CAPSTONE PROJECT

CLASSIFICATION MODEL

TITLE: Cardiovascular Risk Prediction

Prepared By → PRAVEEN.S

(COHORT EVEREST)

AI

Let's go through the Defaulters:

1. Defining the Problem Statement
2. EDA and feature selection
3. Data Visualization
4. Normalization of data
5. Preparing dataset for modeling
6. Applying the model
7. Cross Validation and Hyperparamter Tuning

Problem Statement:

Coronary heart disease (CHD) is the most common type of heart disease.It's known as hardening of the arteries, too. Arteries carry blood and oxygen to your heart.This 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 4,000 records and 15 attributes.The main Objective of this model is to predict the risk of getting CHD in the upcoming 10 years.

Dataset Understanding:

Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.They are

- **Sex:** male or female
- **Age:** Age of the patient
- **is_smoking:** whether or not the patient is a current smoker
- **Cigs Per Day**: the number of cigarettes that the person smoked on average in one day
- **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)

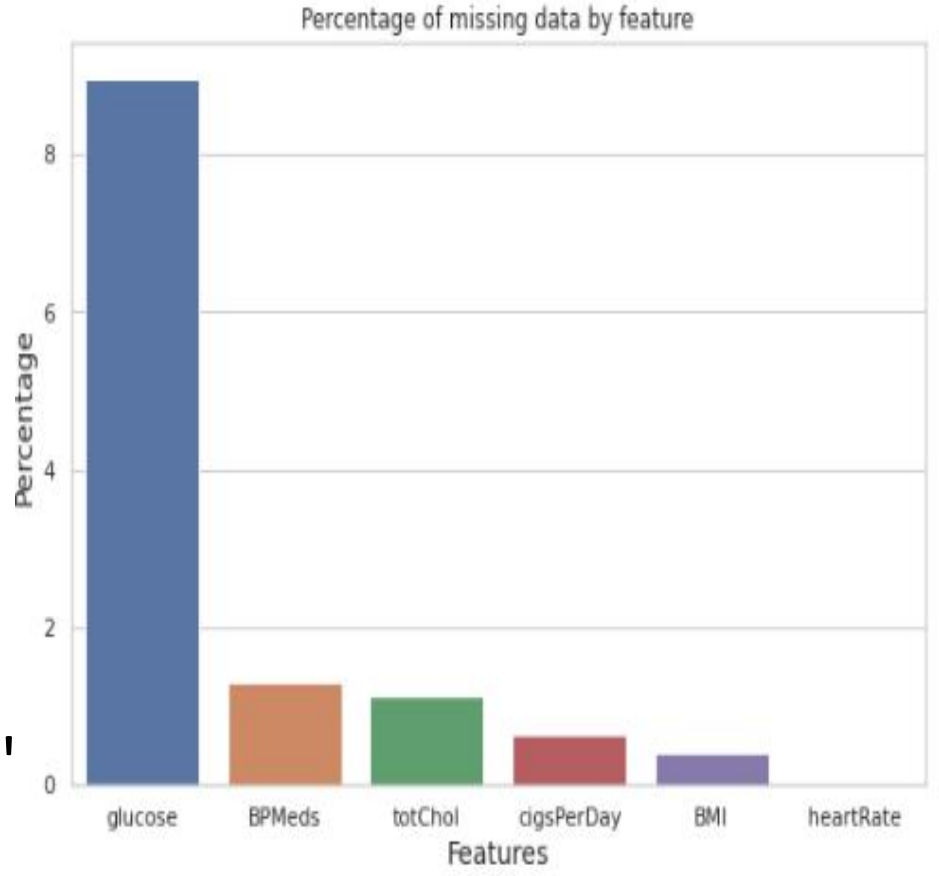
Data Understanding:

- **Prevalent Hyp:** whether or not the patient was hypertensive
- **Tot Chol:** total cholesterol level
- **Sys BP:** systolic blood pressure
- **Dia BP:** diastolic blood pressure
- **BMI:** Body Mass Index
- **Heart Rate:** heart rate
- **Glucose:** glucose level

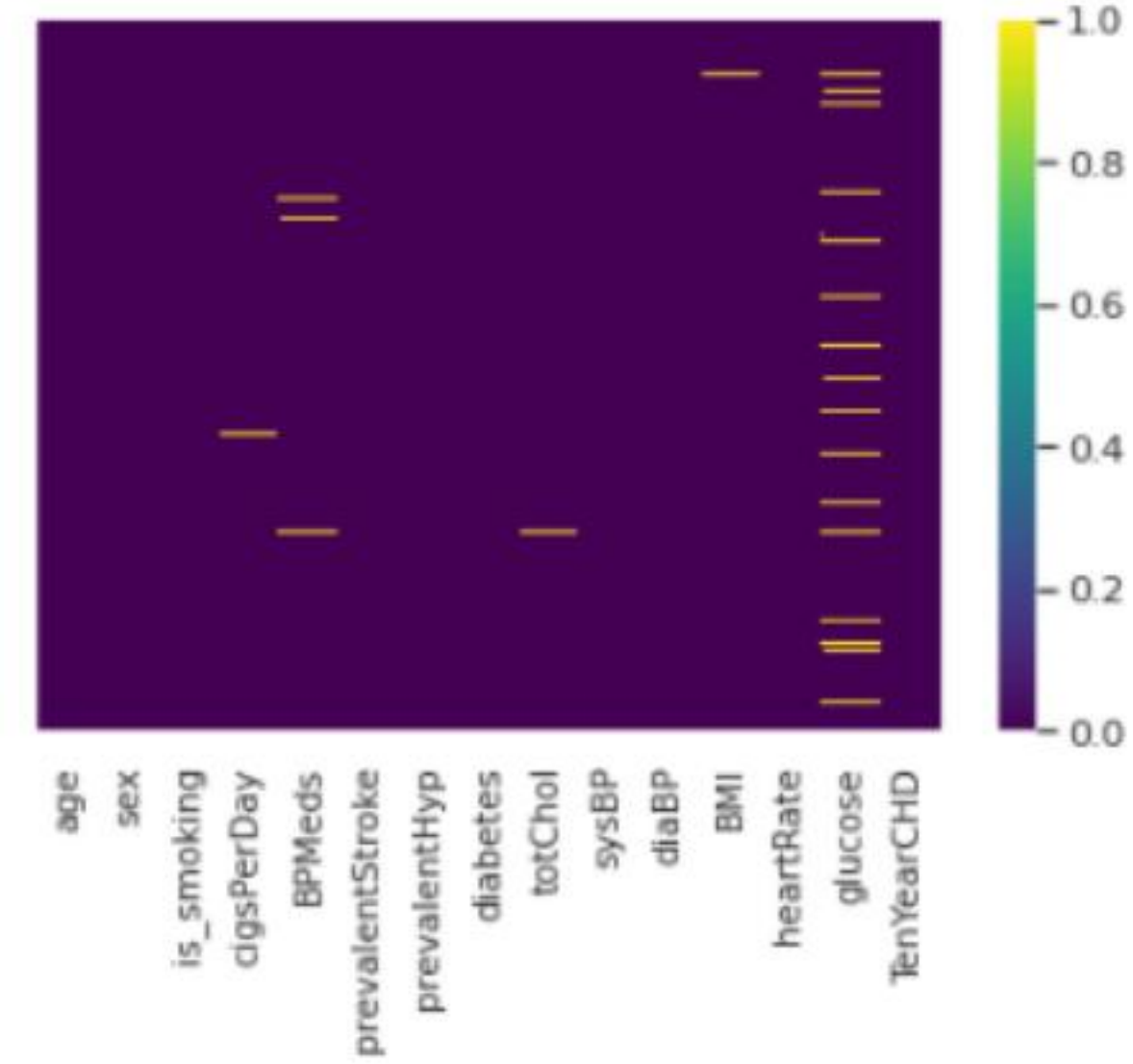
- **10-year risk of coronary heart disease CHD**(binary: “1”, means “Yes”, “0” means “No”)

Data Preprocessing:

- Dropped the “id” and “education” column, As the target value is not dependent of it
- Converted Nominal values such as Yes and No as well as Male and Female into Binary Form
- There were several null values in the dataset especially in the continous value features such as"age","cigsPerDay","totChol","sysBP","diaBP" "BMI","heartRate","glucose".

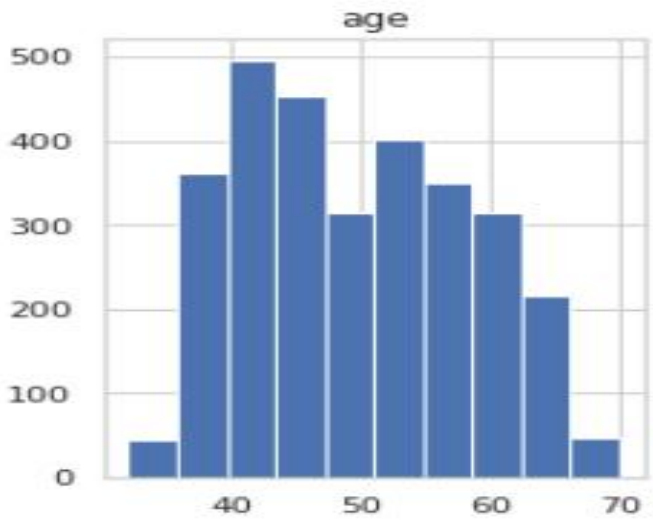


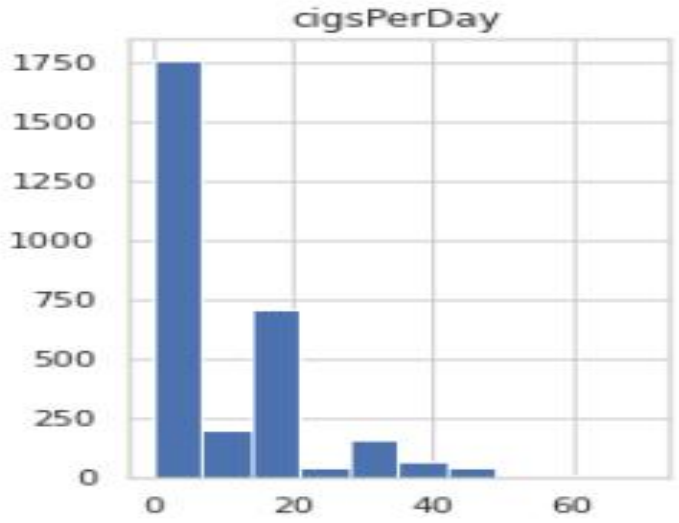
Data Preprocessing:

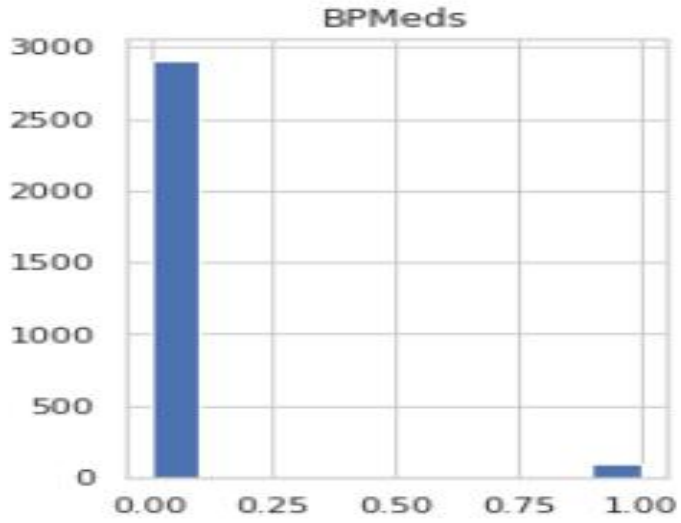


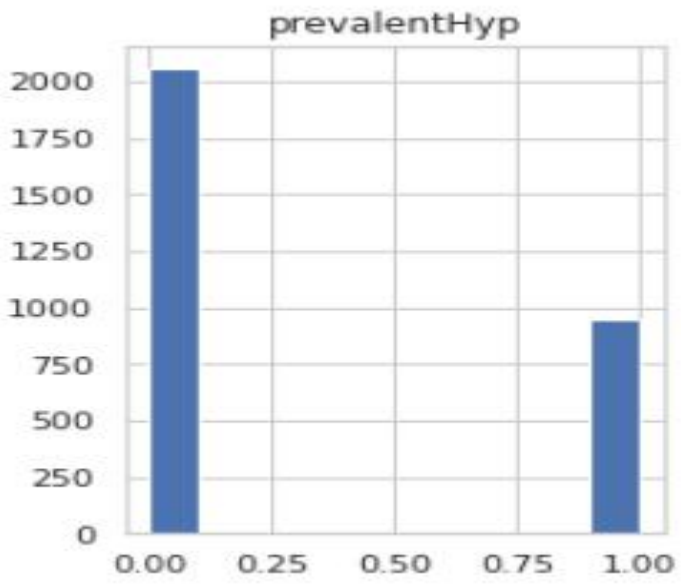
- 12% of the dataset is found to be null values and we have dropped these rows since it is on the low margin side owing to the original dataset.

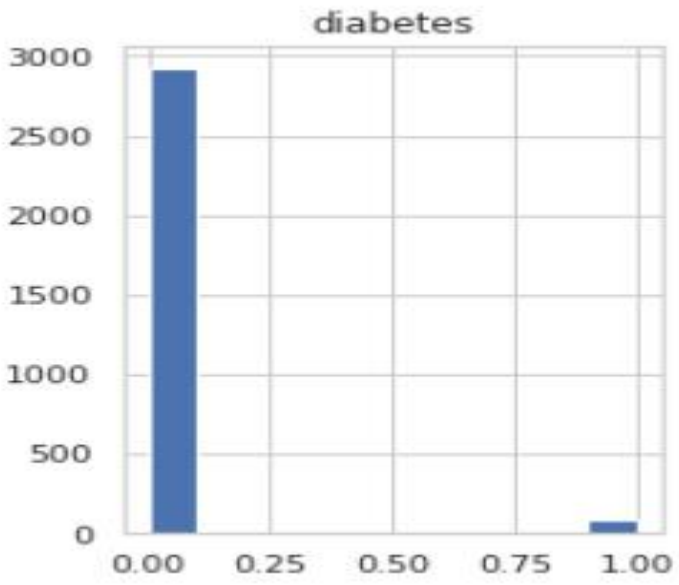
Data Visualization:

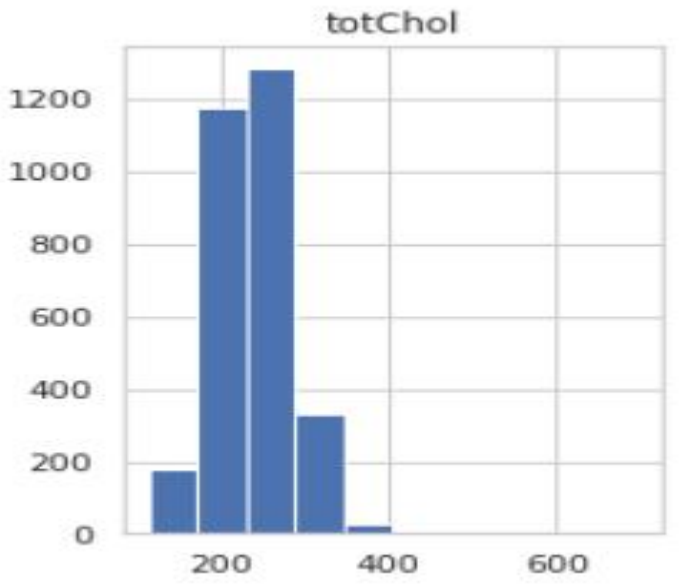




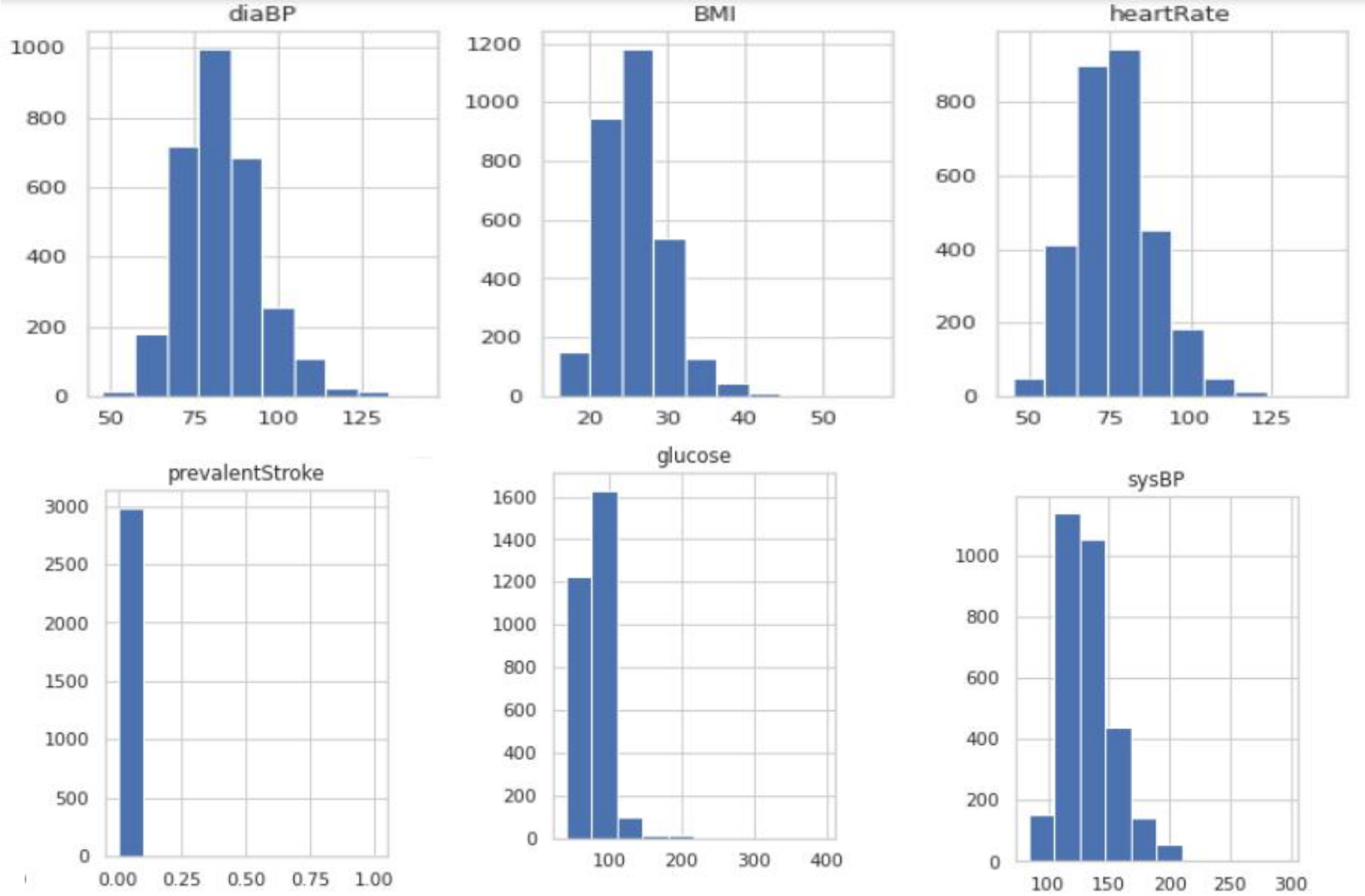






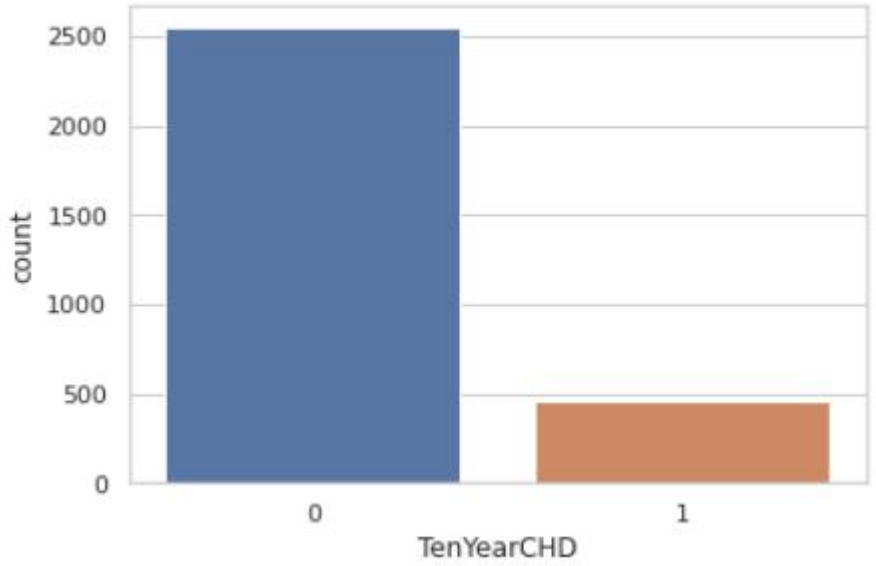


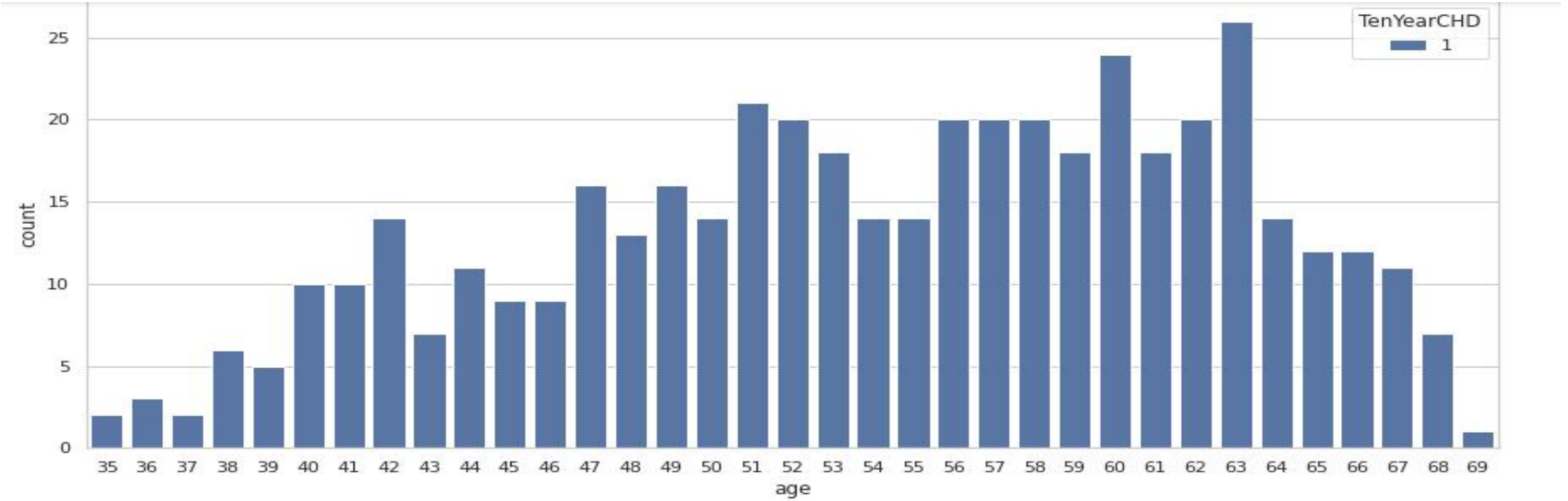
Data Visualization:



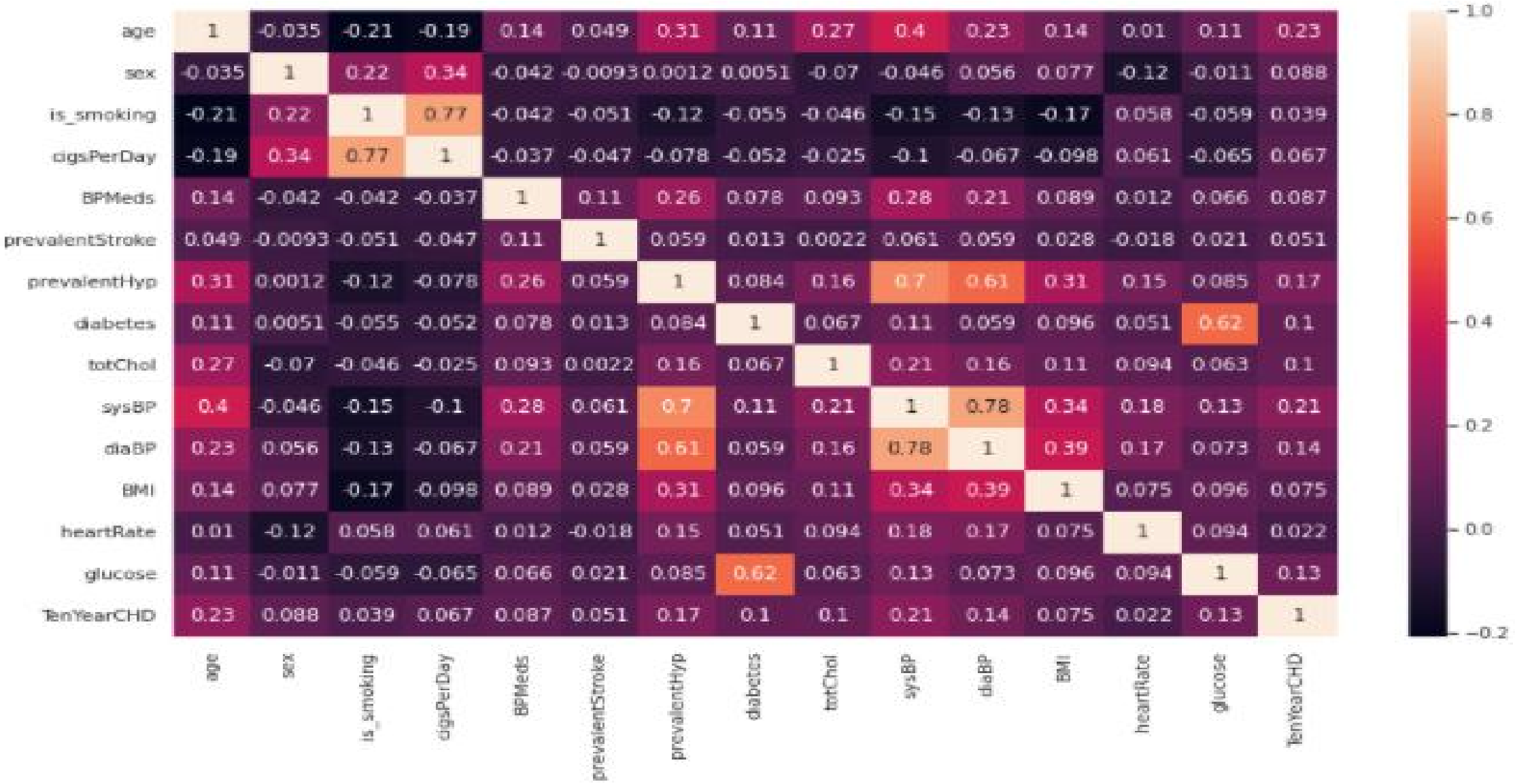
Data Visualization:

Target Variable Distribution





Data Visualization:



Feature Selection:

Using Chisquare stat model:

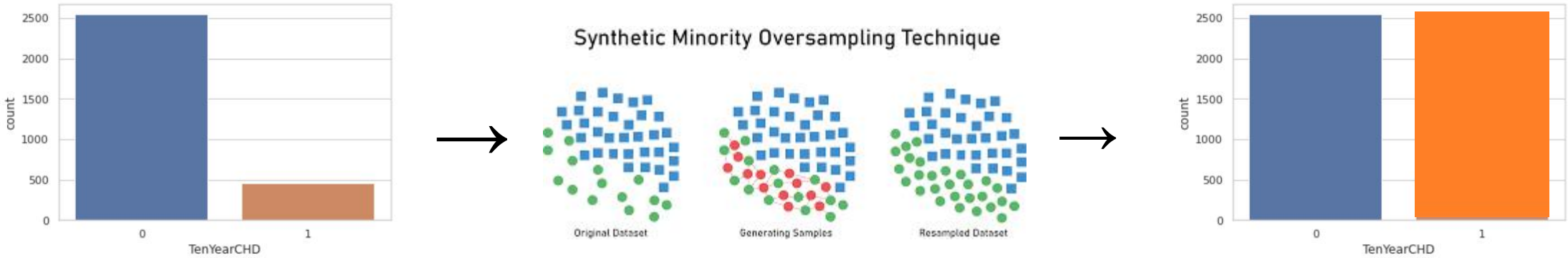
	coef	std err	z	P>|z|	[0.025	0.975]
const	-8.7645	0.761	-11.522	0.000	-10.255	-7.274
age	0.0648	0.007	8.861	0.000	0.050	0.079
sex	0.5148	0.121	4.260	0.000	0.278	0.752
is_smoking	0.2154	0.172	1.252	0.211	-0.122	0.553
cigsPerDay	0.0184	0.007	2.702	0.007	0.005	0.032
BPMeds	0.1011	0.263	0.384	0.701	-0.414	0.616
prevalentStroke	0.9346	0.525	1.779	0.075	-0.095	1.964
prevalentHyp	0.1713	0.153	1.119	0.263	-0.129	0.471
diabetes	-0.0983	0.352	-0.279	0.780	-0.789	0.593
totChol	0.0031	0.001	2.627	0.009	0.001	0.005
sysBP	0.0166	0.004	3.975	0.000	0.008	0.025
diaBP	-0.0079	0.007	-1.139	0.255	-0.022	0.006
BMI	0.0081	0.014	0.584	0.559	-0.019	0.035
heartRate	-0.0036	0.005	-0.786	0.432	-0.013	0.005
glucose	0.0094	0.003	3.749	0.000	0.004	0.014

Using Backward Selection:

	coef	std err	z	P>|z|	[0.025	0.975]
const	-9.2907	0.524	-17.735	0.000	-10.317	-8.264
age	0.0664	0.007	9.313	0.000	0.052	0.080
sex	0.5120	0.118	4.326	0.000	0.280	0.744
cigsPerDay	0.0237	0.005	5.119	0.000	0.015	0.033
totChol	0.0031	0.001	2.586	0.010	0.001	0.005
sysBP	0.0159	0.002	6.719	0.000	0.011	0.021
glucose	0.0090	0.002	4.828	0.000	0.005	0.013

Handling Class Imbalances:

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling.. It potentially performs better than simple oversampling and it is widely used.



Preparing Dataset for Modeling:

	age	sex	cigsPerDay	totChol	sysBP	glucose	TenYearCHD
0	36	1	0.000000	212.000000	168.000000	75.000000	0
1	46	0	10.000000	250.000000	116.000000	94.000000	0
2	50	1	20.000000	233.000000	158.000000	94.000000	1
3	64	0	30.000000	241.000000	136.500000	77.000000	0
4	61	0	0.000000	272.000000	182.000000	65.000000	1
...	...	...	...	...	...	...	...
5089	56	0	0.000000	266.050846	150.847463	108.440699	1
5090	56	0	4.709884	288.822529	188.061425	87.061425	1
5091	55	0	3.653391	217.219130	159.956174	71.872520	1
5092	63	0	1.113869	293.668321	115.805073	87.612628	1
5093	54	0	20.000000	324.356898	159.785633	87.500571	1

train_test_split:

X_train:-(4075, 6)

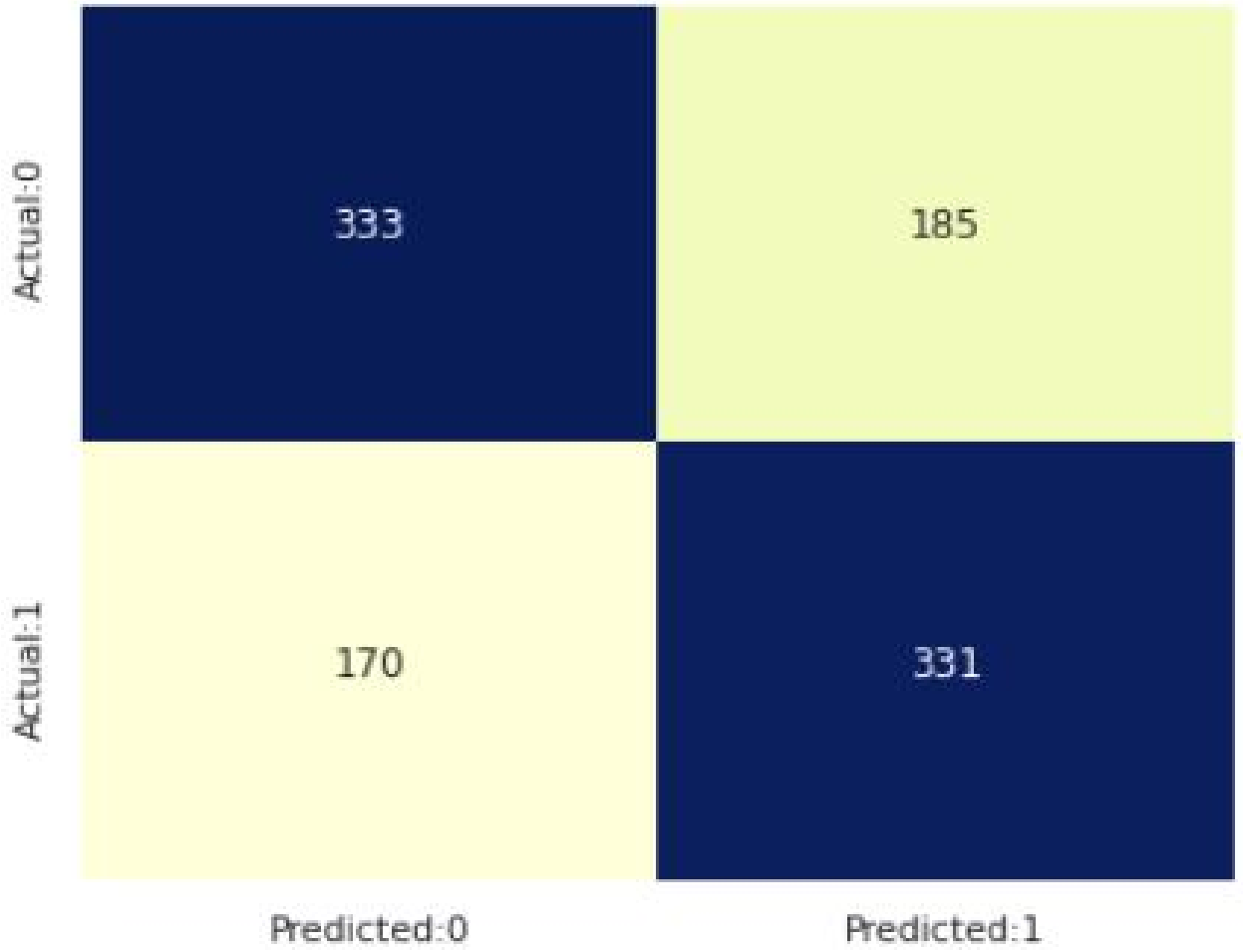
X_test:-(1019, 6)

y_train:-(4075,1)

y_test:-(1019,1)

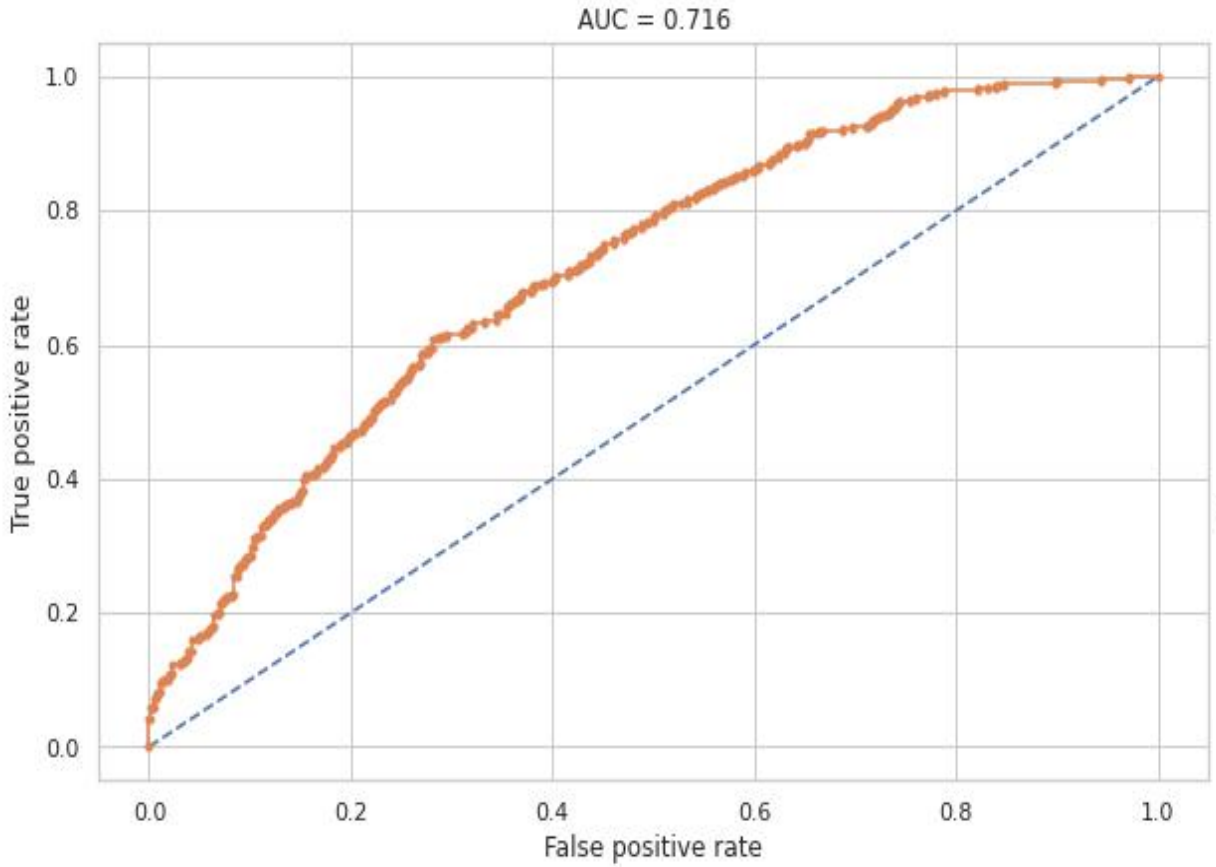
Baseline model:

Logistic Regression:



	precision	recall	f1-score	support
0	0.66	0.64	0.65	518
1	0.64	0.66	0.65	501
accuracy			0.65	1019
macro avg	0.65	0.65	0.65	1019
weighted avg	0.65	0.65	0.65	1019

Model Validation and Selection:




```
GridSearchCV(cv=10, estimator=LogisticRegression(),
             param_grid={'C': [0.01, 0.1, 1, 10, 100],
                         'class_weight': ['balanced', None],
                         'penalty': ['l1', 'l2']})
```

```
{'C': 0.01, 'class_weight': 'balanced', 'penalty': 'l2'}
```

Observations:

Observation1:

As the dataset has so many challenges from null values to class imbalances,it ought to be implement using complex models as they consist of both nominal as well as continous values.

Observation2:

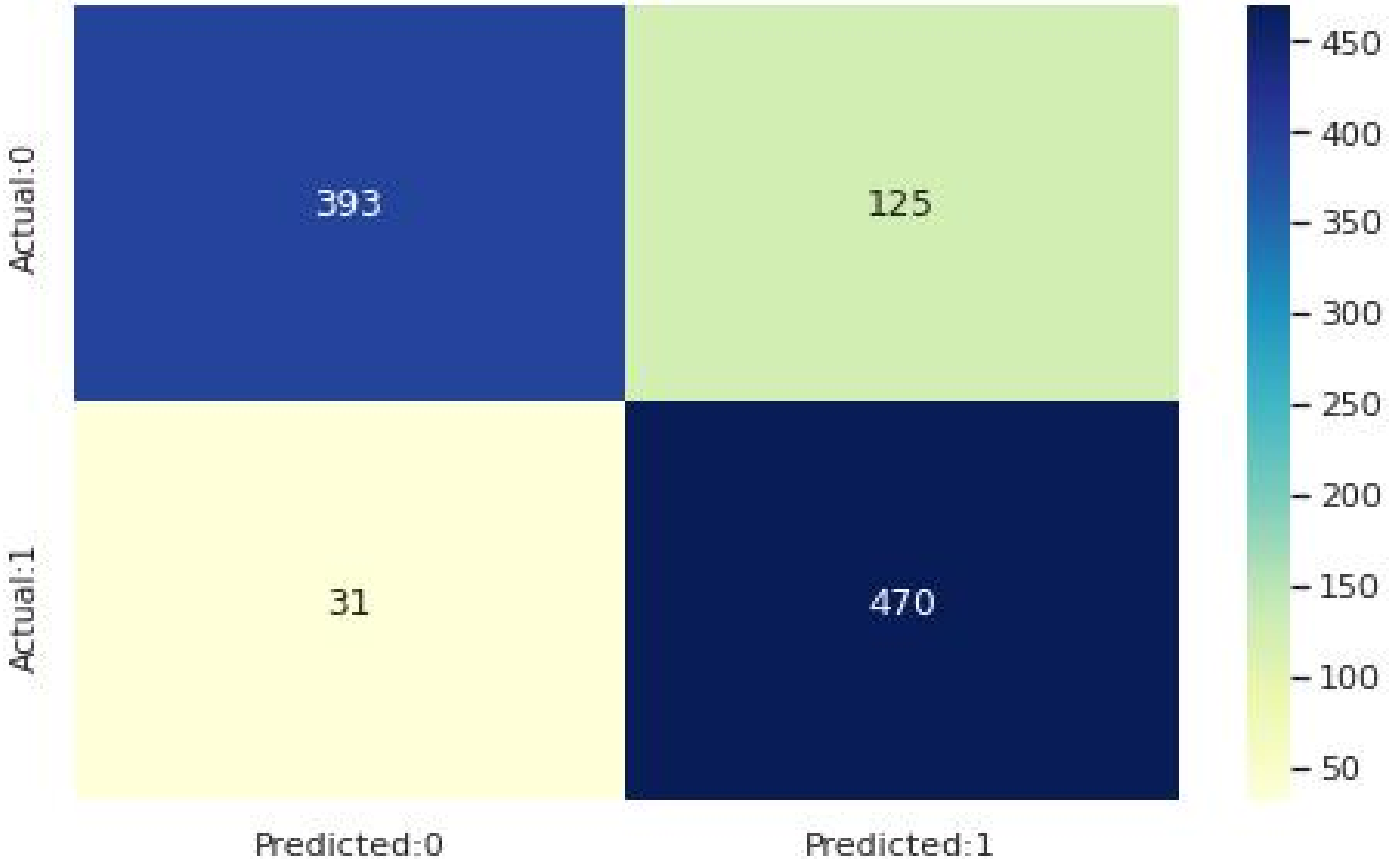
The least intepretable classification model which is the logistic regression had given an score of 65% which is acceptable but this model has to be supported by other models for greater accuracy.

Observation3:

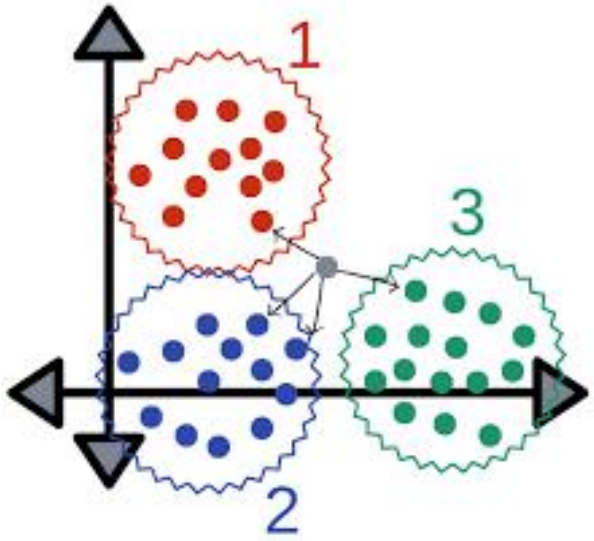
Using pycharet we found that the k-nearest neighbour and Support vector machines to be the perfect fit for this particular dataset and had potentially given greater results in the upcoming slides.

Model Validation and Selection:

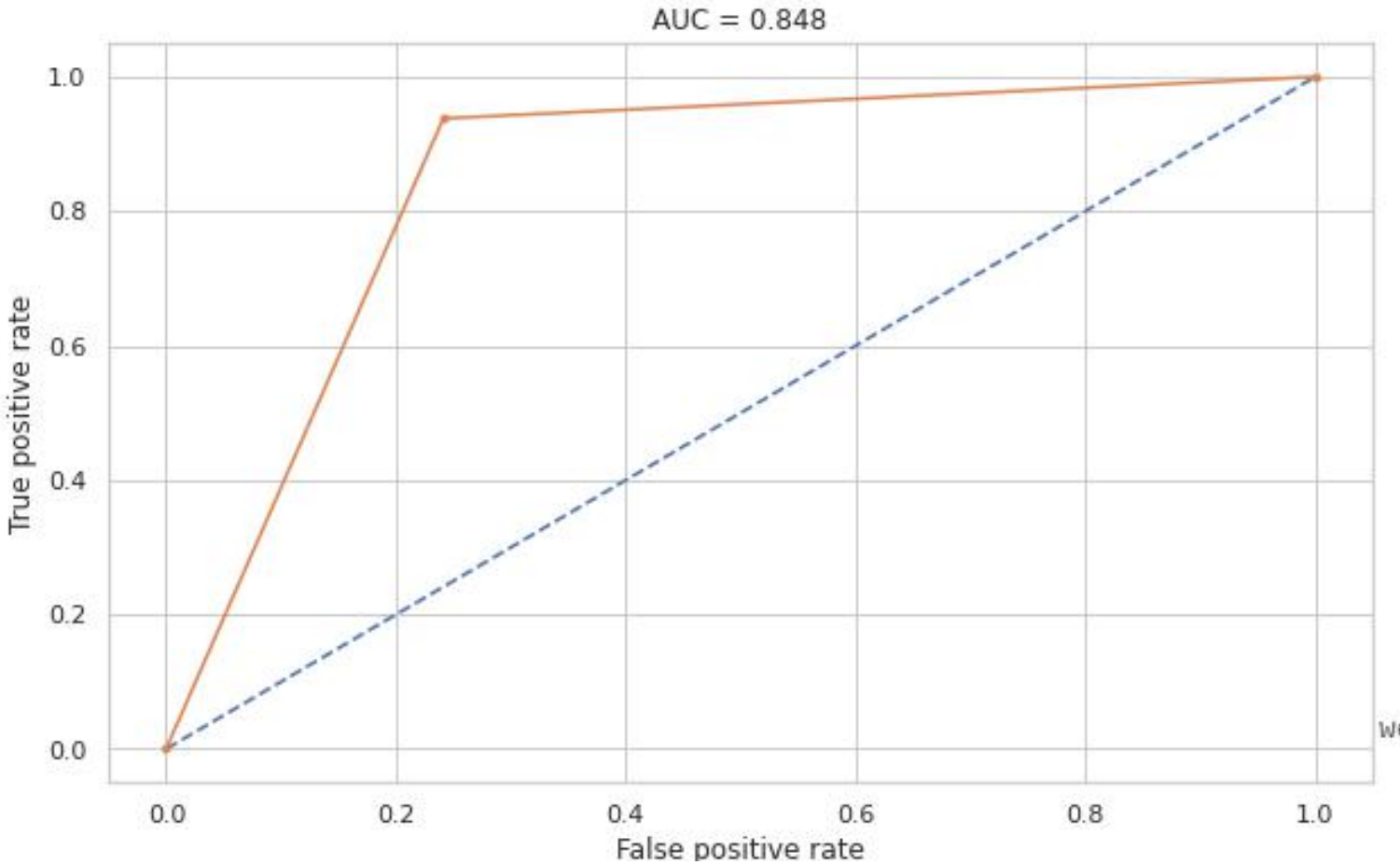
K Nearest Neighbours:




```
{'n_neighbors': 1}
```



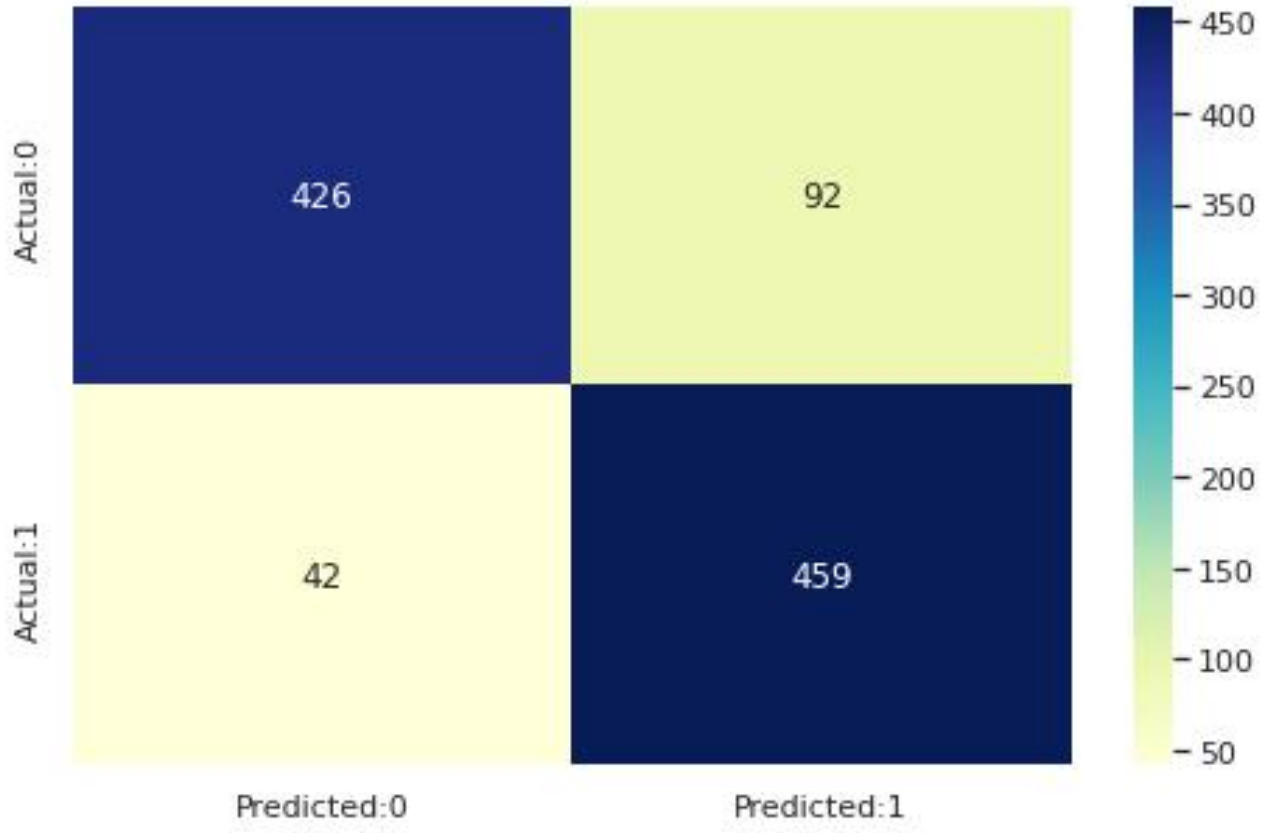
K-Nearest Neighbours:



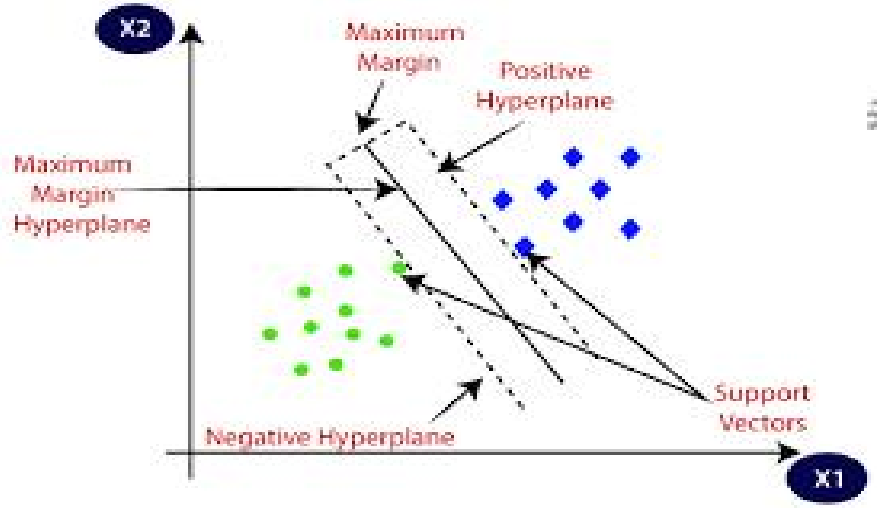
	precision	recall	f1-score	support
0	0.92	0.75	0.83	518
1	0.79	0.94	0.85	501
accuracy			0.84	1019
macro avg	0.85	0.84	0.84	1019
weighted avg	0.86	0.84	0.84	1019

Model Validation and Selection:

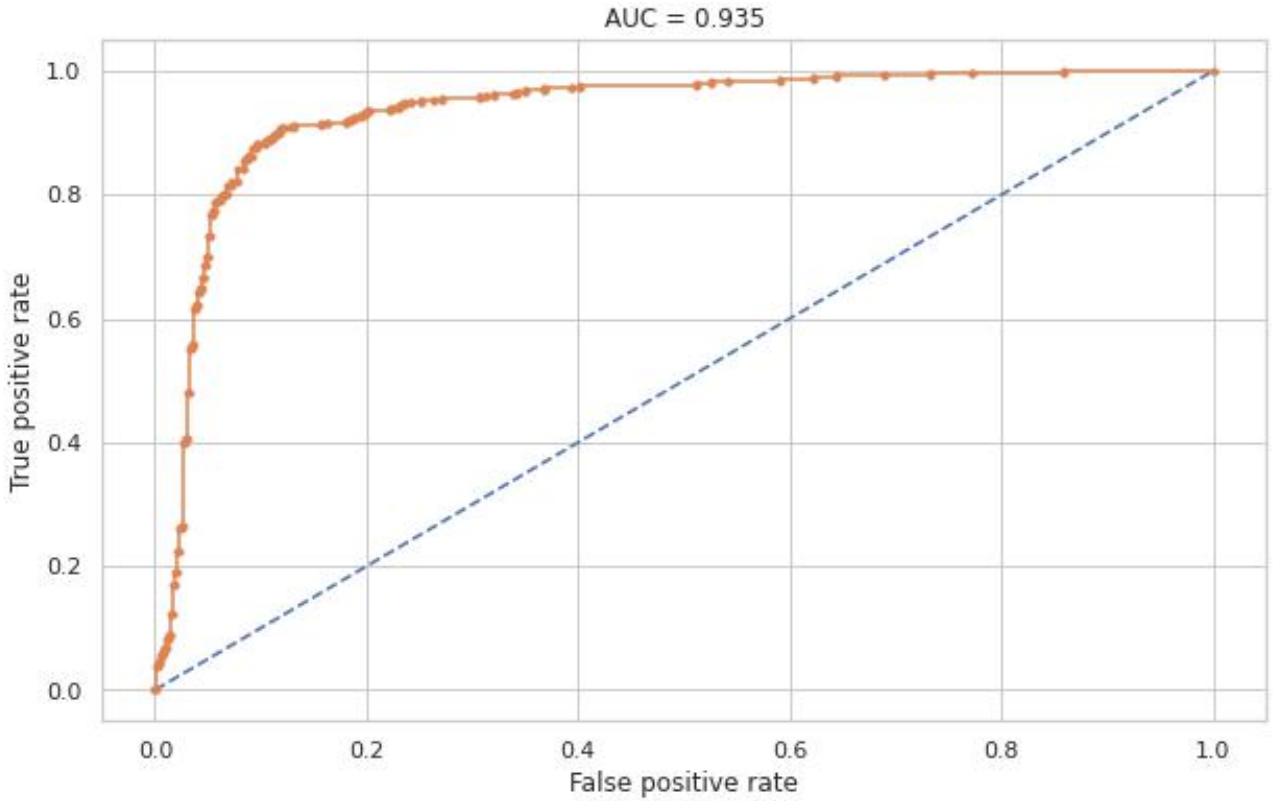
Support Vector Machine




```
{'C': 10, 'gamma': 0.01}
```



Support Vector Machine:



	precision	recall	f1-score	support
0	0.91	0.82	0.86	518
1	0.83	0.92	0.87	501
accuracy			0.87	1019
macro avg	0.87	0.87	0.87	1019
weighted avg	0.87	0.87	0.87	1019

Conclusion:

	Accuracy	AUC	F1 score
Logistic regression	0.658489	0.719912	0.661479
K-nearest neighbours	0.842983	0.844512	0.854281
Support vector machine	0.868499	0.934779	0.872624

- The most important features in predicting the ten year risk of developing CHD were age and systolic blood pressure
- The Support vector machine with the radial kernel was the best performing model in terms of accuracy and the F1 score. Its high AUC shows that it has a high true positive rate.
- Balancing the dataset by using the SMOTE technique helped in improving the models' sensitivity, this is when compared to the performance metrics of other models
- With more data(especially that of the minority class) better models can be built

Challenges:

- Selecting the perfect technique for Feature selection and handling class imbalances did took a lot of research and was time consuming.
- About 12% of the dataset was found to be consisting of null values which we had dropped could alter the accuracy of the models we have built.
- Training the Support vector model did take a lot of computational technique.

Thank You

That's a WRAP!!!