Capstone Project – III

Cardiovascular Risk Prediction

NAVED MANSURI
Data Science Trainee, Almabetter

Point of Discussion

- ❑ Problem statement
- ❑ Data summary
- ❑ EDA
- ❑ Feature engineering
- ❑ Machine learning model
    - ❑ Logistics Regression
    - ❑ Random Forest
    - ❑ Support Vector Machine(SVM)
    - ❑ K-Nearest Neighbor(KNN)
    - ❑ XGBoost
- ❑ Model comparison
- ❑ Conclusion

Problem statement

- ❑ Cardiovascular diseases (CVDs) are the leading cause of death globally.
- ❑ An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.
- ❑ 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 3,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioural, and medical risk factors.

Data summary

- ❑ There are 3000+ rows and 17 columns in the data set and 15 are numeric features and 2 categorical.
- ❑ **TenYearCHD** is dependent variable.
- ❑ Data information
  - ❑ Demographic:

    Sex: male or female("M" or "F")

    Age: Age of the patients (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

    Education: The level of education of the patient (categorical values - 1,2,3,4)
  - ❑ Behavioural:

    is_smoking: whether or not the patient is a current smoker ("YES" or "NO")

    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.)

Data summary

- ❑ 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) Medical(current)

  Tot Chol: total cholesterol level (Continuous)

  Sys BP: systolic blood pressure (Continuous)

  Dia BP: diastolic blood pressure (Continuous)

  BMI: Body Mass Index (Continuous)

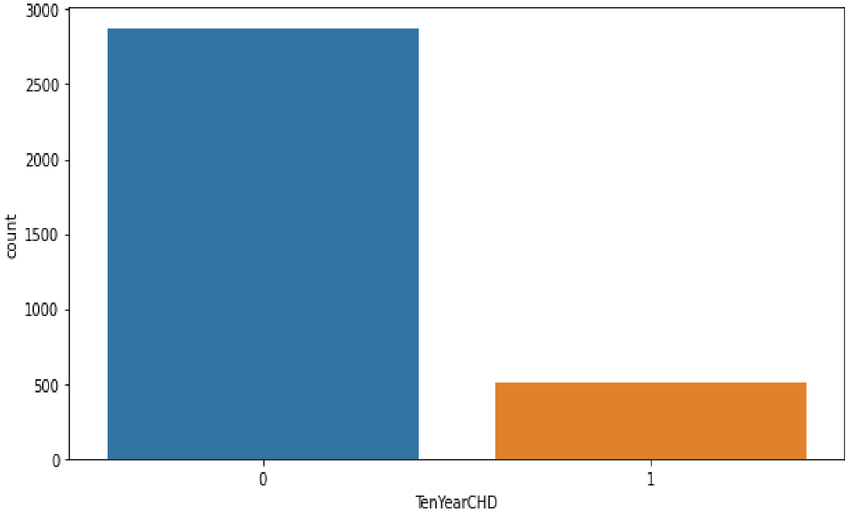
  Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)

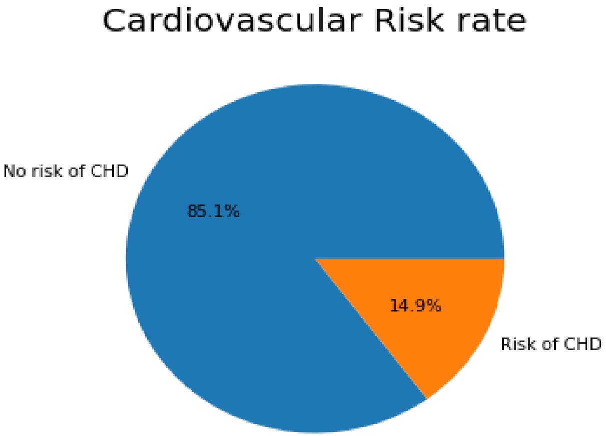
  Glucose: glucose level (Continuous) Predict variable (desired target)

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

Exploratory Data Analysis

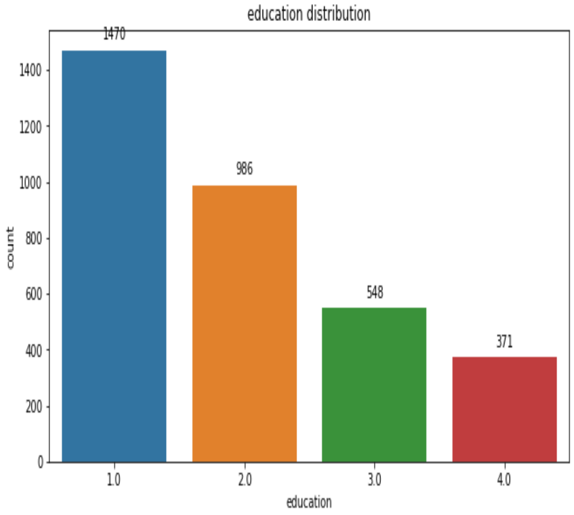
Analyzing Dependent Variable

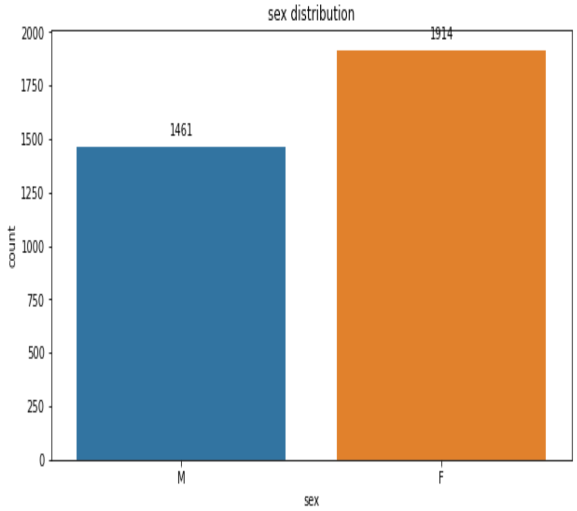


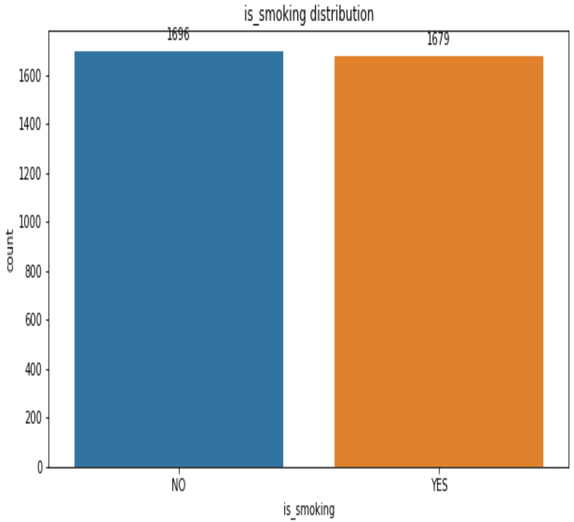


- ❑ Dependent variable(Ten year CHD) is binary, its only consist two values 0 or 1.
- ❑ Ten year CHD is imbalanced with 15% of risk CHD.

Analyzing Independent Variable





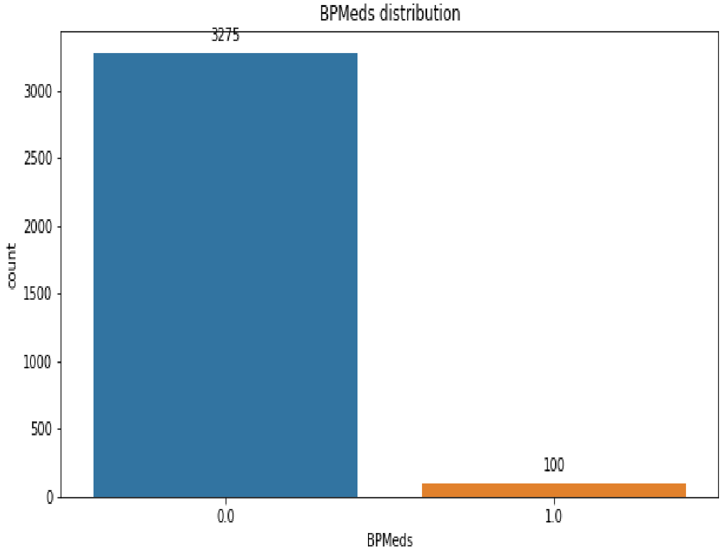


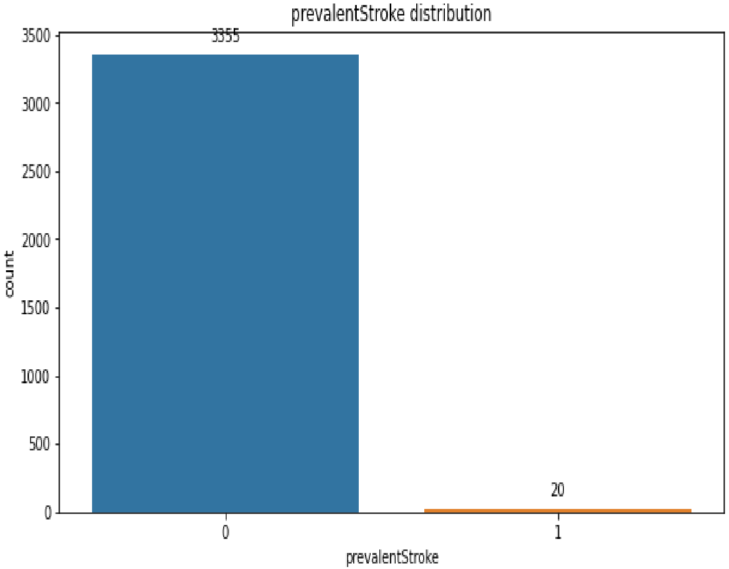
❑ Female are more compare to male's.

❑ Equally number of smackers.

❑ Most people are education level 1.

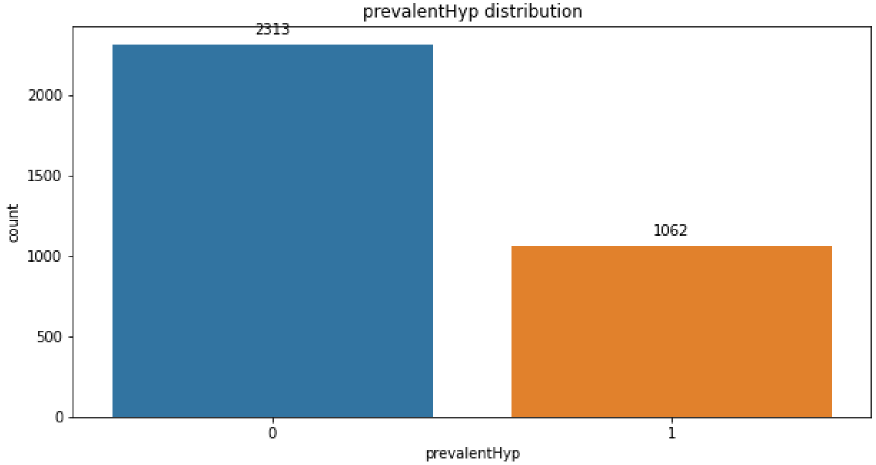
Analyzing Independent Variable

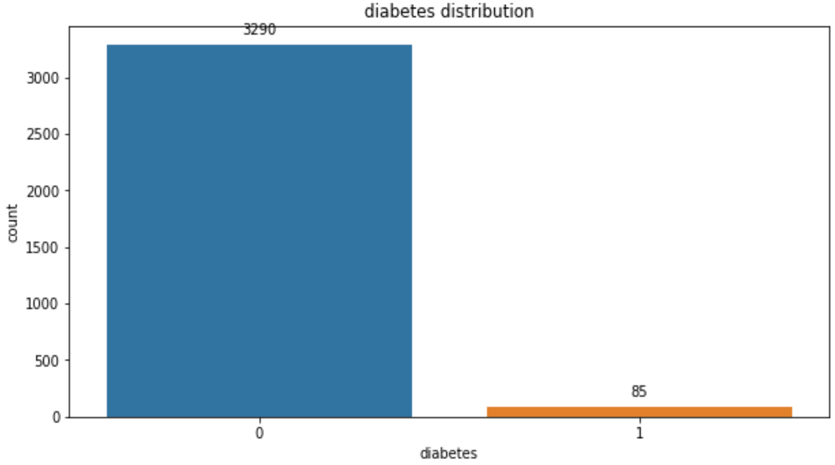




❑ Very less number of people having past blood pressure and hark stoke.

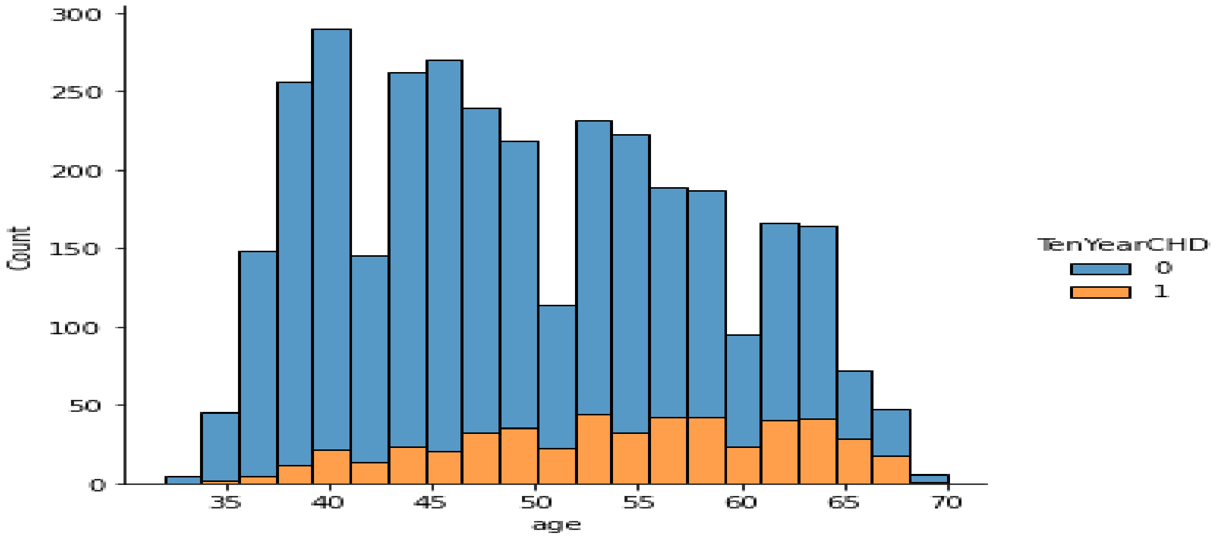
Analyzing Independent Variable





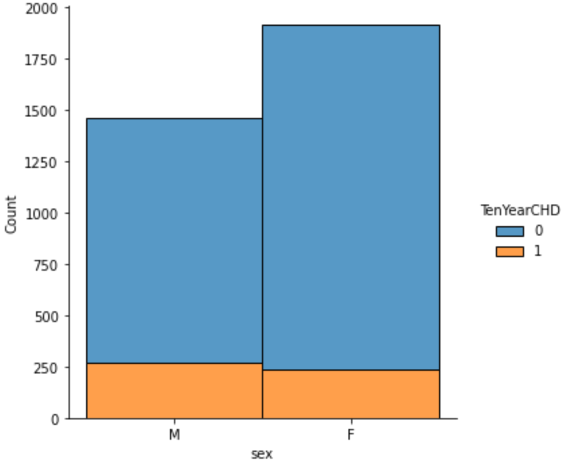
- ❑ 1000+ people having hypertension.
- ❑ A few peoples suffering from diabetes.

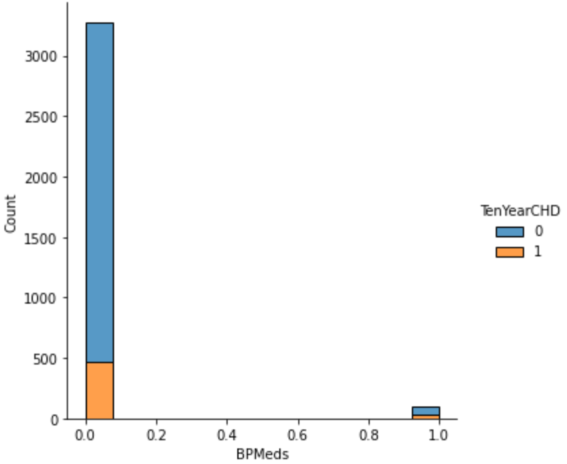
Analyzing Relationship Between Dependent And Independent Variables

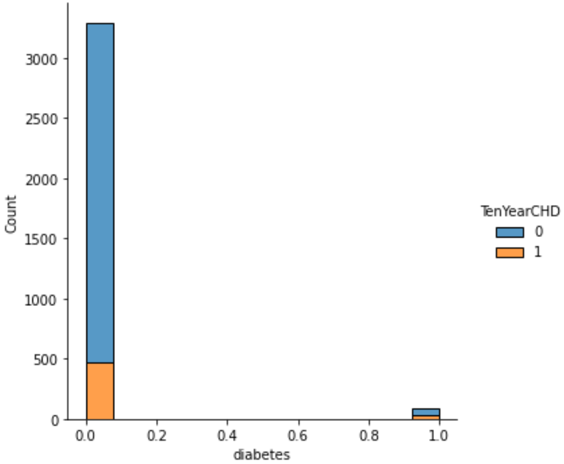


- ❑ Ages of 45 and 65 have the highest risk of acquiring heart disease

Analyzing Relationship Between Dependent And Independent Variables

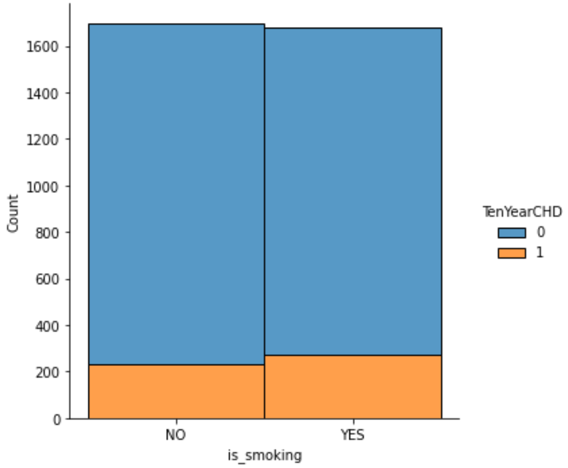


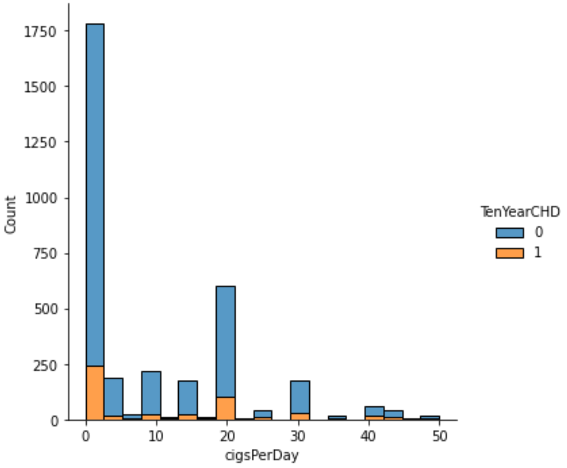




❑ Cardiovascular heart disease affects slightly more men than women.

Analyzing Relationship Between Dependent And Independent Variables

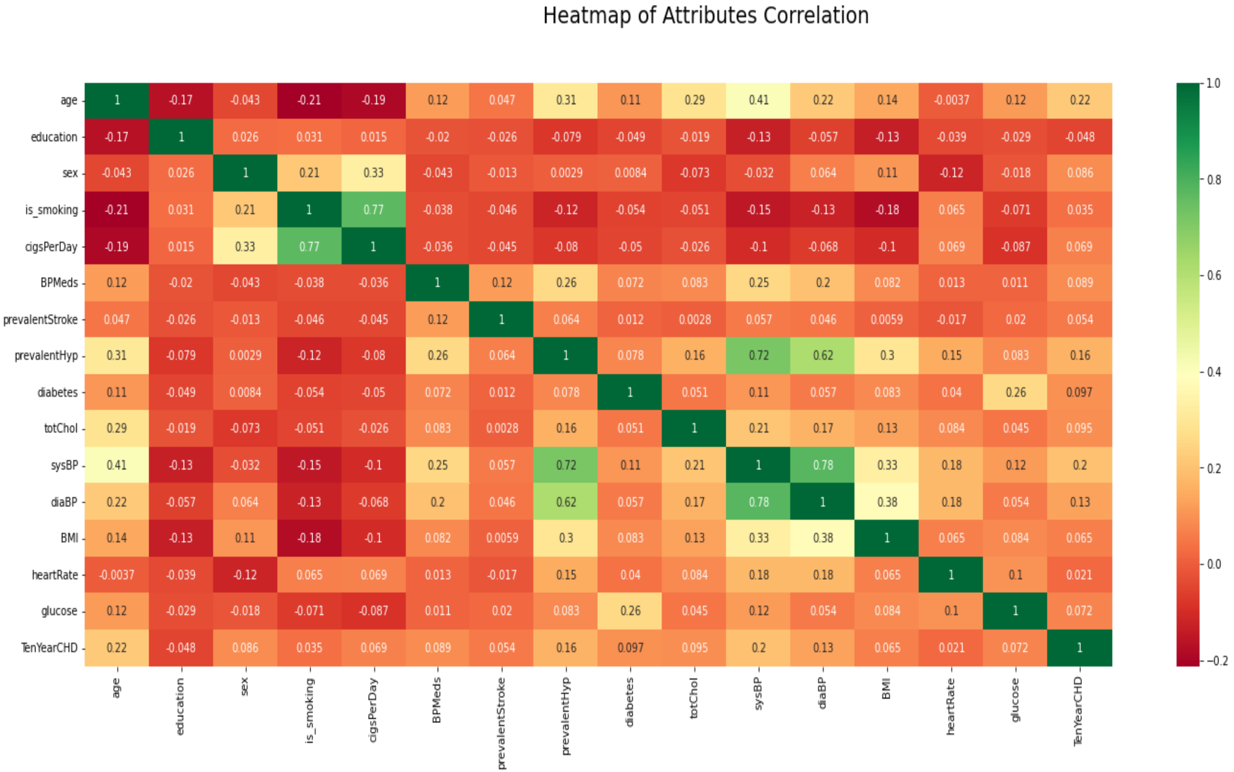




❑ Cardiovascular heart disease affects nearly equal numbers of smokers and non-smokers.

Correlation map

- Highest correlation between **systolic BP** and **diastolic BP**.
- **Systolic BP** and **Diastolic BP** shows a high correlation with **hypertension**.
- **cigarette** smoking and the number of cigarettes **smoked per day**.
- **Systolic BP** and **age** have a positive correlation.



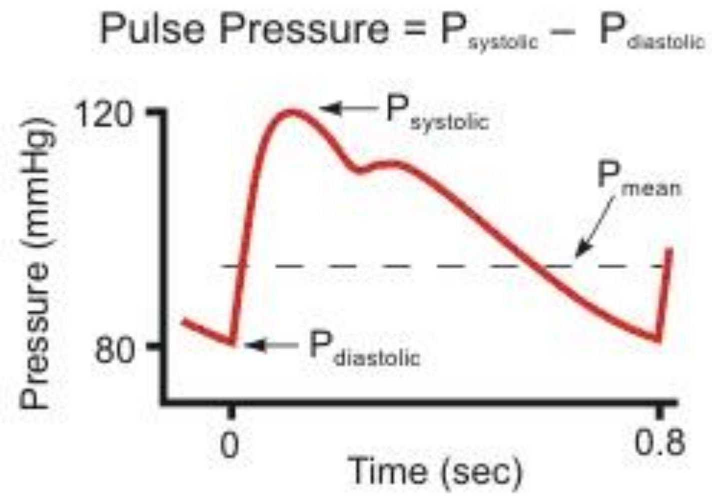
Heatmap of Attributes Correlation

	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
age	1	-0.17	-0.043	-0.21	-0.19	0.12	0.047	0.31	0.11	0.29	0.41	0.22	0.14	-0.0037	0.12	0.22
education	-0.17	1	0.026	0.031	0.015	-0.02	-0.026	-0.079	-0.049	-0.019	-0.13	-0.057	-0.13	-0.039	-0.029	-0.048
sex	-0.043	0.026	1	0.21	0.33	-0.043	-0.013	0.0029	0.0084	-0.073	-0.032	0.064	0.11	-0.12	-0.018	0.086
is_smoking	-0.21	0.031	0.21	1	0.77	-0.038	-0.046	-0.12	-0.054	-0.051	-0.15	-0.13	-0.18	0.065	-0.071	0.035
cigsPerDay	-0.19	0.015	0.33	0.77	1	-0.036	-0.045	-0.08	-0.05	-0.026	-0.1	-0.068	-0.1	0.069	-0.087	0.069
BPMeds	0.12	-0.02	-0.043	-0.038	-0.036	1	0.12	0.26	0.072	0.083	0.25	0.2	0.082	0.013	0.011	0.089
prevalentStroke	0.047	-0.026	-0.013	-0.046	-0.045	0.12	1	0.064	0.012	0.0028	0.057	0.046	0.0059	-0.017	0.02	0.054
prevalentHyp	0.31	-0.079	0.0029	-0.12	-0.08	0.26	0.064	1	0.078	0.16	0.72	0.62	0.3	0.15	0.083	0.16
diabetes	0.11	-0.049	0.0084	-0.054	-0.05	0.072	0.012	0.078	1	0.051	0.11	0.057	0.083	0.04	0.26	0.097
totChol	0.29	-0.019	-0.073	-0.051	-0.026	0.083	0.0028	0.16	0.051	1	0.21	0.17	0.13	0.084	0.045	0.095
sysBP	0.41	-0.13	-0.032	-0.15	-0.1	0.25	0.057	0.72	0.11	0.21	1	0.78	0.33	0.18	0.12	0.2
diaBP	0.22	-0.057	0.064	-0.13	-0.068	0.2	0.046	0.62	0.057	0.17	0.78	1	0.38	0.18	0.054	0.13
BMI	0.14	-0.13	0.11	-0.18	-0.1	0.082	0.0059	0.3	0.083	0.13	0.33	0.38	1	0.065	0.084	0.065
heartRate	-0.0037	-0.039	-0.12	0.065	0.069	0.013	-0.017	0.15	0.04	0.084	0.18	0.18	0.065	1	0.1	0.021
glucose	0.12	-0.029	-0.018	-0.071	-0.087	0.011	0.02	0.083	0.26	0.045	0.12	0.054	0.084	0.1	1	0.072
TenYearCHD	0.22	-0.048	0.086	0.035	0.069	0.089	0.054	0.16	0.097	0.095	0.2	0.13	0.065	0.021	0.072	1

Feature engineering

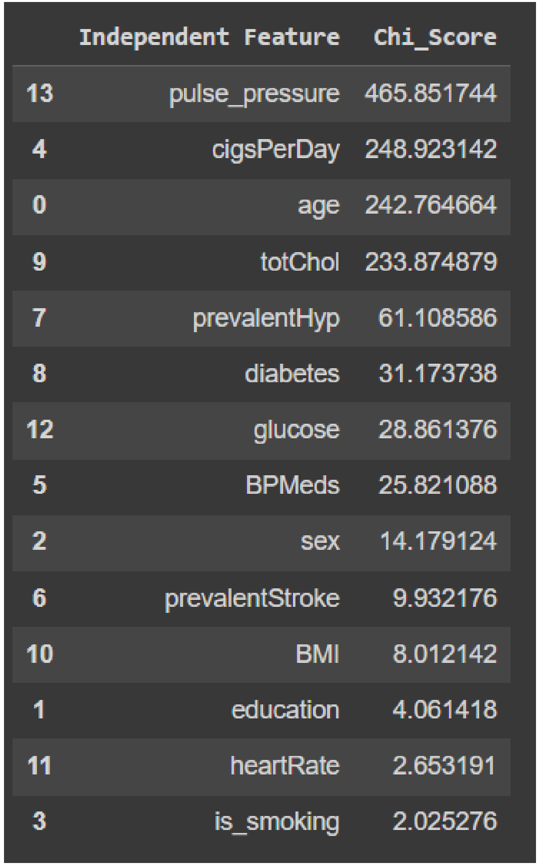
There is a high correlation between **sysBP (Systolic BP)** and **diaBP (Diastolic BP)**, and both of them influence our target variable to a greater extent, so we cannot drop them directly, but rather must find a parameter that can formulate these parameters together in such a way that we can add a single feature without experiencing multicollinearity or **pulse pressure**.

Pulse Pressure = Systolic BP - Diastolic BP



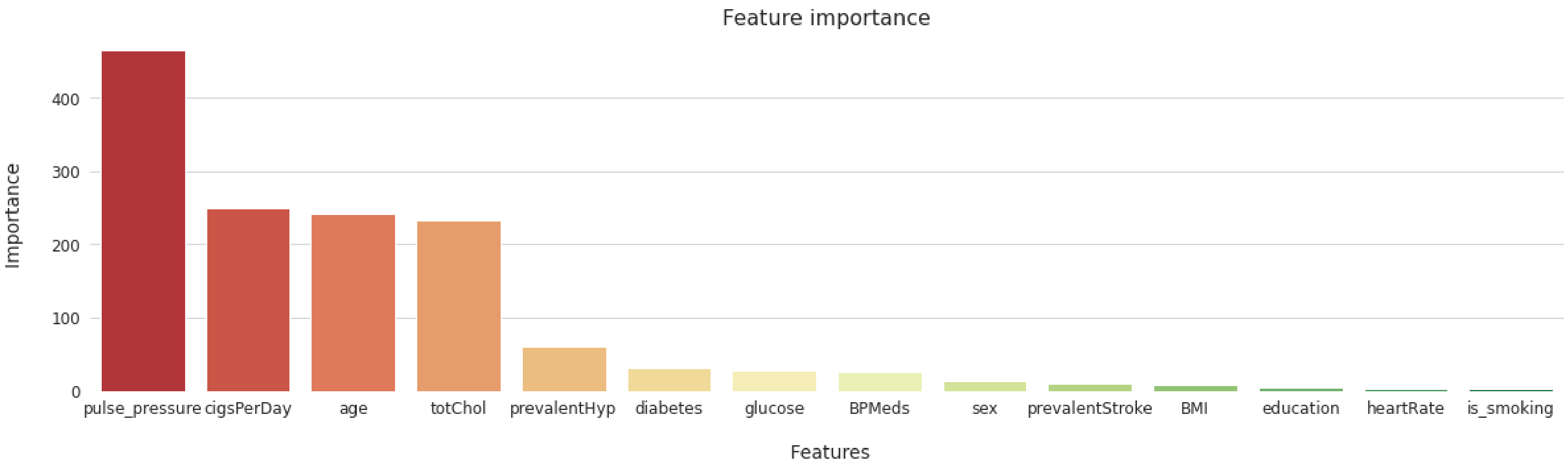
Feature engineering

- ❑ **Feature selection** is the process of reducing the number of input variables when developing a predictive model.
- ❑ It is desirable to reduce the number of input variables to both reduce the computational cost of modelling and, in some cases, to improve the performance of the model.
- ❑ In this model we using **Chi-Square** test for selecting the features that influence the most.



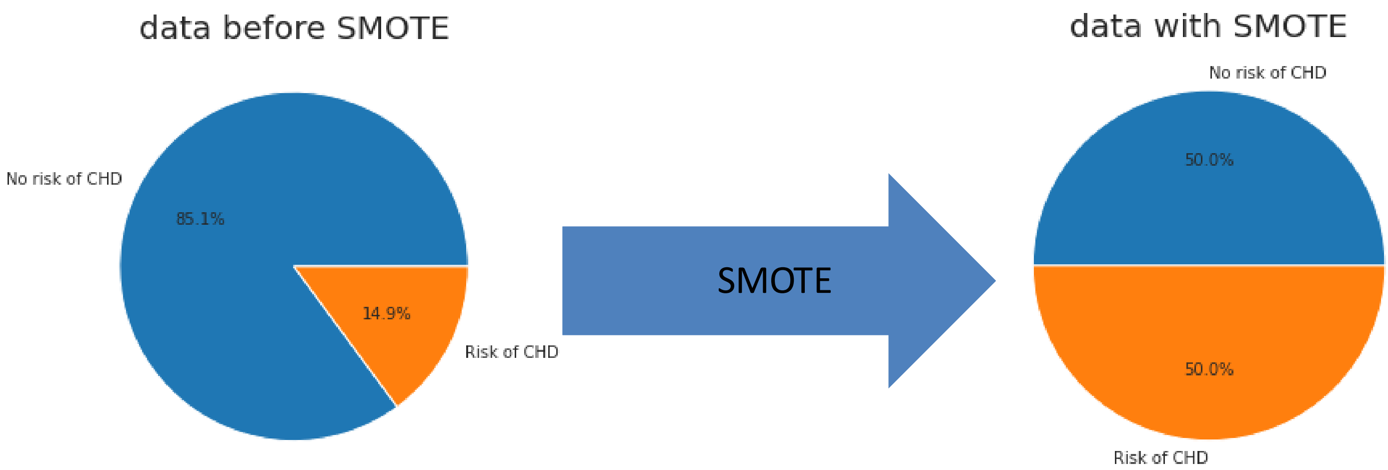
	Independent Feature	Chi_Score
13	pulse_pressure	465.851744
4	cigsPerDay	248.923142
0	age	242.764664
9	totChol	233.874879
7	prevalentHyp	61.108586
8	diabetes	31.173738
12	glucose	28.861376
5	BPMeds	25.821088
2	sex	14.179124
6	prevalentStroke	9.932176
10	BMI	8.012142
1	education	4.061418
11	heartRate	2.653191
3	is_smoking	2.025276

Feature engineering



❑ we observe **BMI**, **education**, **heartrate**, **sex** and **is smoking** very less chi2 score. hence remove those columns.

Handling Imbalanced Data



data before SMOTE

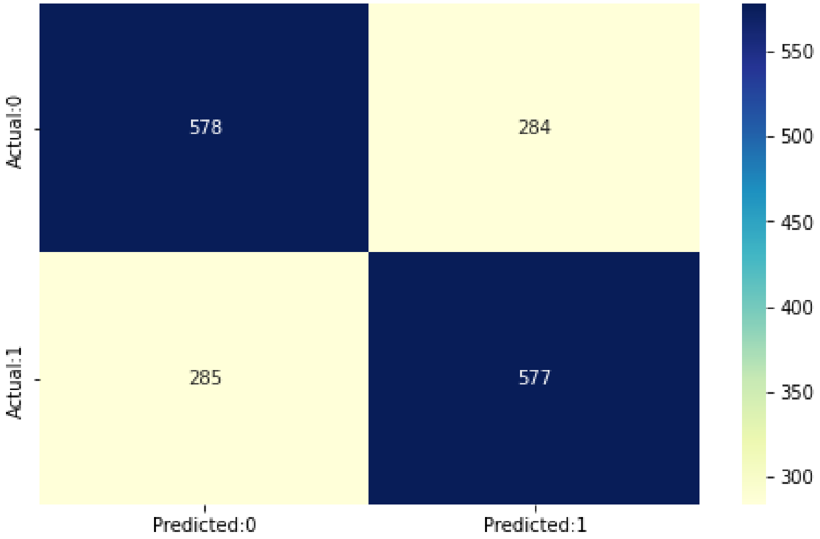
data with SMOTE

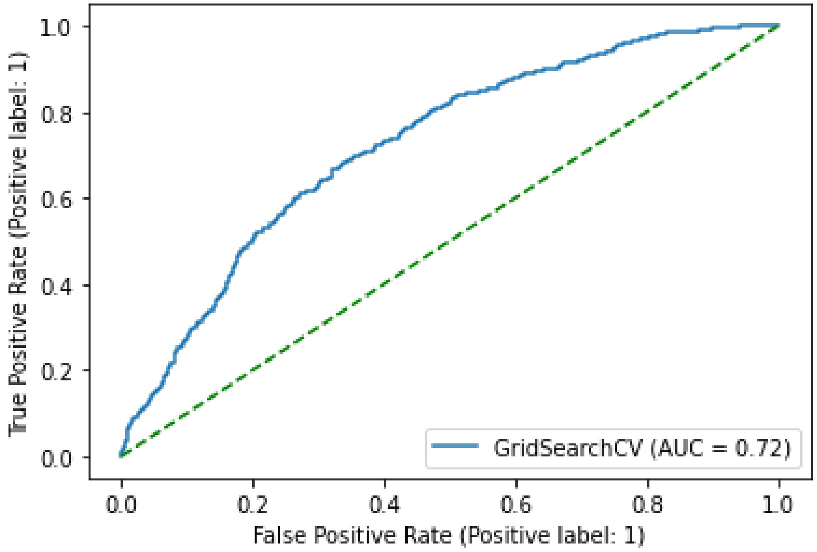
❑ Since our dataset is imbalanced, with more negative cases than positive cases, we may end up with a classifier that is biased towards the negative cases. The classifier may have high accuracy, but poor precision and recall.

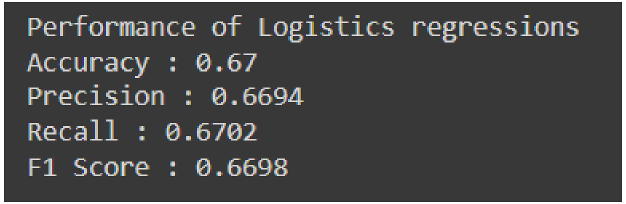
❑ **We have successfully oversampled the minority class using SMOTE. Now, the model we build will be able to learn from both classes without any bias.**

Machine learning model

Logistics Regression

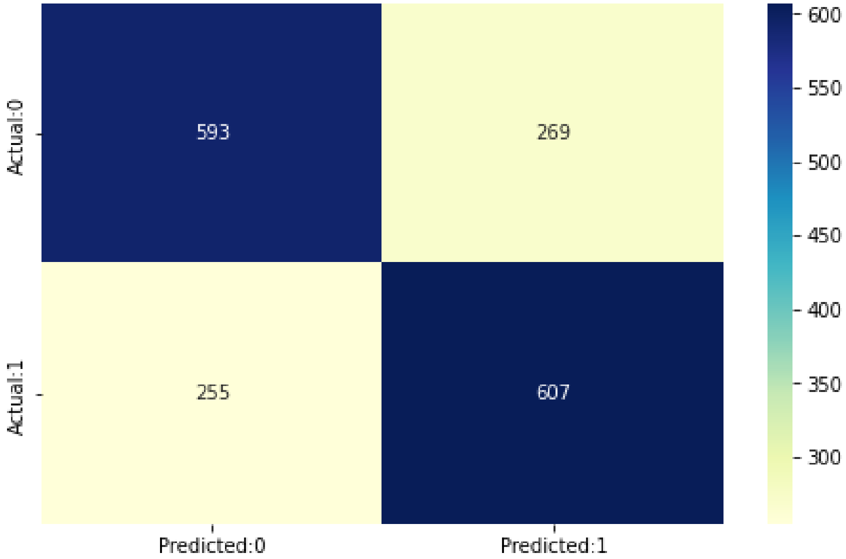


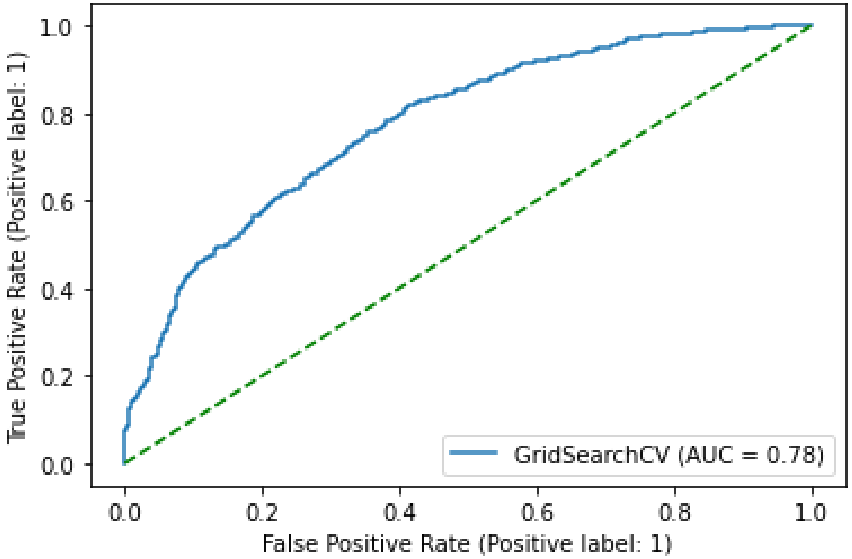


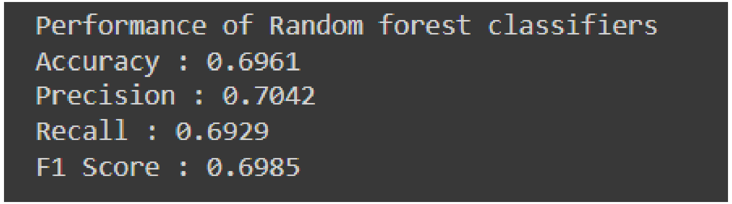


```
Performance of Logistics regressions
Accuracy : 0.67
Precision : 0.6694
Recall : 0.6702
F1 Score : 0.6698
```

Random Forest

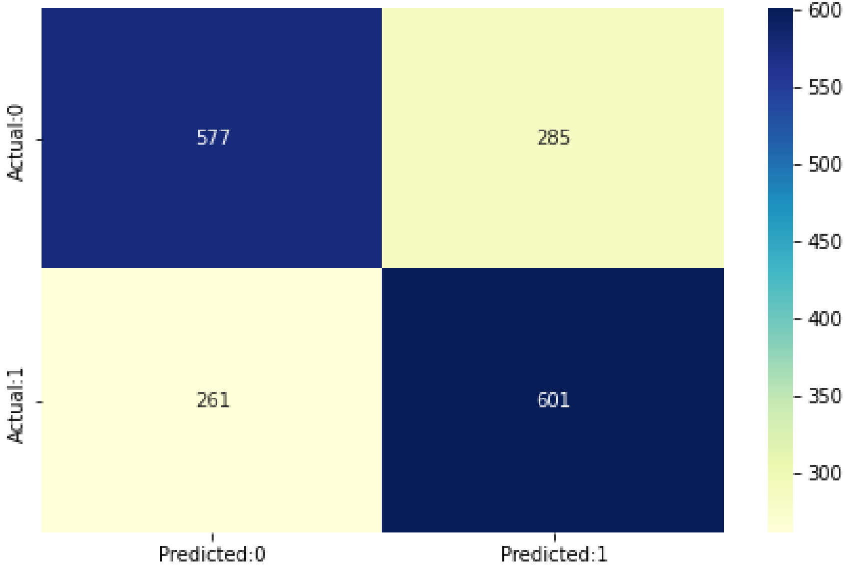


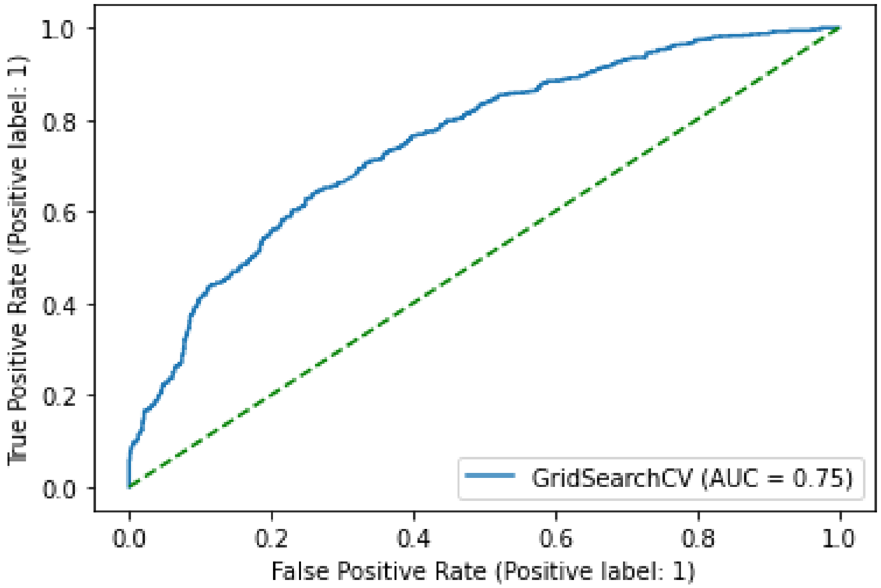


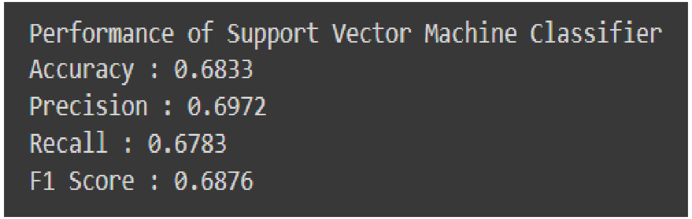


```
Performance of Random forest classifiers
Accuracy : 0.6961
Precision : 0.7042
Recall : 0.6929
F1 Score : 0.6985
```

Support Vector Machine(SVM)

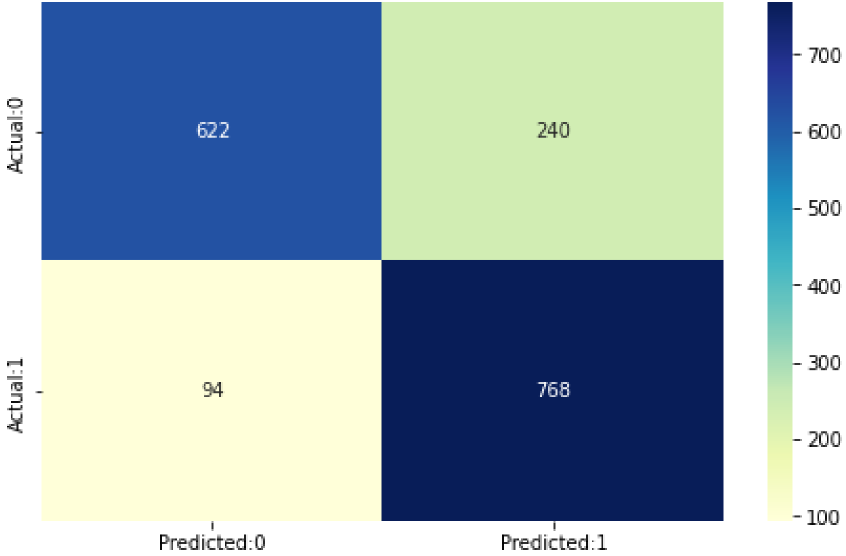


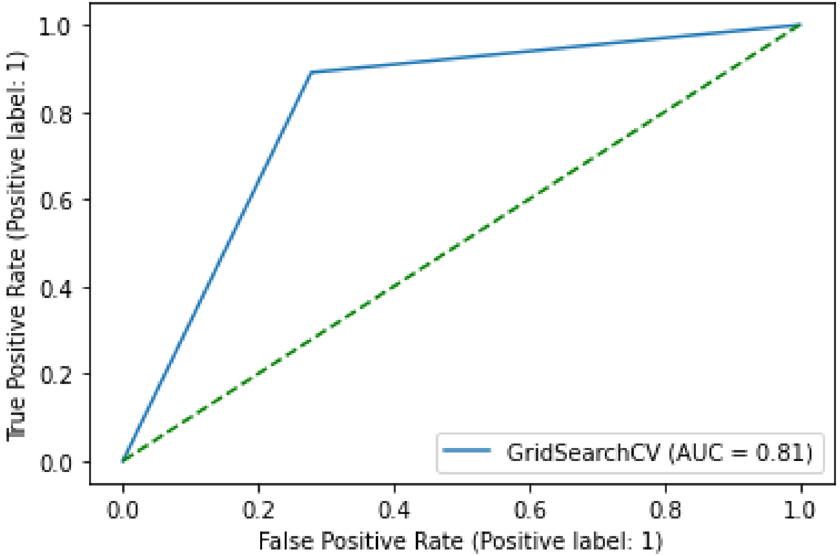


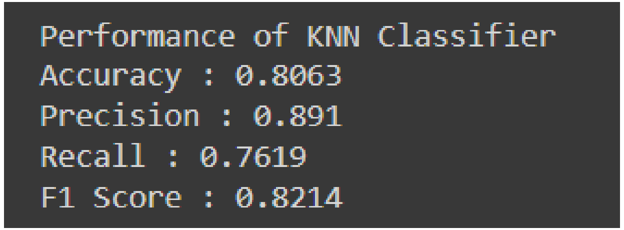


```
Performance of Support Vector Machine Classifier
Accuracy : 0.6833
Precision : 0.6972
Recall : 0.6783
F1 Score : 0.6876
```

KNN

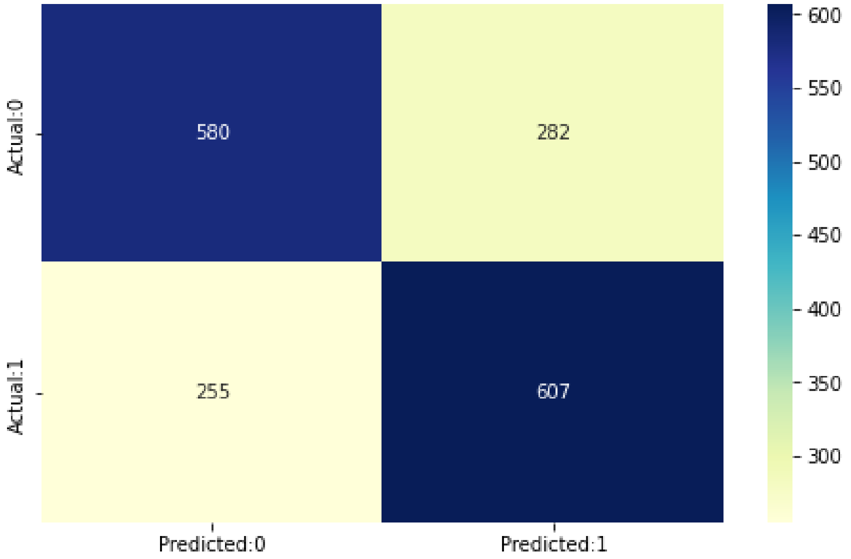


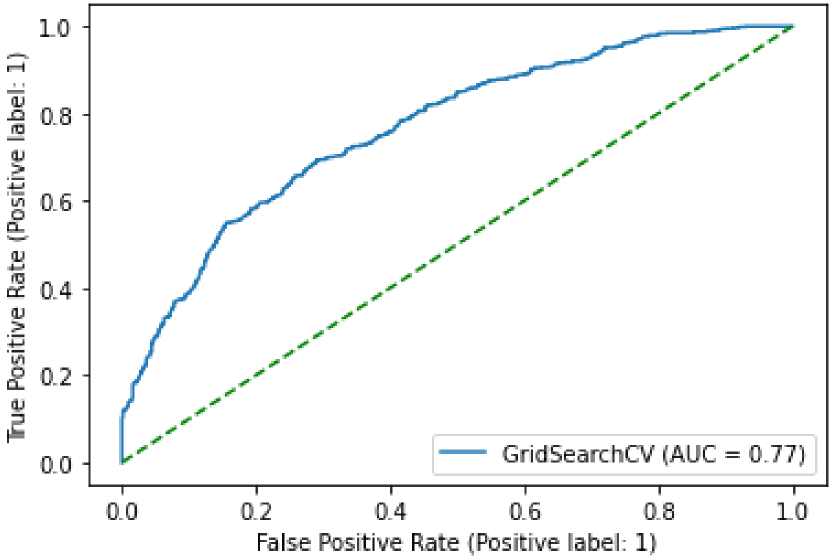


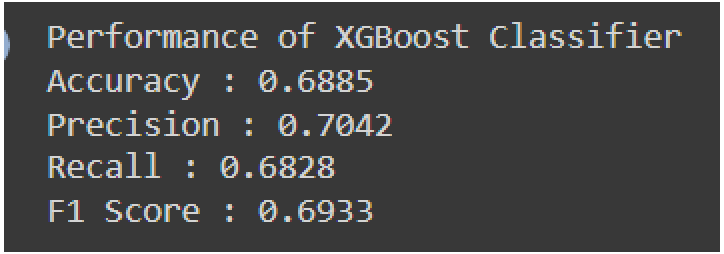


```
Performance of KNN Classifier
Accuracy : 0.8063
Precision : 0.891
Recall : 0.7619
F1 Score : 0.8214
```

XGBoost

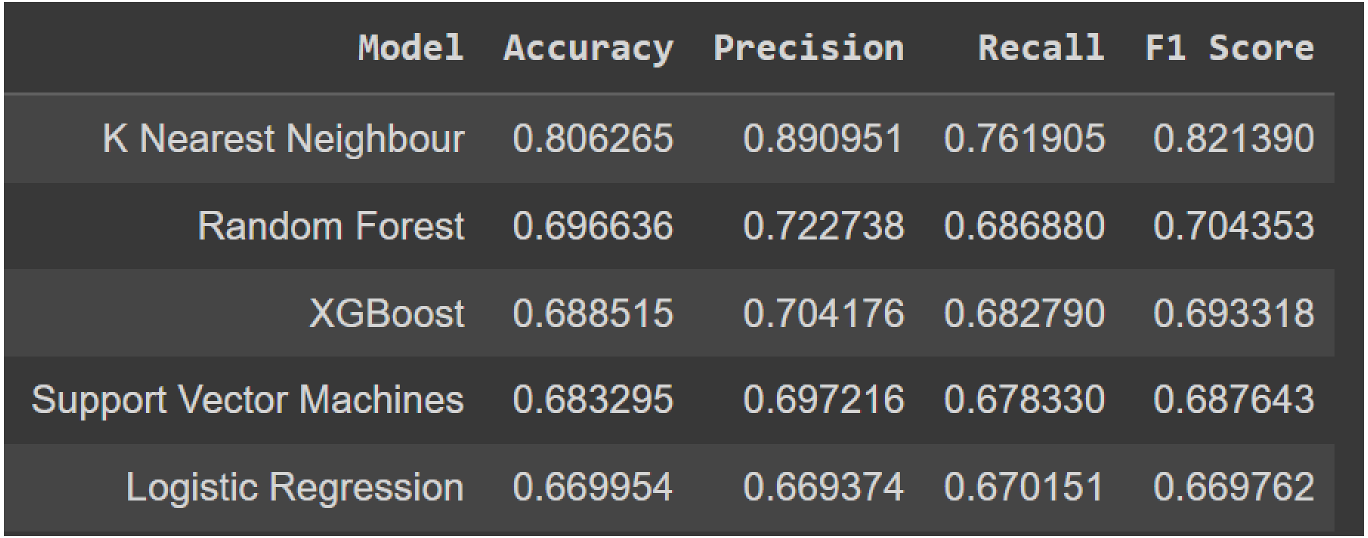






```
Performance of XGBoost Classifier
Accuracy : 0.6885
Precision : 0.7042
Recall : 0.6828
F1 Score : 0.6933
```

Model comparison



Model	Accuracy	Precision	Recall	F1 Score
K Nearest Neighbour	0.806265	0.890951	0.761905	0.821390
Random Forest	0.696636	0.722738	0.686880	0.704353
XGBoost	0.688515	0.704176	0.682790	0.693318
Support Vector Machines	0.683295	0.697216	0.678330	0.687643
Logistic Regression	0.669954	0.669374	0.670151	0.669762

❑ **The K Nearest Neighbour is proved to be best accuracy (80%), it can be used for risk prediction of Cardiovascular heart disease.**

Conclusion

- we trained 5 Machine Learning models, and hyperparameter adjustment was utilised models to increase model performance.
- The training dataset was oversampled using SMOTE to reduce bias on one outcome, missing values were handled, feature engineering, and feature selection were performed.
- Cardiovascular heart disease affects a similar number of smokers and non-smokers.
- **Age**, **total cholesterol**, **systolic blood** and **diastolic blood pressure**, **BMI**, **heart rate**, and **glucose** are the main factors in determining a person's 10-year chance of having cardiovascular heart disease.
- **The K Nearest Neighbour is proved to be best algorithms can be used for the risk prediction of Cardiovascular heart disease.**
- We chose the oversampling technique because the data provided to us had fewer records. But since there will be a lot of unbalanced and large amounts of health data, we can try to work on cost-sensitive learning, which, rather than changing the data records, only gives more weight to the minority and focuses on the individuals at high risk for heart disease.

QnA

Thank you