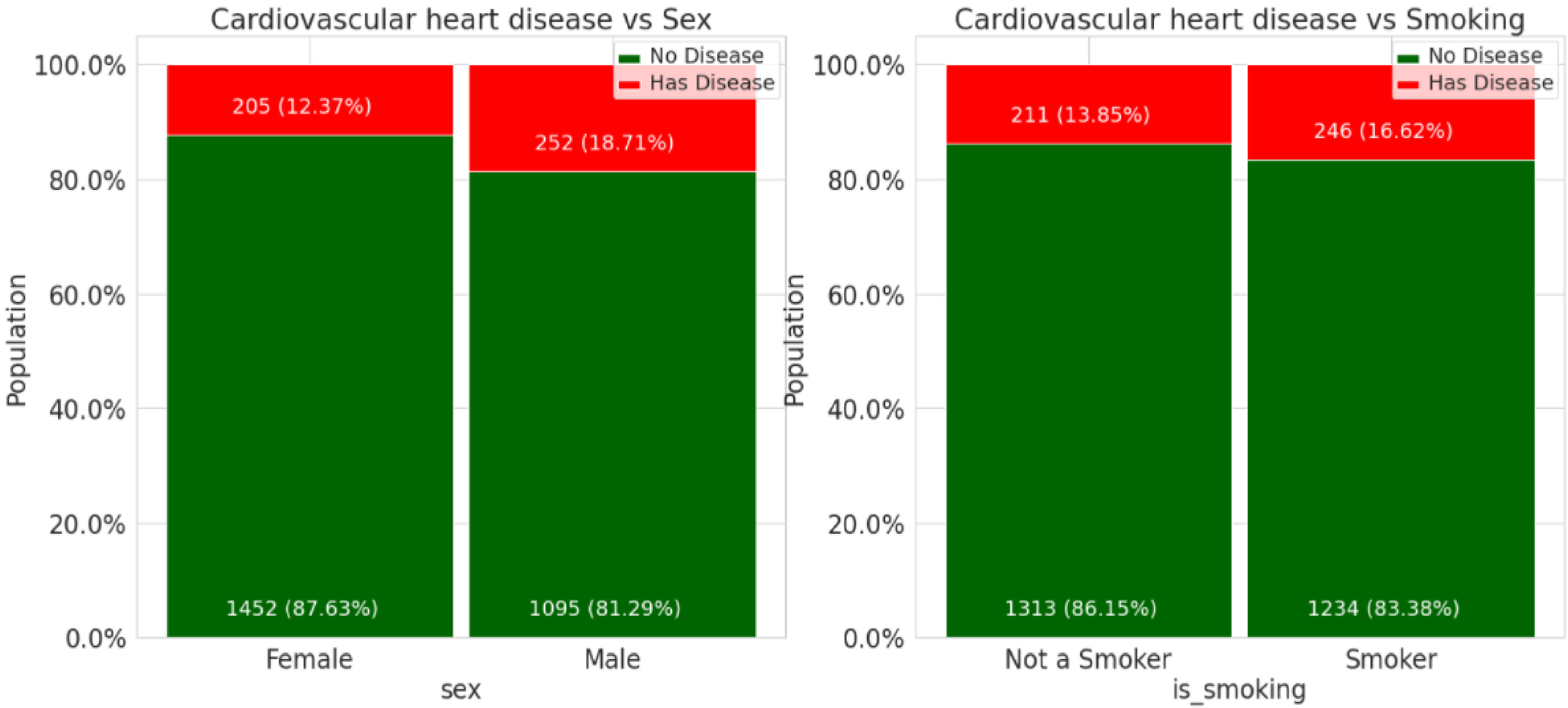


- **Target variable Analysis**: We have the imbalanced data set as the number of people without the disease greatly exceeds the number of people with the disease.

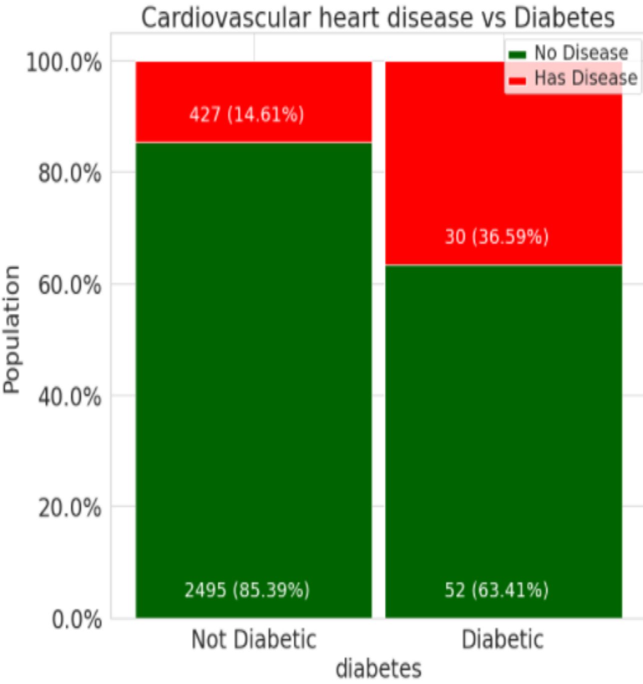
Visualize the target and age variable:



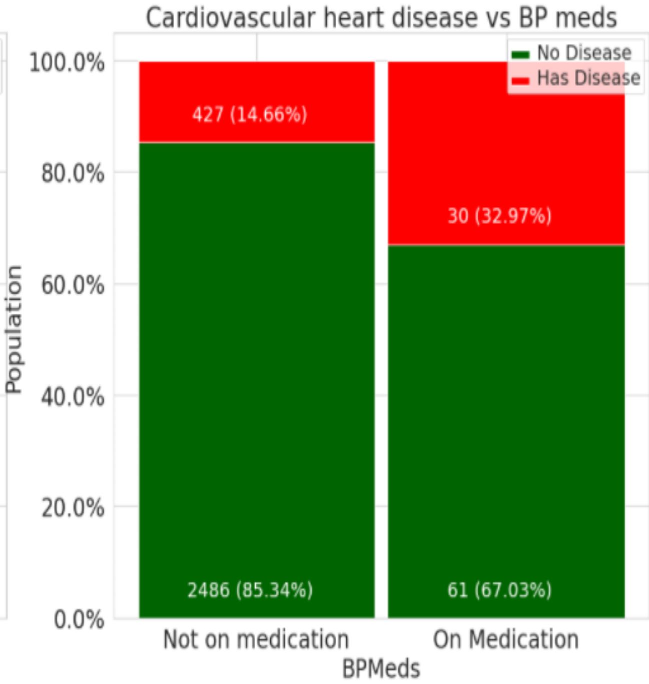
Analysis of categorical variables with respect to target variable



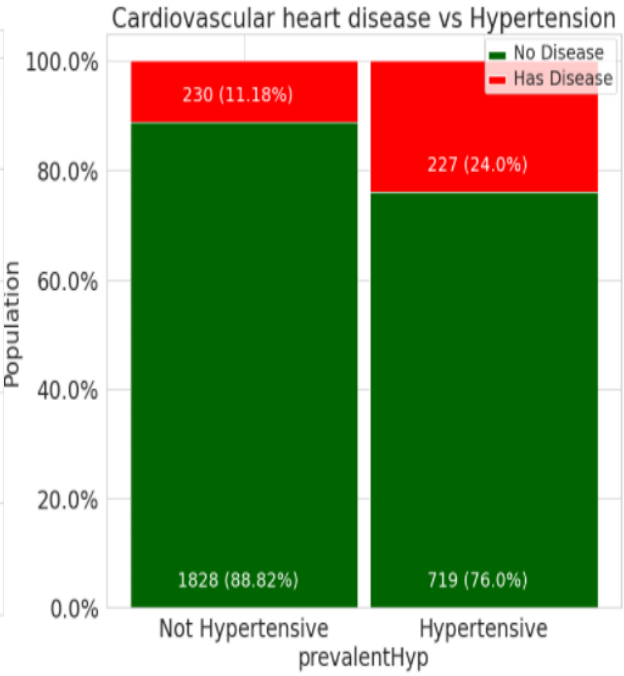
Contd..



Cardiovascular heart disease vs Diabetes



Cardiovascular heart disease vs BP meds



Cardiovascular heart disease vs Hypertension

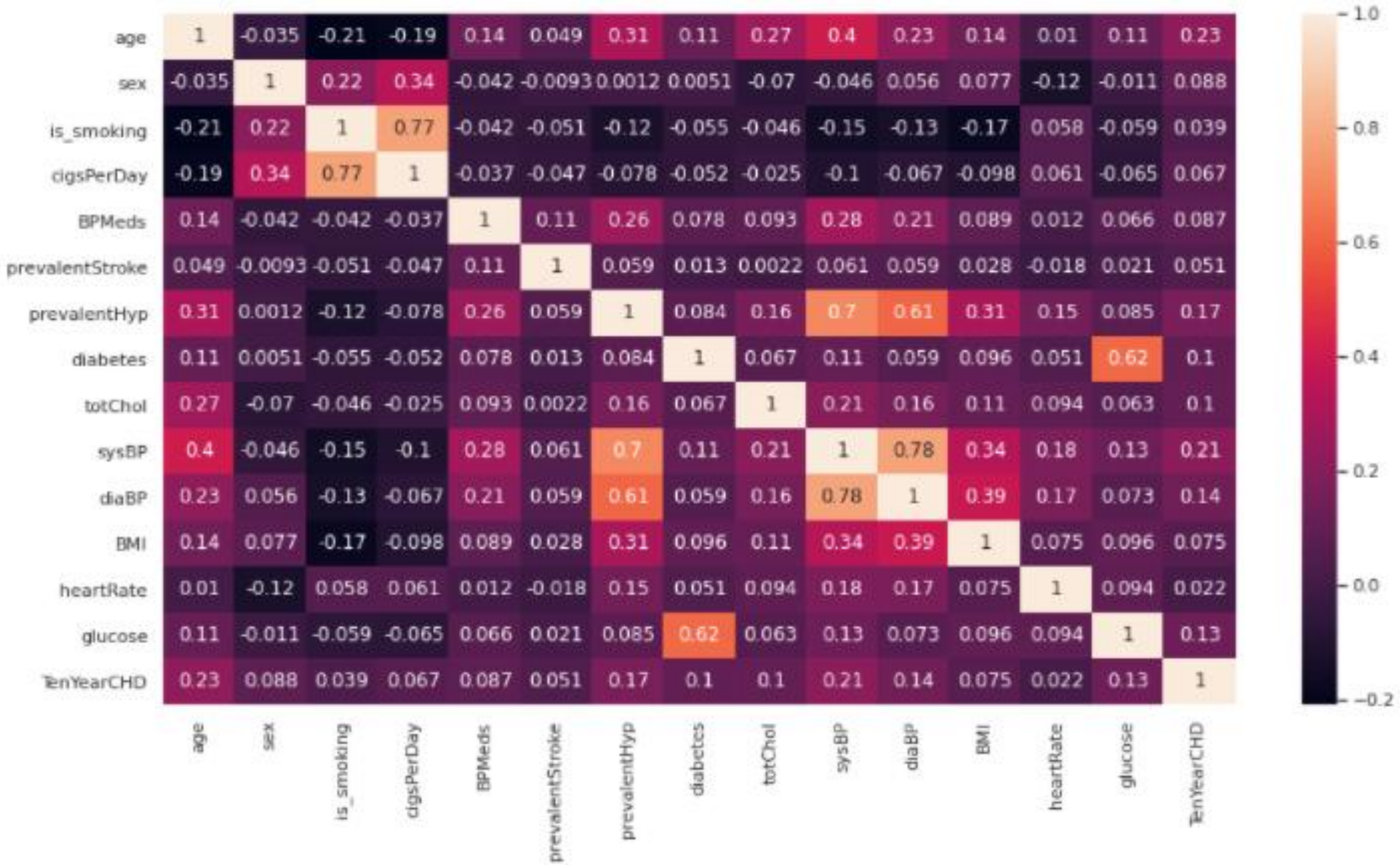
- Slightly more males are suffering from Cardiovascular heart disease than females.

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

- The percentage of people who have Cardiovascular heart disease is higher among the diabetic patients and also those patients with prevalent hypertension have more risk of Cardiovascular heart disease compare to those who don't have hypertensive problem.

- The percentage of people who are on medication of blood pressure have more risk of Cardiovascular heart disease compare to those who are not on medication.

Correlation between the variables:



	age	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
age	1	-0.035	-0.21	-0.19	0.14	0.049	0.31	0.11	0.27	0.4	0.23	0.14	0.01	0.11	0.23
sex	-0.035	1	0.22	0.34	-0.042	-0.0093	0.0012	0.0051	-0.07	-0.046	0.056	0.077	-0.12	-0.011	0.088
is_smoking	-0.21	0.22	1	0.77	-0.042	-0.051	-0.12	-0.055	-0.046	-0.15	-0.13	-0.17	0.058	-0.059	0.039
cigsPerDay	-0.19	0.34	0.77	1	-0.037	-0.047	-0.078	-0.052	-0.025	-0.1	-0.067	-0.098	0.061	-0.065	0.067
BPMeds	0.14	-0.042	-0.042	-0.037	1	0.11	0.26	0.078	0.093	0.28	0.21	0.089	0.012	0.066	0.087
prevalentStroke	0.049	-0.0093	-0.051	-0.047	0.11	1	0.059	0.013	0.0022	0.061	0.059	0.028	-0.018	0.021	0.051
prevalentHyp	0.31	0.0012	-0.12	-0.078	0.26	0.059	1	0.084	0.16	0.7	0.61	0.31	0.15	0.085	0.17
diabetes	0.11	0.0051	-0.055	-0.052	0.078	0.013	0.084	1	0.067	0.11	0.059	0.096	0.051	0.62	0.1
totChol	0.27	-0.07	-0.046	-0.025	0.093	0.0022	0.16	0.067	1	0.21	0.16	0.11	0.094	0.063	0.1
sysBP	0.4	-0.046	-0.15	-0.1	0.28	0.061	0.7	0.11	0.21	1	0.78	0.34	0.18	0.13	0.21
diaBP	0.23	0.056	-0.13	-0.067	0.21	0.059	0.61	0.059	0.16	0.78	1	0.39	0.17	0.073	0.14
BMI	0.14	0.077	-0.17	-0.098	0.089	0.028	0.31	0.096	0.11	0.34	0.39	1	0.075	0.096	0.075
heartRate	0.01	-0.12	0.058	0.061	0.012	-0.018	0.15	0.051	0.094	0.18	0.17	0.075	1	0.094	0.022
glucose	0.11	-0.011	-0.059	-0.065	0.066	0.021	0.085	0.62	0.063	0.13	0.073	0.096	0.094	1	0.13
TenYearCHD	0.23	0.088	0.039	0.067	0.087	0.051	0.17	0.1	0.1	0.21	0.14	0.075	0.022	0.13	1

Contd..

- There are no features with more than 0.5 correlation with the Ten year risk of developing CHD and this shows that the features a 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.

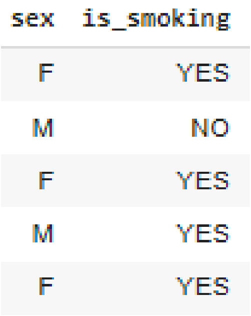
These includes:

- Blood glucose and diabetes;
- systolic and diastolic blood pressures;
- cigarette smoking and the number of cigarettes smoked per day

Data Preprocessing:

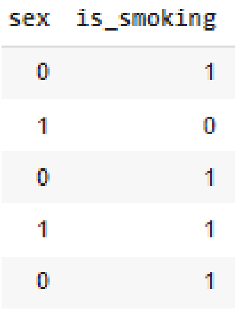
- We have the features like ‘id’ and ‘education’ which does not provide much more information so we remove that columns.
- We’ve the columns ‘**sex**’ and ‘**is_smoking**’ which are of string type so we convert them into integer by applying the function which converts the following:

1. In **sex** feature **M(Male)** will be converted to 1 and **F(Female)** will be converted to 0.
2. In **is_smoking** feature **YES** will be converted to 1 and **NO** will be converted to 0.



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

String converted to Integer



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

Contd..

- Since, we’ve the overall 11% of missing values in our dataset, so we can exclude these entries without losing most of the data/information.

No. of rows before operation	No. of rows after operation
3390	3004

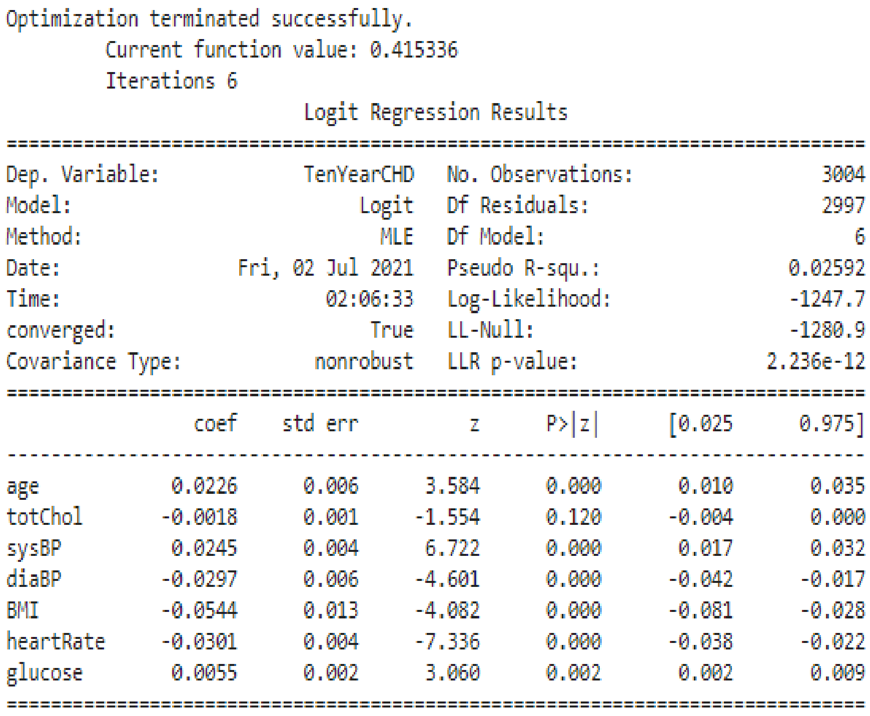
Feature Engineering/ Selection:

- For feature selection we’ve used **Tree-based: SelectFromModel** which is an embedded methods use algorithms that have built-in feature selection methods.
- We have used **RandomForest()** to select features based on feature importance. We calculate feature importance using node impurities in each decision tree.
- In Random forest, the final feature importance is the average of all decision tree feature importance.

After performing the feature selection the important features are:

- **['age', 'totChol', 'sysBP', 'diaBP', 'BMI', 'heartRate', 'glucose']**

Statistics on top/ important features:



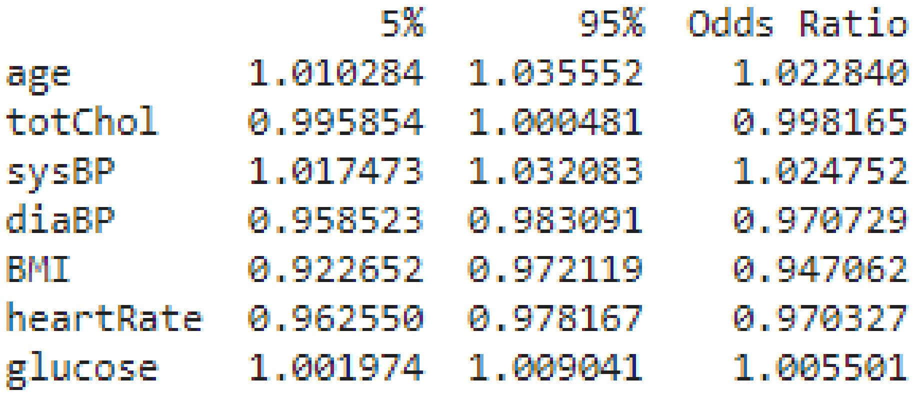
```
Optimization terminated successfully.
         Current function value: 0.415336
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:             TenYearCHD   No. Observations:                 3004
Model:                          Logit   Df Residuals:                     2997
Method:                           MLE   Df Model:                            6
Date:                Fri, 02 Jul 2021   Pseudo R-squ.:                 0.02592
Time:                        02:06:33   Log-Likelihood:                -1247.7
converged:                       True   LL-Null:                       -1280.9
Covariance Type:            nonrobust   LLR p-value:                 2.236e-12
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
age            0.0226      0.006      3.584      0.000       0.010       0.035
totChol       -0.0018      0.001     -1.554      0.120      -0.004       0.000
sysBP          0.0245      0.004      6.722      0.000       0.017       0.032
diaBP         -0.0297      0.006     -4.601      0.000      -0.042      -0.017
BMI           -0.0544      0.013     -4.082      0.000      -0.081      -0.028
heartRate     -0.0301      0.004     -7.336      0.000      -0.038      -0.022
glucose        0.0055      0.002      3.060      0.002       0.002       0.009
==============================================================================
```

Here, **Iterations** refer to the number of times the model iterates over the data, trying to optimize the model.

Explanation of some of the terms in the summary table:

- **coef**: It is the coefficients of the independent variables in the regression equation.
- **Log-Likelihood**: the natural logarithm of the Maximum Likelihood Estimation(MLE) function. MLE is the optimization process of finding the set of parameters which result in best fit.
- **LL-Null**: the value of log-likelihood of the model when no independent variable is included(only an intercept is included).
- **Pseudo R-squ.**: It is the ratio of the log-likelihood of the null model to that of the full model.

Checking the odds ratio of top features:



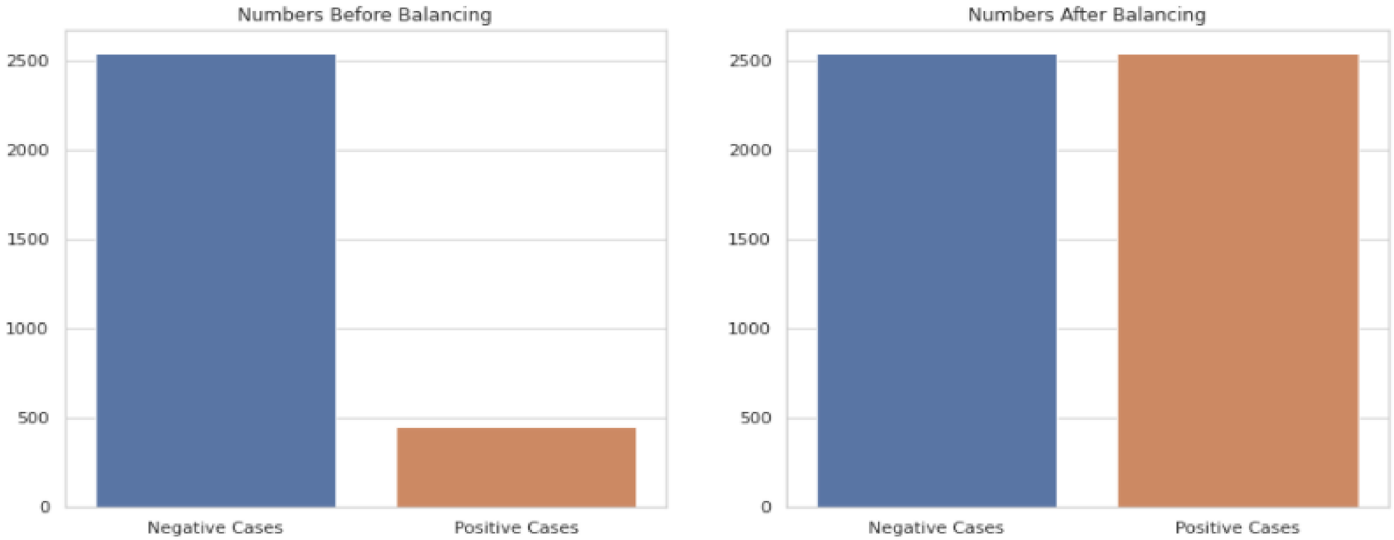
```
                 5%       95%  Odds Ratio
age        1.010284  1.035552    1.022840
totChol    0.995854  1.000481    0.998165
sysBP      1.017473  1.032083    1.024752
diaBP      0.958523  0.983091    0.970729
BMI        0.922652  0.972119    0.947062
heartRate  0.962550  0.978167    0.970327
glucose    1.001974  1.009041    1.005501
```

From above table we can conclude that,

- The odds of getting cardiovascular heart disease risk increases with about 2% for every increase in **age** and **systolic blood pressure(sysBP)**.
- The other factors show no significant positive odds.

Model building and predictions

- First we balance our dataset because for every positive case there are about **5-6** negative cases.
- To handle this problem we will balance the dataset using the **Synthetic Minority Oversampling Technique (SMOTE)**

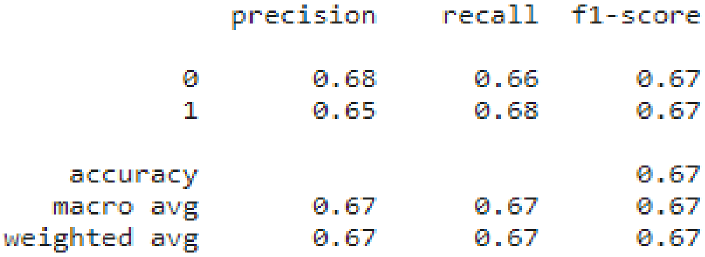


Algorithms Used:

Here, we’ll be using this 4 algorithms along with **GridsSearchCV** for finding optimum parameters:

1. **Logistic Regression**
2. **Random Forrest**
3. **XG-Boost**
4. **Support Vector Machine**

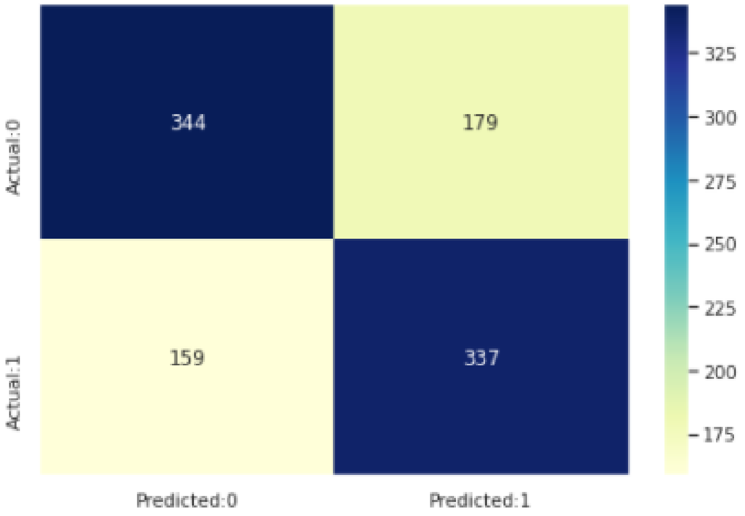
1. Logistic regression

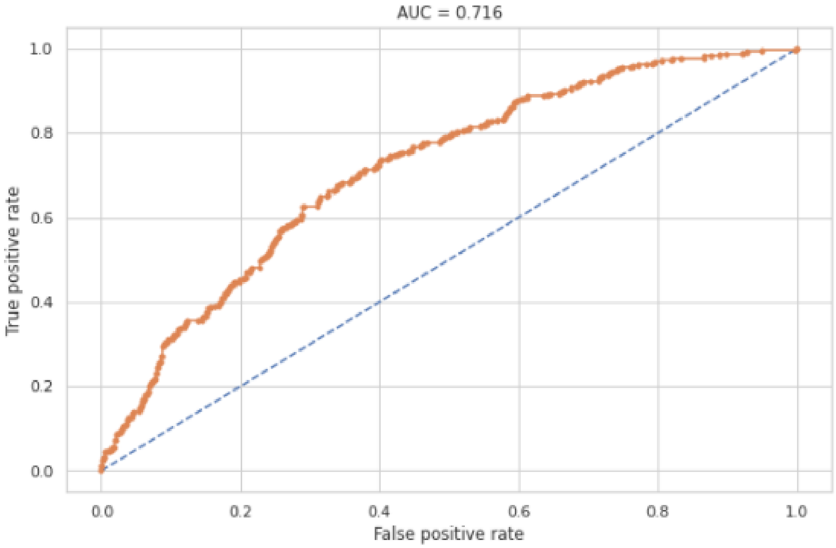


```
              precision    recall  f1-score

           0       0.68      0.66      0.67
           1       0.65      0.68      0.67

    accuracy                           0.67
   macro avg       0.67      0.67      0.67
weighted avg       0.67      0.67      0.67
```

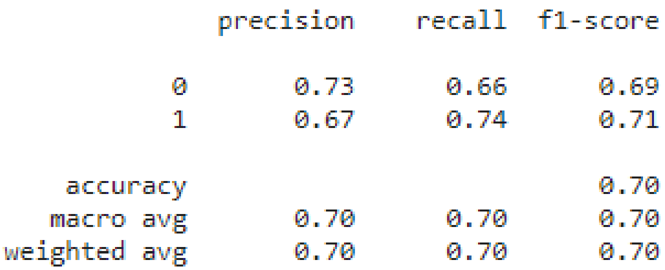
Best parameters:

{'C': 10,
 'class_weight': None,
 'penalty': 'l2'}

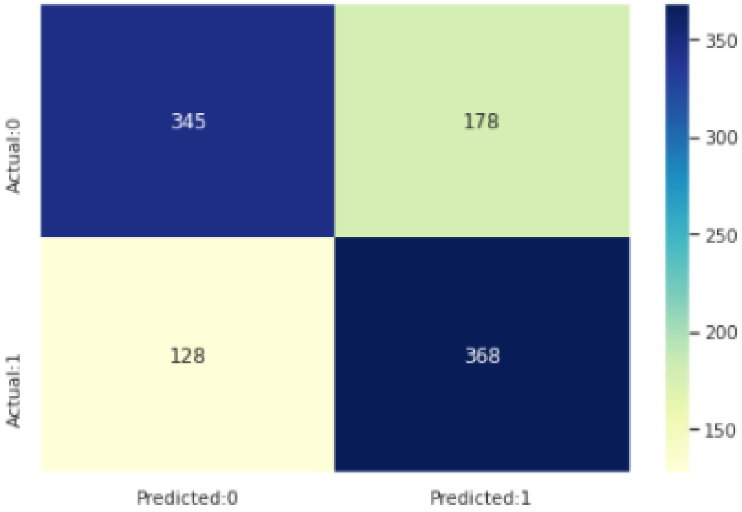




2. Random Forrest

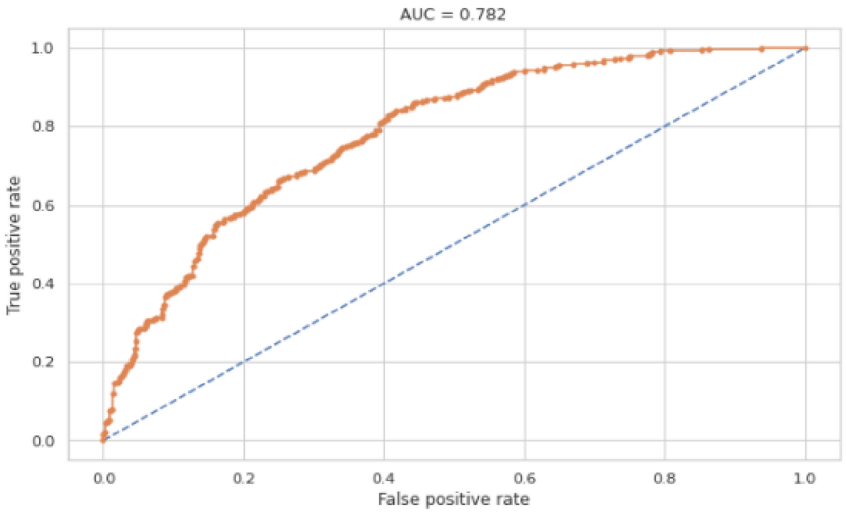


```
              precision    recall  f1-score

           0       0.73      0.66      0.69
           1       0.67      0.74      0.71

    accuracy                           0.70
   macro avg       0.70      0.70      0.70
weighted avg       0.70      0.70      0.70
```

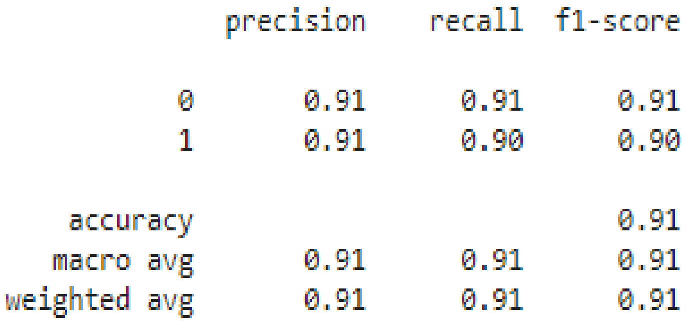


Best parameters:

{'max_depth': 8,
 'min_samples_leaf': 40,
'min_samples_split': 50,
'n_estimators': 80}



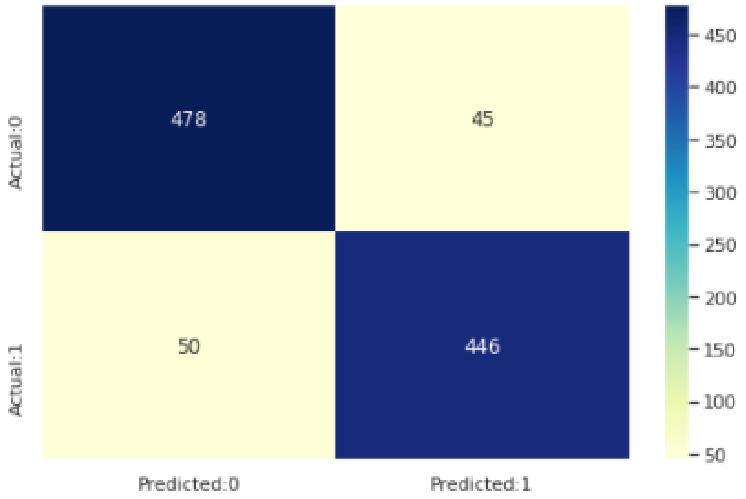
3. XGBoost

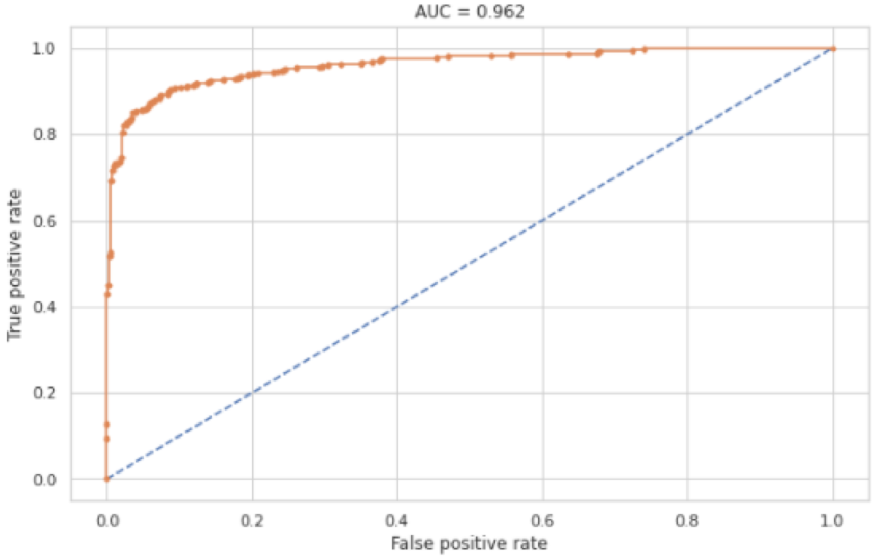


```
              precision    recall  f1-score

           0       0.91      0.91      0.91
           1       0.91      0.90      0.90

    accuracy                           0.91
   macro avg       0.91      0.91      0.91
weighted avg       0.91      0.91      0.91
```

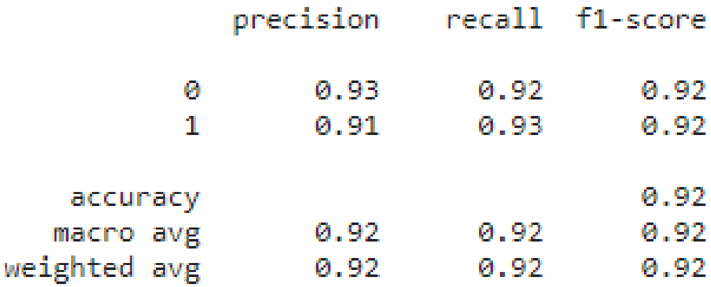
Best parameters:

{'learning_rate': 0.1,
'max_depth': 11,
'n_estimators': 200}

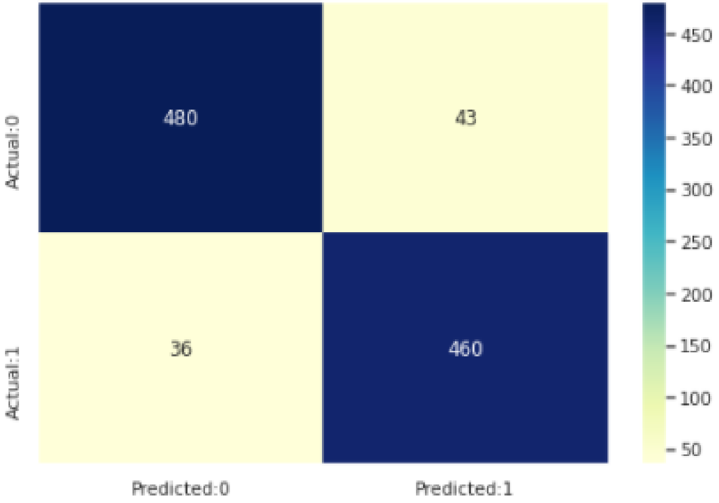




4. Support Vector Machine

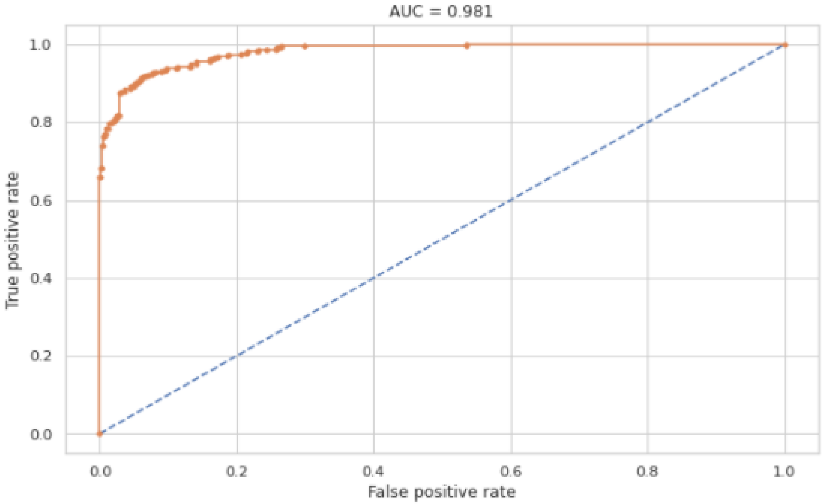


```
              precision    recall  f1-score

           0       0.93      0.92      0.92
           1       0.91      0.93      0.92

    accuracy                           0.92
   macro avg       0.92      0.92      0.92
weighted avg       0.92      0.92      0.92
```

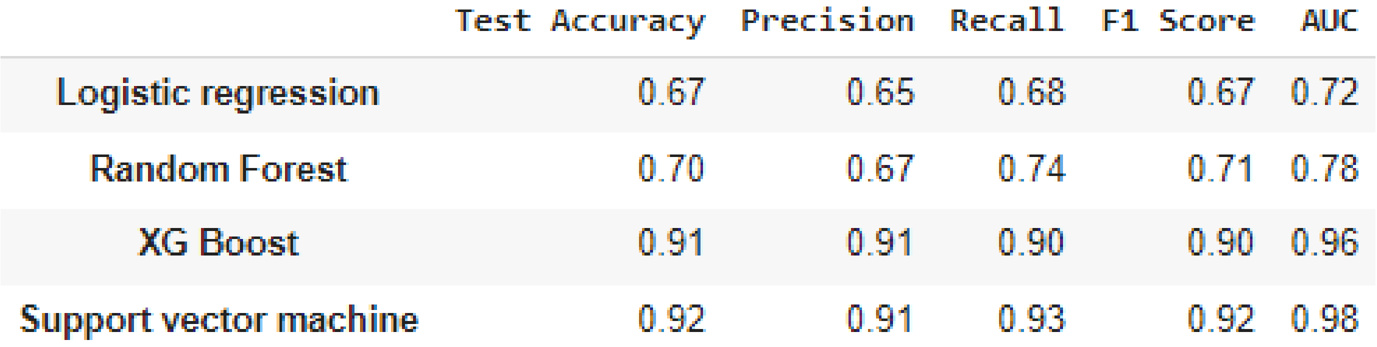


Best parameters:

{'learning_rate': 0.1,
'max_depth': 11,
'n_estimators': 200}



Comparison of Models Performance Metrics

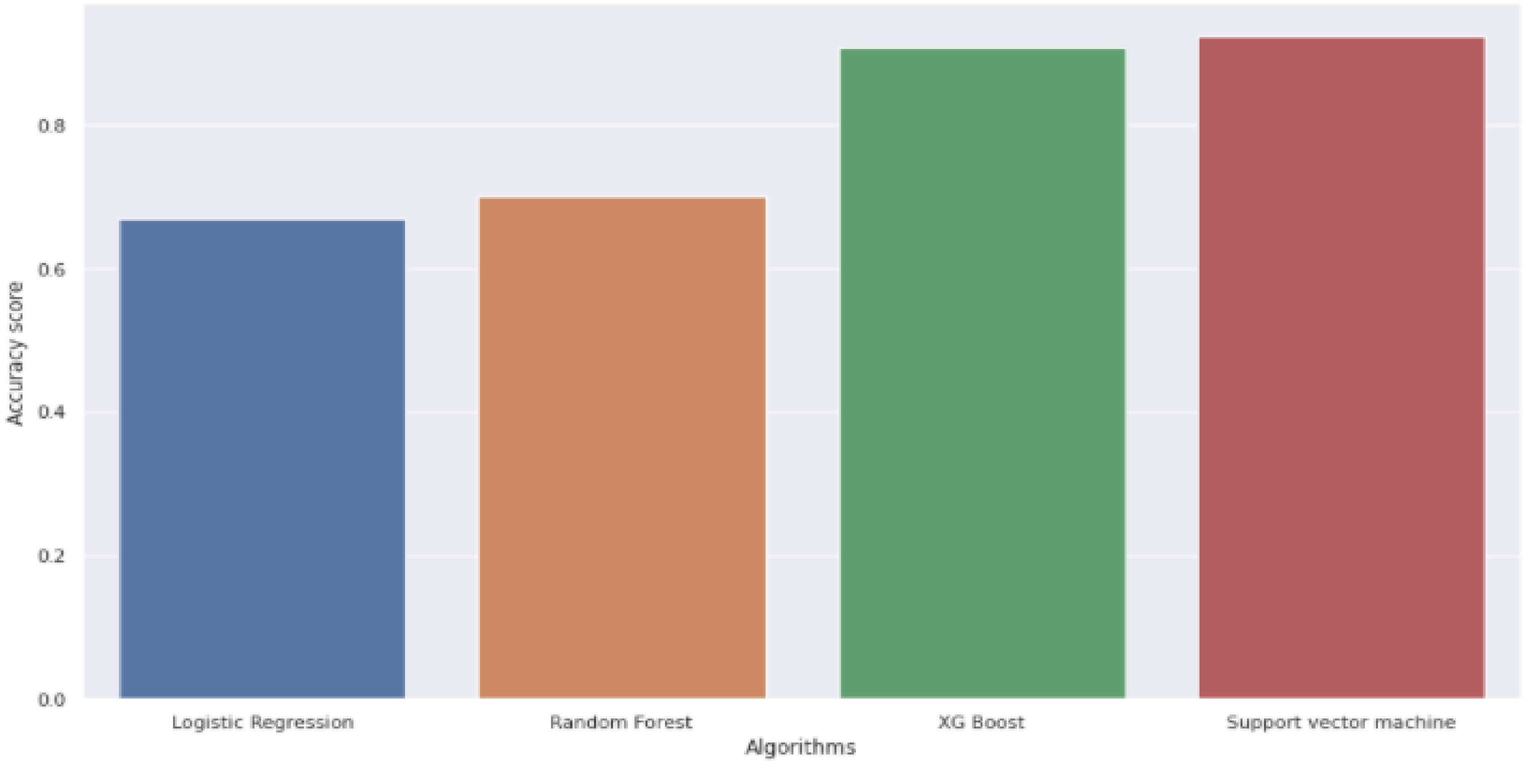


	Test Accuracy	Precision	Recall	F1 Score	AUC
Logistic regression	0.67	0.65	0.68	0.67	0.72
Random Forest	0.70	0.67	0.74	0.71	0.78
XG Boost	0.91	0.91	0.90	0.90	0.96
Support vector machine	0.92	0.91	0.93	0.92	0.98

- Observation from above table:
- **XG Boost**, **Support vector machine** gives highest Accuracy, Recall, Precision and AUC score.
- Highest recall and AUC score is given by **Support vector machine.**

Overall we can say that **Support vector machine** is the best model that can be used for the risk prediction of Cardiovasular Heart Disease.

Plotting Accuracy score with respect to each models



Challenges

- Handling the missing values.
- Making data more accurate.
- Selection of important features.

Conclusion

- The 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.

Q & A