Capstone Project

Cardiovascular Disease Risk Prediction

Introduction

- In recent times, there has been an increase in the number of people suffering from cardiovascular diseases.
- 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.

Steps in the project

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

- The 5 steps are as follows :-

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

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, behavioural, 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.

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)

1. 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.

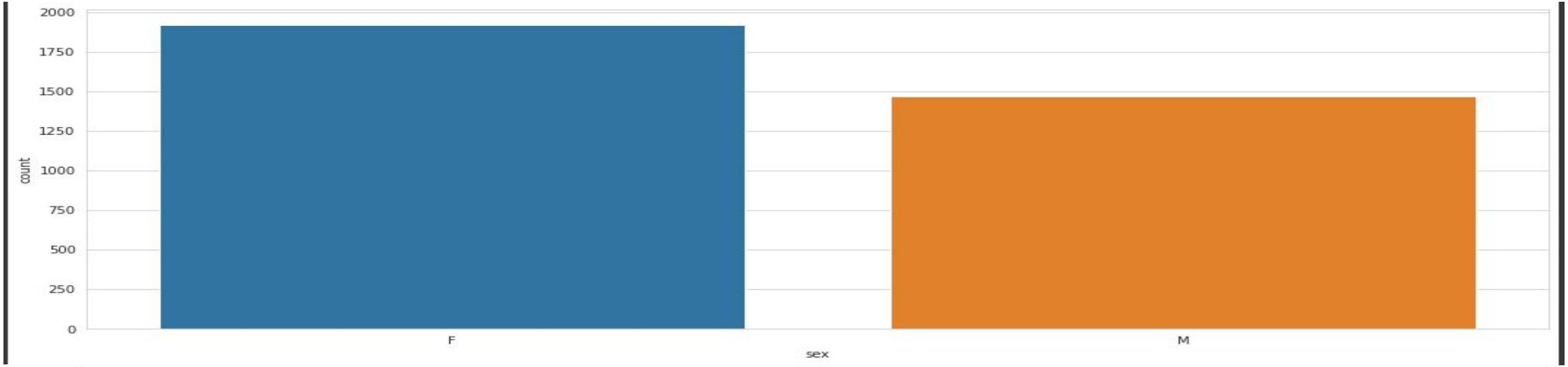
- Null values in the columns were filled with the mode and median for categorical and continuous variables respectively. Median was chosen due to outliers.

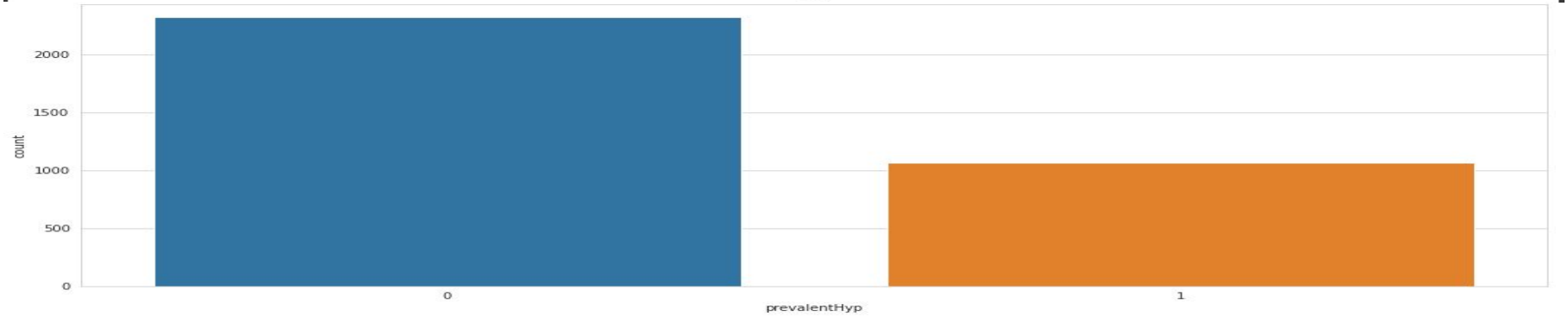
- Columns of smoking and cigarettes per day did not match in some cases, and these were corrected.

- 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.

2. Exploratory Data Analysis

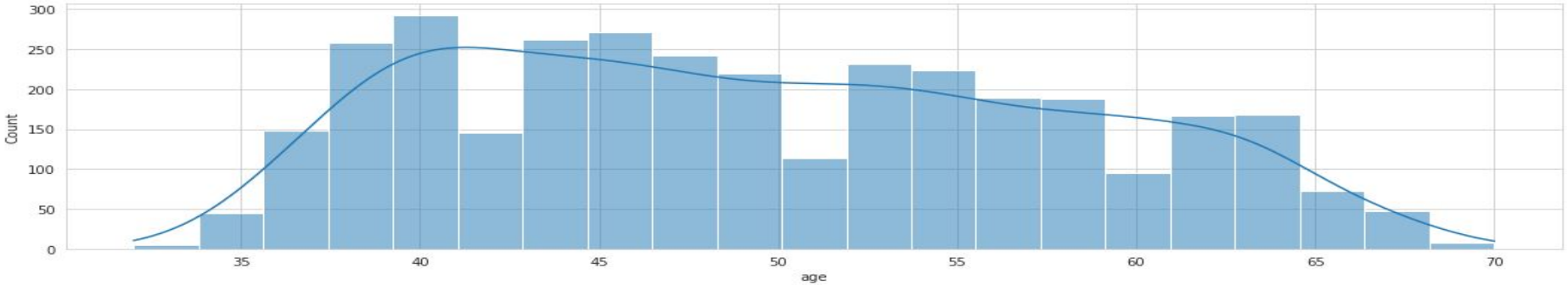
- Univariate analysis

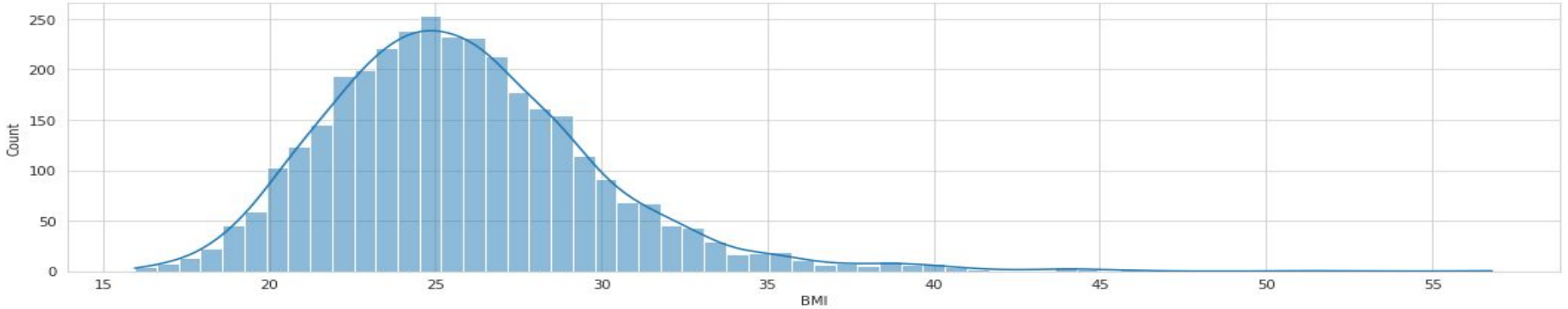




2. Exploratory Data Analysis

- Univariate analysis

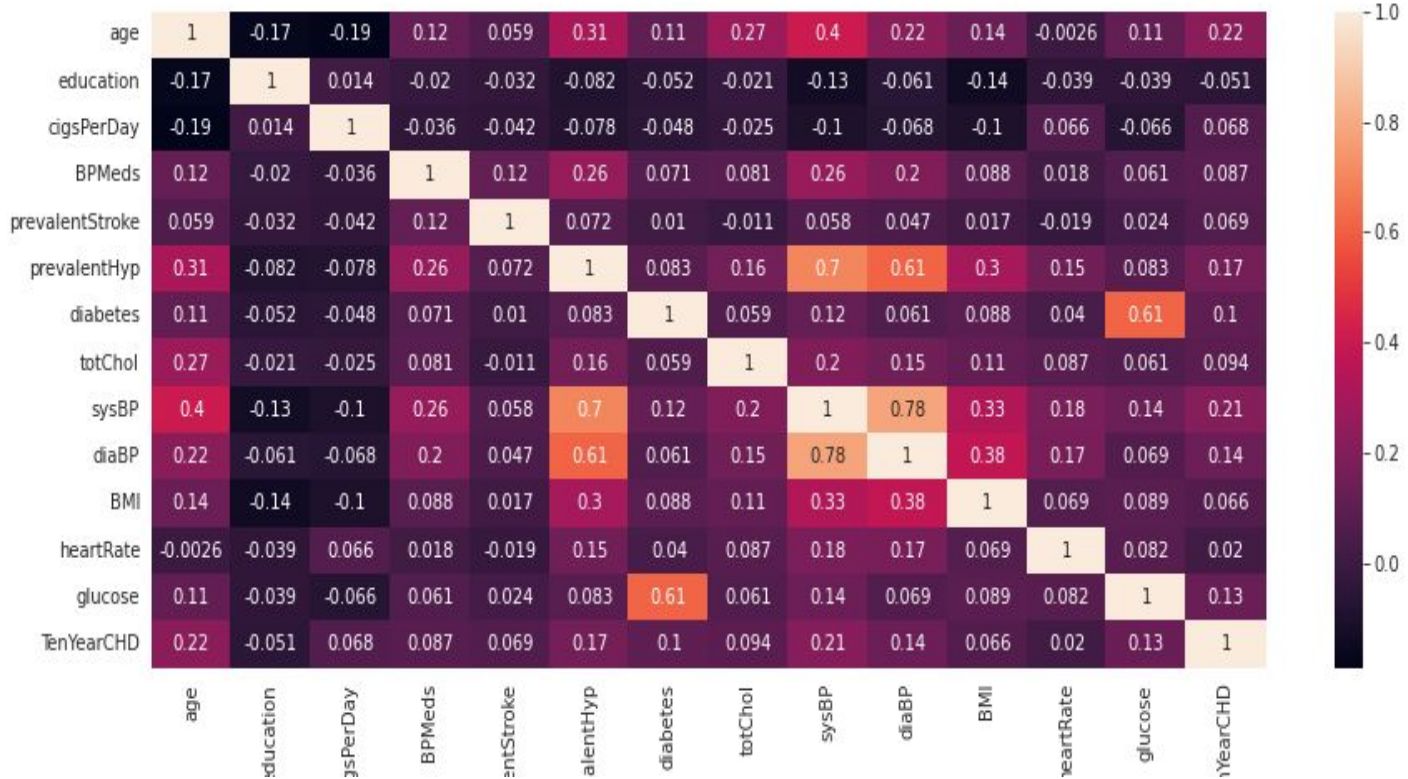




Findings of Univariate Analysis

- Number of females are more than men by a small margin.
- Non smokers are in majority.
- More than 3000 people are not on BP medication.
- Only a small number of people have previously suffered a stroke.
- A large number of people are non-diabetic.
- Most of the numerical columns have a normal distribution.
- Age mostly ranges from 35 to 70 years.
- BMI ranges from 16 to 40.
- Among smokers, the most common number of cigarettes smoked in a day is 20.
- BP numbers range from 100 to 200 and 60 to 120 for systolic and diastolic respectively.

- Bivariate analysis

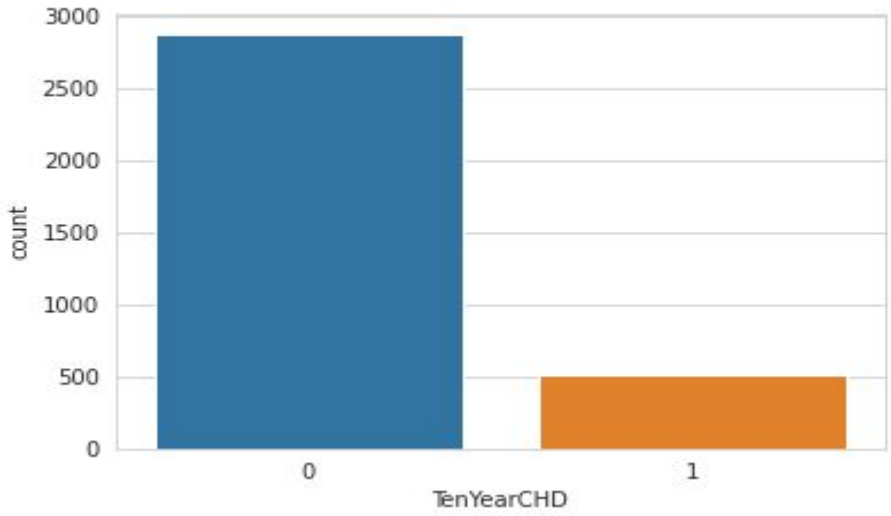


Findings of Bivariate Analysis

- For bivariate analysis focus is mainly on numerical variables.
- Correlation heatmap and pairplot is used.
- Correlation heatmap shows the relationship between the variables.
- Most of the variables are positively related to each other.
- Naturally, systolic and diastolic BP have a positive relationship.
- Cholesterol and glucose have a positive relationship.
- BP is positively related to prevalent hypertension.
- Variables such as age, prevalent hypertension, systolic BP, diastolic BP, and glucose influence the risk of heart disease mainly.

Dependent Variable

- The dependent variable is the Ten Year CHD column.



- As seen there is an imbalance in the classes. There are more people with no risk than people with risk of CHD.

3. Data Transformation

- Some research on systolic and diastolic BP helped in finding an important metric called, pulse pressure.
- Pulse pressure is the difference between systolic and diastolic BP.
- Pulse pressure was added and had the combined value of systolic and diastolic BP.
- ‘Is_smoking’ column was dropped because the ‘cigsPerDay’ column provided the same meaning in a much detailed manner
- Dummy variables were created for the ‘sex’ column to convert the categorical variables to binary values.
- SMOTE was used to remove imbalance from both the classes of the dependent variable.
- Min Max Scaler was used to scale all the values to a similar range between 0 and 1.

3. Data Transformation

- Some research on systolic and diastolic BP helped in finding an important metric called, pulse pressure.
- Pulse pressure is the difference between systolic and diastolic BP.
- Pulse pressure was added and had the combined value of systolic and diastolic BP.
- ‘Is_smoking’ column was dropped because the ‘cigsPerDay’ column provided the same meaning in a much detailed manner
- Dummy variables were created for the ‘sex’ column to convert the categorical variables to binary values.
- SMOTE was used to remove imbalance from both the classes of the dependent variable.
- Min Max Scaler was used to scale all the values to a similar range between 0 and 1.

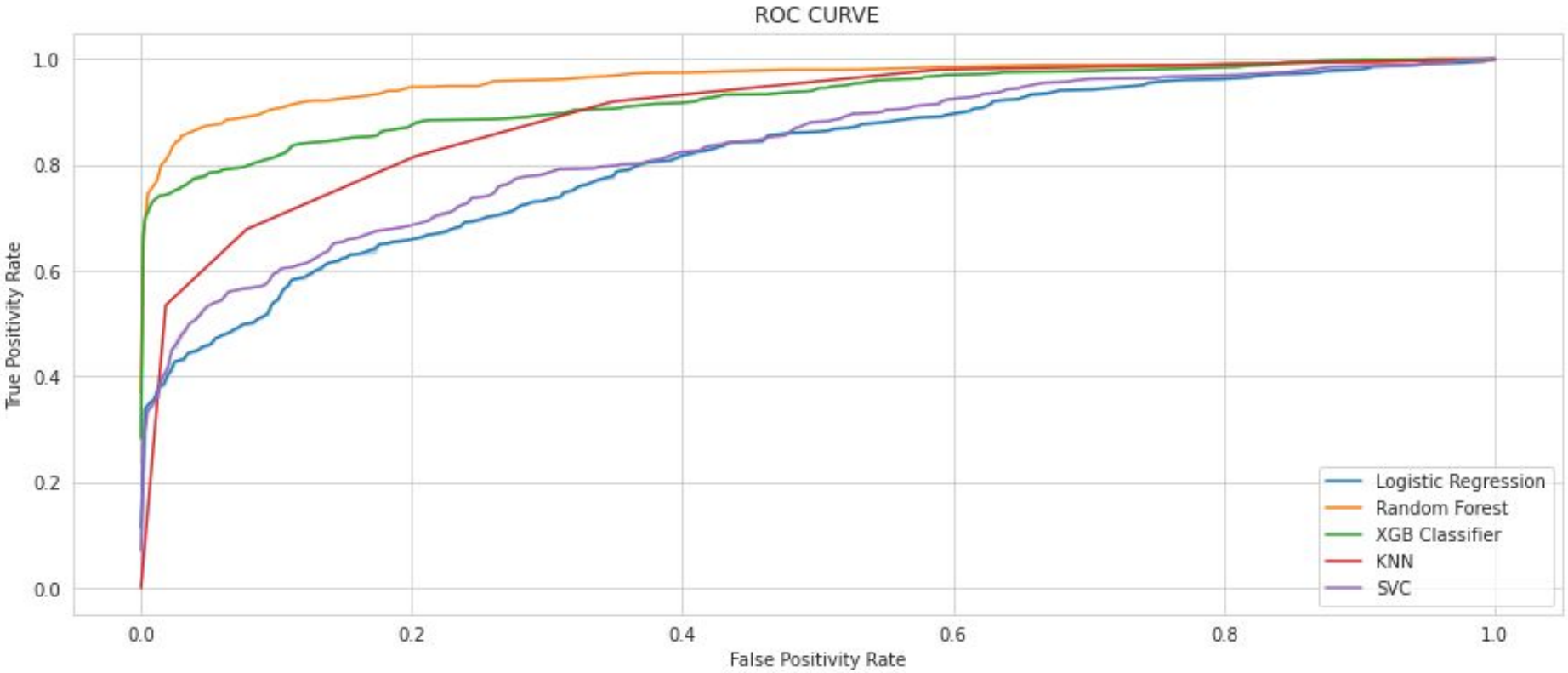
4. Model Building and Evaluation

- For this project, 5 models have been experimented with.

	Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train ROC AUC	Test ROC AUC
0	Logistic Regression	0.746201	0.727431	0.785714	0.736000	0.684843	0.669091	0.746901	0.724911
1	Random Forest	1.000000	0.912326	1.000000	0.927619	1.000000	0.885455	1.000000	0.911166
2	XGB Classifier	0.881676	0.862847	0.942022	0.901639	0.816230	0.800000	0.882423	0.860133
3	KNN	0.881025	0.806424	0.875897	0.786340	0.890940	0.816364	0.880912	0.806853
4	SVC	0.767477	0.757812	0.830705	0.801782	0.678403	0.654545	0.768494	0.753352

- Random forest is overfitting on train data, but at the same time is performing better than other models on the test data. It is closely followed by XGB Classifier.

4. Model Building and Evaluation



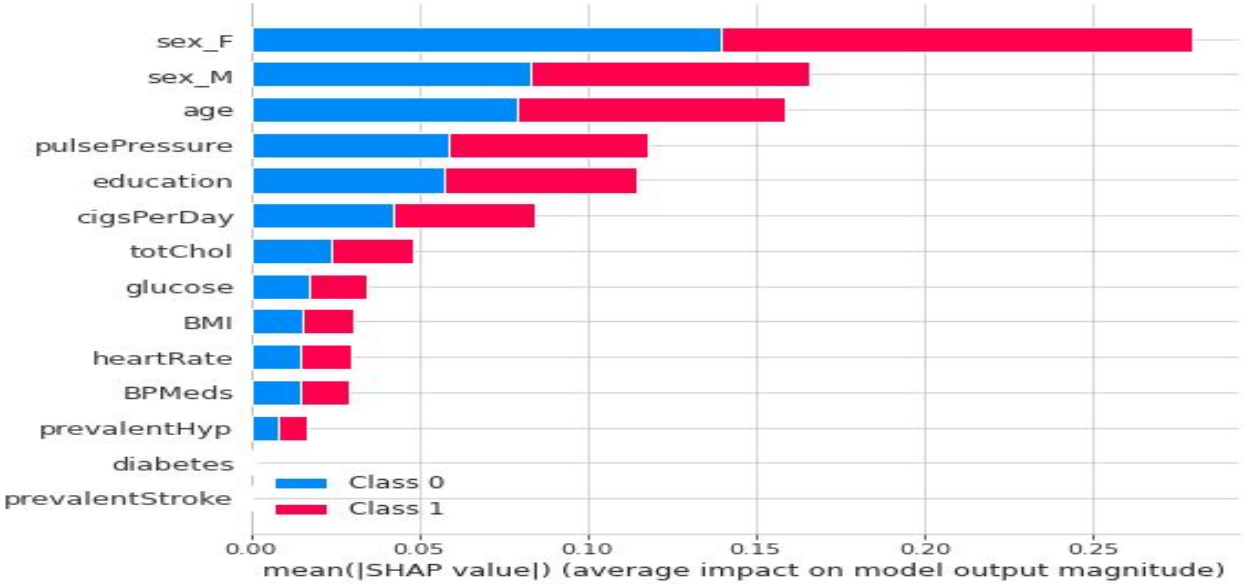
5. Hyperparameter Tuning

- Hyperparameter tuning was performed on the Random Forest model using Grid Search CV. An attempt was done to reduce overfitting and improve results further.

	Model	Train Accuracy	Test Accuracy	Train Precision	Test Precision	Train Recall	Test Recall	Train ROC AUC	Test ROC AUC
0	Random Forest	0.982197	0.888889	0.983226	0.893657	0.981537	0.870909	0.982205	0.888112

- The result of the hyperparameter tuning explain that, overfitting was reduced to a small extent, but results did not improve drastically, the results have remained more or less the same after hyperparameter tuning.

SHAP Feature Importance



- The above figure explains that, features such as gender, age and pulse pressure are highly influencing the possibility of Coronary Heart Disease (CHD).

Conclusion

- A model is developed with almost 89% accuracy, 89% precision and 87% recall on the test data.
- With more knowledge from an expert in the field of cardiovascular health, new variables could be developed to enhance the predictions further.
- Better models other than the ones used here could be used to improve predictions.
- This project has provided experience in an important field of healthcare and has clearly illustrated the application of machine learning in this field.
- Machine learning can help make lives better and with early predictions of diseases, casualties can be avoided.

THANK YOU!