**Cardiovascular Risk Prediction**

**Mrityunjay Dubey**

**Data science trainee,**

**AlmaBetter, Bangalore**

**Abstract:**

This project will lead you through the creation of a screening tool for determining a patient's 10-year risk of developing coronary heart disease (CHD) using various machine learning techniques.

**Problem statement:**

Globally, heart disease is the leading cause of morbidity and mortality, killing more people each year than any other disease combined. The WHO estimates that 17.9 million deaths worldwide from heart disease occurred in 2016, accounting for 31% of all fatalities. More than 75% of these fatalities occurred in developing and middle-income nations.

Coronary heart disease, also known as a heart attack, is by far the most prevalent and lethal of all heart conditions. For instance, it is believed that someone in the United States experiences a heart attack every 40 seconds, and that there are roughly 805,000 heart attacks there each year (CDC 2019).

Our given dataset has 3390 records and 16 attributes.

**Dataset Summary:**

• **Sex**: Gender of person whether male or female("M" or "F")

• **Age**: Age of the patient;(Continuous - Although the recorded ages are whole numbers as they are only years, the concept of age is continuous)

• **Education:** Ordinal, high values represents highly educated.

• **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.)

• **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)

• **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**: Continuous, in medical research, variables such as heart rate though in fact discrete, are considered continuous because of a large number of possible values.

• **Glucose**: The glucose level in blood.

• **TenYearCHD**: Abbreviation for Ten Year coronary heart disease, nominal and our target variable as well.

**EDA:**

**Exploration of data**

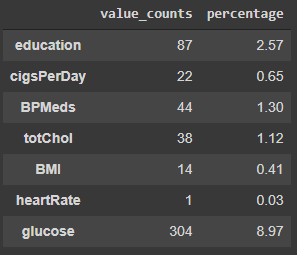
After importing our data, we looked at it in greater detail and tried to better comprehend it so that we could consider what information our dataset might include. 3390 rows and 16 columns make up the dataset in question.

**Data scrubbing**

We used this procedure to convert our raw data into the correct format and delete unwanted columns in order to provide results that were more accurate and handled more complex data in a shorter amount of time.

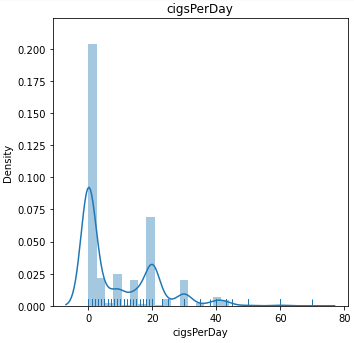
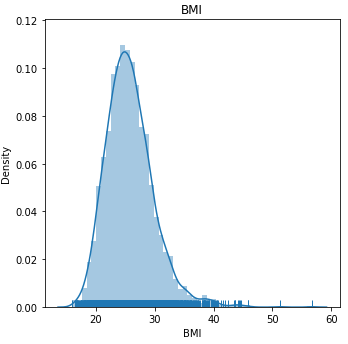
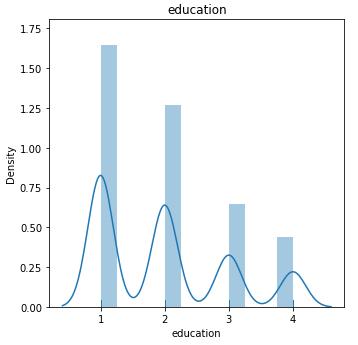
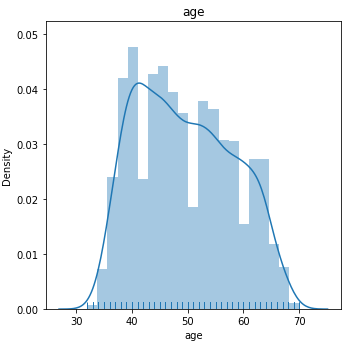
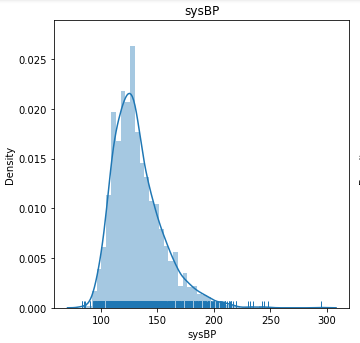
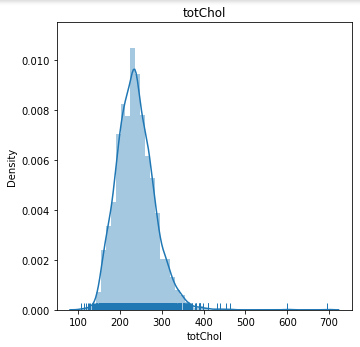
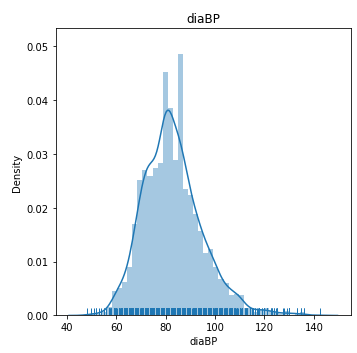
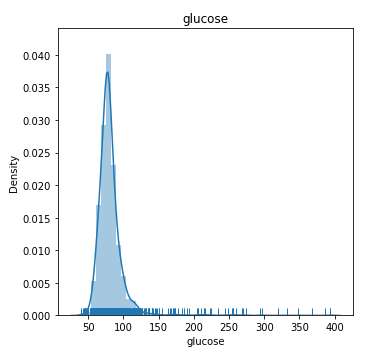
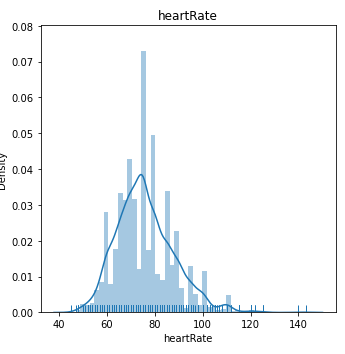
**Handling Null values**

Our dataset has few null values that’s need to be treated



As almost all the data are right skewed that’s why we imputed null values.

**Distributions of continuous variables**

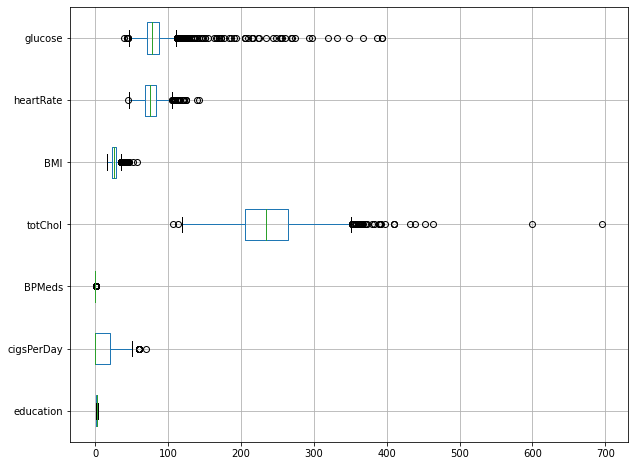
** **

We can see that several of the variables are right-skewed from this point. This might be as a result of outliers, as we'll see in the section on outlier handling.

**Treating Null Values:**

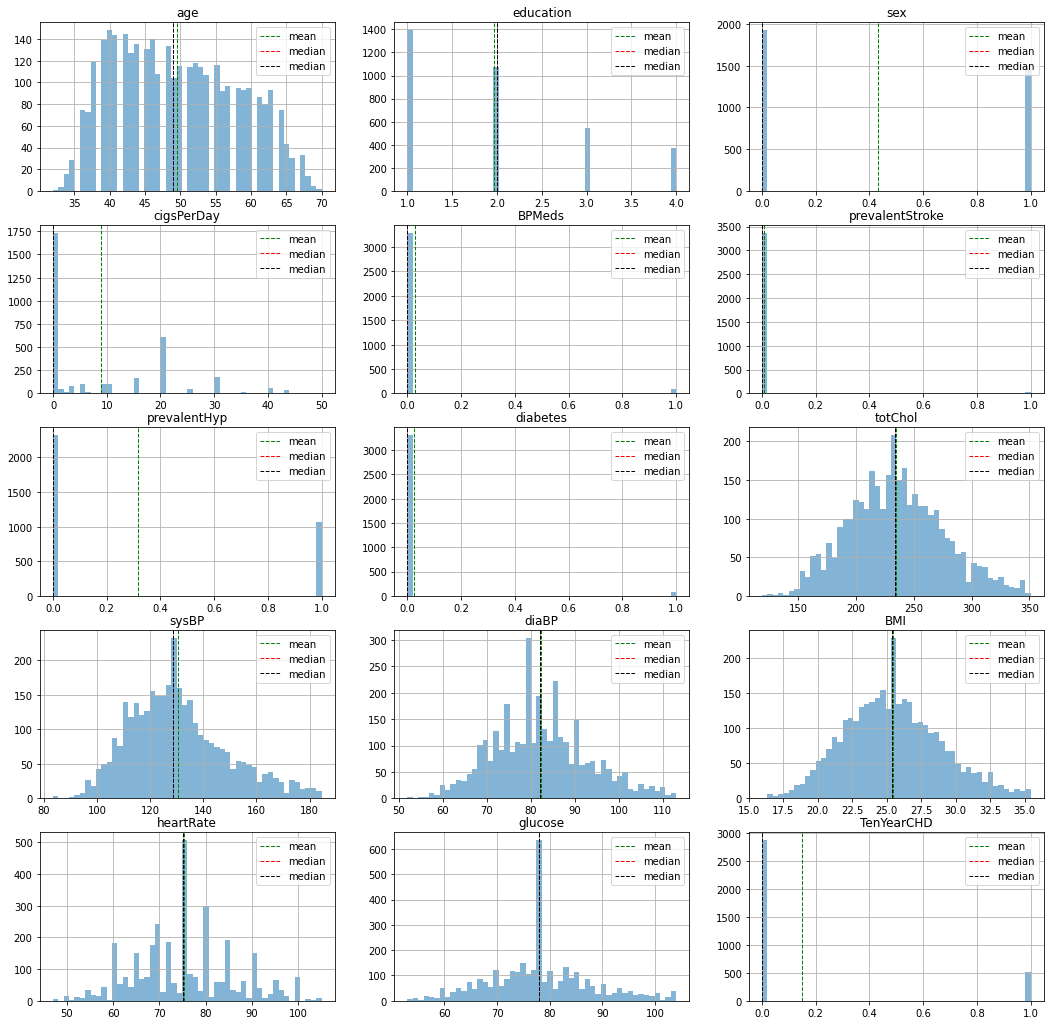
Education, CigsPerDay, BPMeds, totChol, BMI, heartrate, and Glucose all have null values which needs to be treated

We can use either of mean , mode or median to impute the null values we can decide which one to use after having a look at boxplot.



We shouldn't enter the mean since outliers have a significant impact on the mean and there are outliers in the variables.

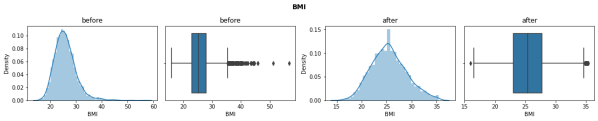
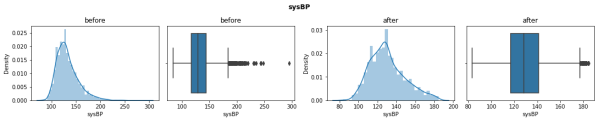
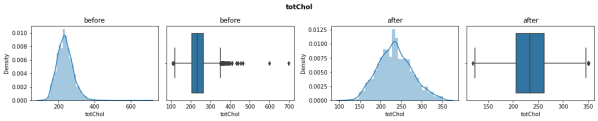
We can see that the mode and median are equal for all the variables. Therefore, since the mode and median are unaffected by outliers, we can utilise either one.



There are no longer any null values in the dataset as a result

**Treating Outliers**

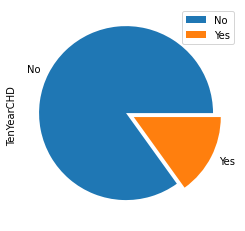
All values outside of the lower and upper limits are replaced with the variable's median

This is how our variables looks like before and after treatment of outliers.

**HandlingClass Imbalance:**

The goal variable, "TenYearCHD," is clearly very unbalanced, as can be shown. All others variables are more or less are balanced. We can fix this by SMOTE, Tomelinks, Penalizing misclassification.



We try to increase the number of data points using SMOTE because our dataset is tiny. We then use Tomek to reduce the number of data points. Finally, we penalise the loss function for different classifier models (RandomForest, SVC).

**Fixing Non-Numeric Variables:**

Only two variables in our model, sex and is smoking, have non-numeric values.

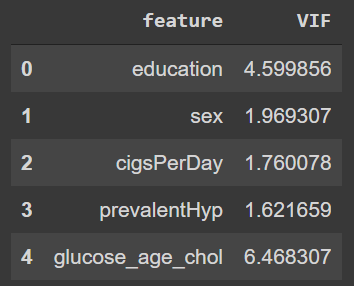
We remove is smoking since its data is derived from the daily cigarette consumption column.

Next, we swap Male and Female with 0 and 1, respectively.

**Preparing Data for logistic regression.**

We create a column named glucose age chol that is a function of glucose, age, and chol to ensure that our data is not multicollinear. where cholestral and age have the most roles.

The following are the final VIF values for the variables:



**Train Test split and oversampling for logistic regression**

Multicollinearity needed to be eliminated for logistic regression, which was done. The data must then be scaled using the MinMaxScaler.

Specifically, we substitute each value for each data point at each variable with

(datapoint – minimum)/ (maximum- minimum)

Following that, we divide the data into a train set and a test set, with the remaining data going to the test set. The test set has 102 values, all of which are 1. 2303 of the values in the train set are 0s, while 409 are ones.

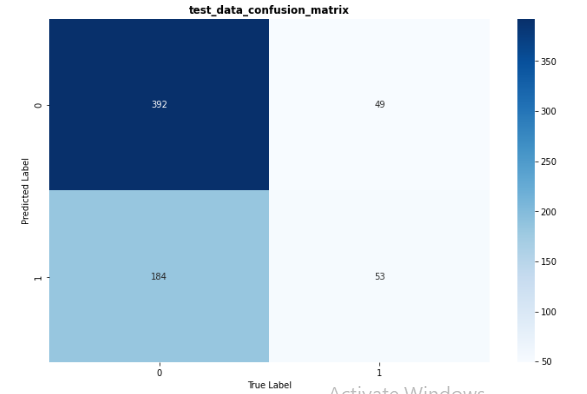
In the training dataset, we must now balance the 1s. Our training set has 1905 values of zeros and ones after applying SMOTE and Tomeklinks.

For the Naive Bayes model, the same dataset is also utilised.

**Applying Classification Model**

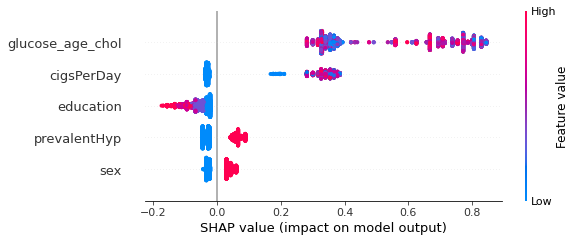
Prior to using any model, we must choose our goal. We will give priority to predicting the 1s since our main goal should be to determine whether a person will get heart disease over the next 10 years or not. The scoring tool of choice is the ROC-AUC curve.

1. Logistic Regression: We obtain a C value of 0.77 and a penalty of l2 using GridSearchCV. The resulting confusion matrix is as follows.



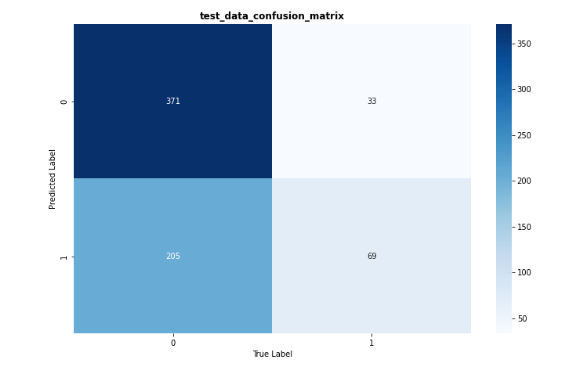
Our goal is to enhance this number, which currently allows us to forecast 60% of those who may acquire heart disease.

Additionally, the ROC-AUC curve and Shap values are discovered.



2. Gaussian Naive Bayes: Using the Naive Bayes model. Since we can't generate better outcomes, we mainly ignore this model.

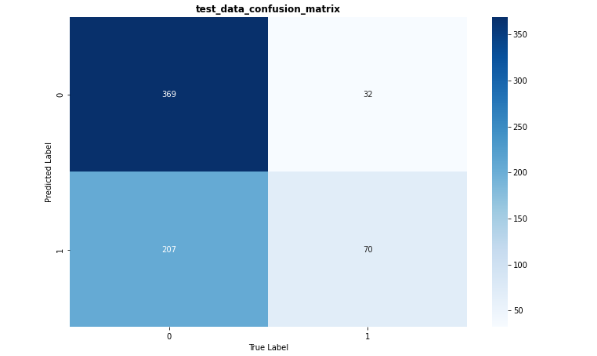
3. Suppor Vector Classifie1 : We prepare the dataset before performing SVC. In order to do this, we use the dataset that was multicollinearity-free. Use GridSearchCV to apply a support vector classifier to obtain this confusion matrix.



In this case, Logistic Regression did not produce the best results. As before, we also get the categorization report and ROC-AUC curve.

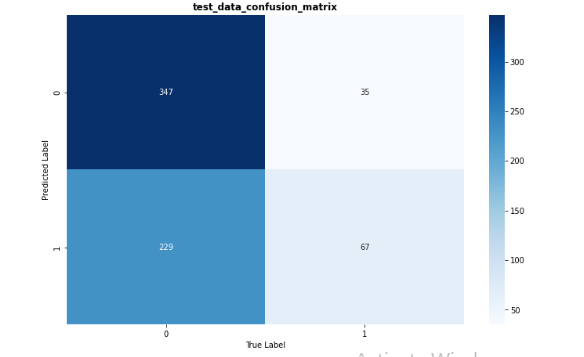
4. Support Vector Classifier 2: To tackle the class imbalance, we employ two methods: first, we create new data points using smote and tomek linkages, and second, we employ a loss function that severely penalises incorrect classifications.

The confusion matrix is then obtained by applying SVC with GridSearchCV to obtain C = 40, gamma = 0.01 and kernel = rbf.

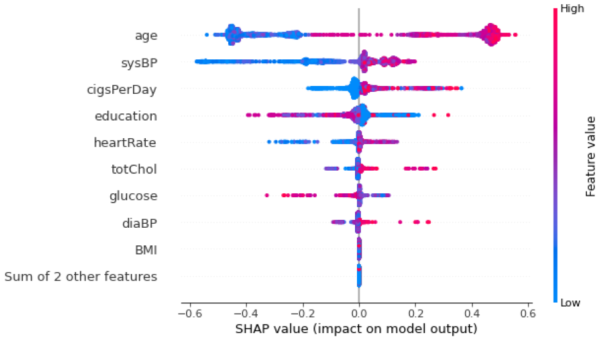


5. Ensemble Tree Models: For ensemble tree models, we get the data ready. We generate the dataset and train test split in accordance with these since scaling is not necessary.

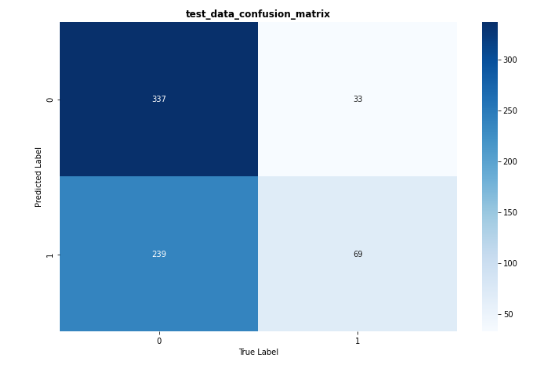
5.1. LGBM Implementation: To find the ideal hyperparameters, we use GridSearchCV (maximum depth, number of estimators and learning rate). The resulting confusion matrix is

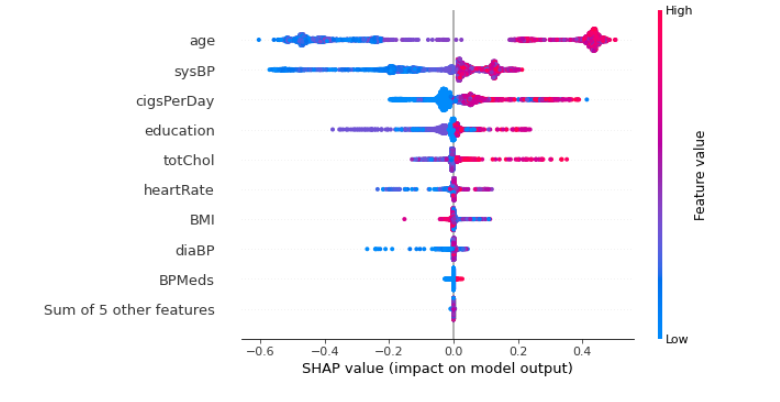


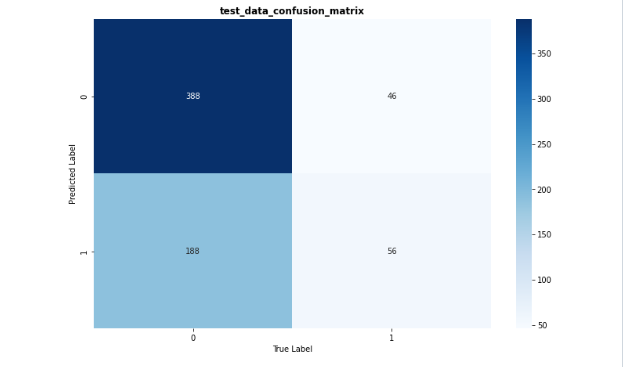
Over 60% of the data might be appointed in this case. Additionally obtained are the ROC-AUC curve and SHAP plots.



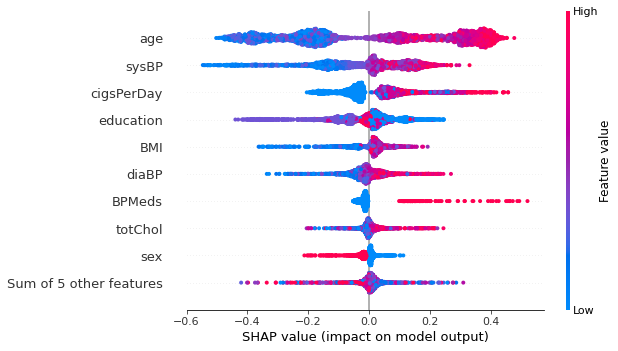
5.2 Extreme Gradient Random Forest Boosting Model (XGRFB): Using GridSearchCV, we implement the XGRFB model and adjust the hyperparameters. The resulting confusion matrix is

 Additionally, we get the ROC-AUC curves and SHAP values.

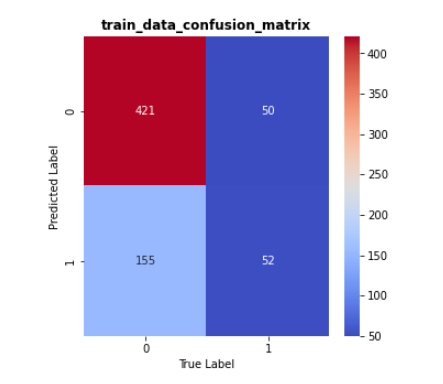


5.3. Random Forest Implementation: The Random Forest Model is applied as previously, and GridSearchCV is used to fine-tune the hyperparameters. The resulting confusion matrix is 

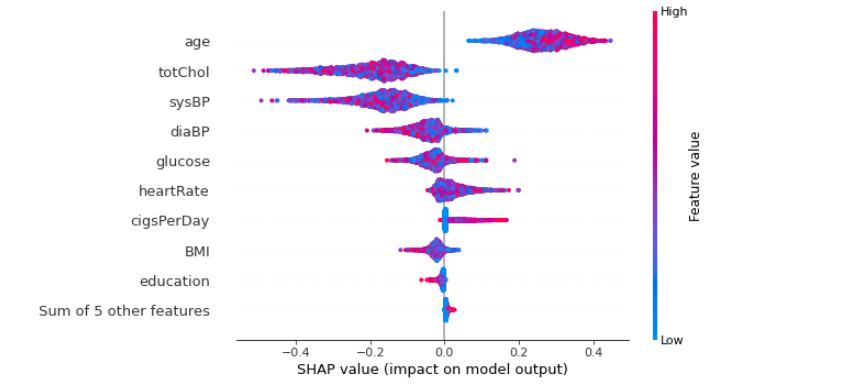
Additionally, we get the ROC-AUC curves and SHAP values.



6. Neural Network: We use a neural network model with one output vertex with early stopping and two hidden layers. The resulting confusion matrix is



The SHAP Value of NN is



**10.Conclusion.**

If we want simplicity, we can utilise logistic regression among the others. We can utilise the SVC(2) or XGRFB models, among others, if we want to get high results. We favour the XGRFB model because it effectively captures the core of the data.