Capstone Project – 3

Cardiovascular Risk Prediction Supervised ML-Classification

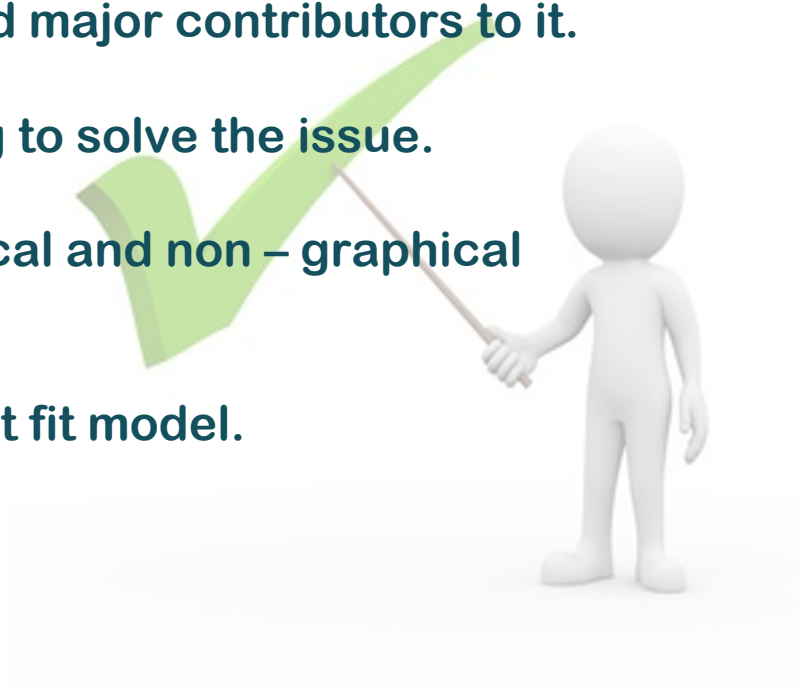
Team Members

Sanjeev Kumar Thakur

Jogapritam Sahu

Objectives

- Understanding the problem at hand and major contributors to it.
- Elucidating what and how are we trying to solve the issue.
- Exploratory data analysis using graphical and non – graphical approach.
- Applying models and validating the best fit model.
- Conclusion.



Data Pipeline

- **Data Processing** - In this part we removed unnecessary features and peculiar observations. Since there were some null observations we got rid of those too.

- **EDA** - Then we did some exploratory data analysis on the features from the previous actions to visualize the pattern.

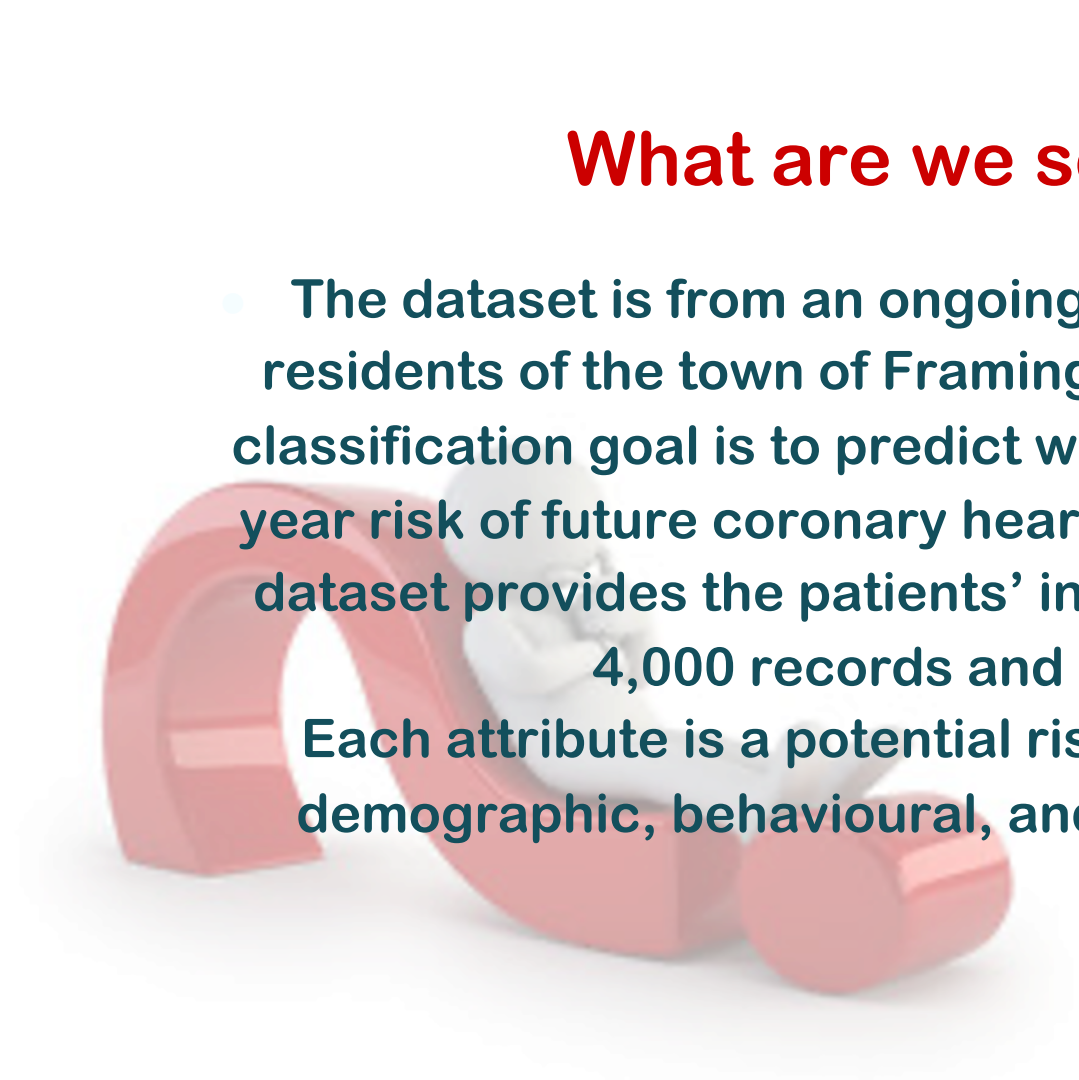
- **Data Balancing** – Since our data was class imbalanced we tried to balance it using SMOTE(Synthetic minority oversampling technique).

- **Model evaluation** - Finally we prepared our features to train various models and evaluate their performance in an iterative manner starting from the simplest model and gradually increasing the complexity.

What are we solving ?

- 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) or not. The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

Each attribute is a potential risk factor. These include demographic, behavioural, and medical risk factors.



Coronary Heart Disease

If we were to define Coronary Heart Disease(CHD) in a nutshell then : Coronary Heart Disease comes in the category of Cardiovascular diseases that are a group of disorders of the heart and blood vessels.

Specifically, Coronary heart disease is the term that describes what happens when your heart's blood supply is blocked or interrupted by a build-up of fatty substances in the coronary arteries.

Coronary heart disease cannot be cured but treatment can help manage the symptoms and reduce the chances of problems such as heart attacks.

Data Features

Demographic

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

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.

Medical(history)

- **BP Meds**: whether or not the patient was on blood pressure medication
- **Prevalent Stroke**: whether or not the patient had previously had a stroke
- **Prevalent Hyp**: whether or not the patient was hypertensive
- **Diabetes**: whether or not the patient had diabetes

contd.

Data Features

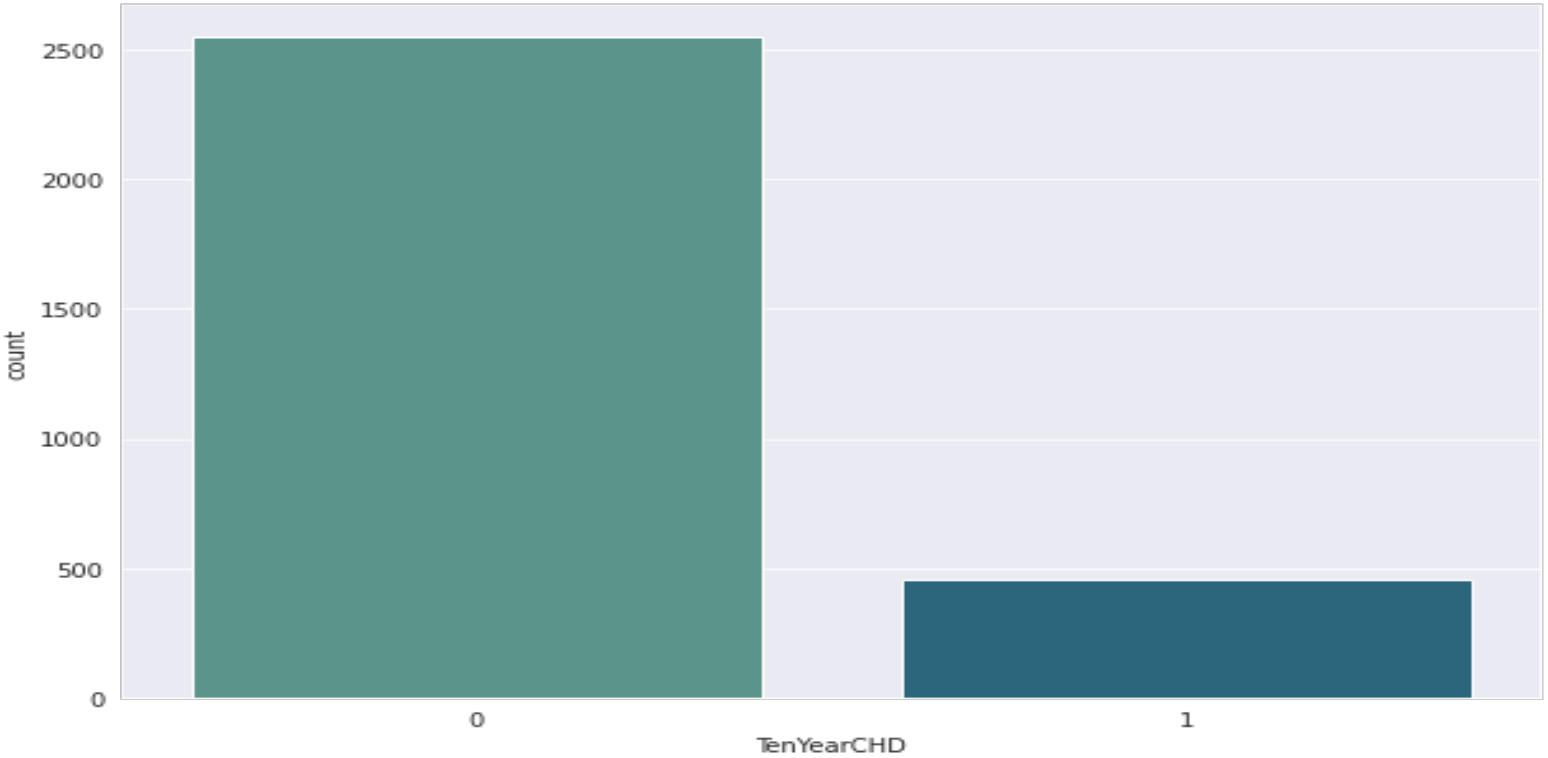
Medical(current)

- **Tot Chol**: total cholesterol level
- **Sys BP**: systolic blood pressure
- **Dia BP**: diastolic blood pressure
- **BMI**: Body Mass Index
- **Heart Rate**: heart rate (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

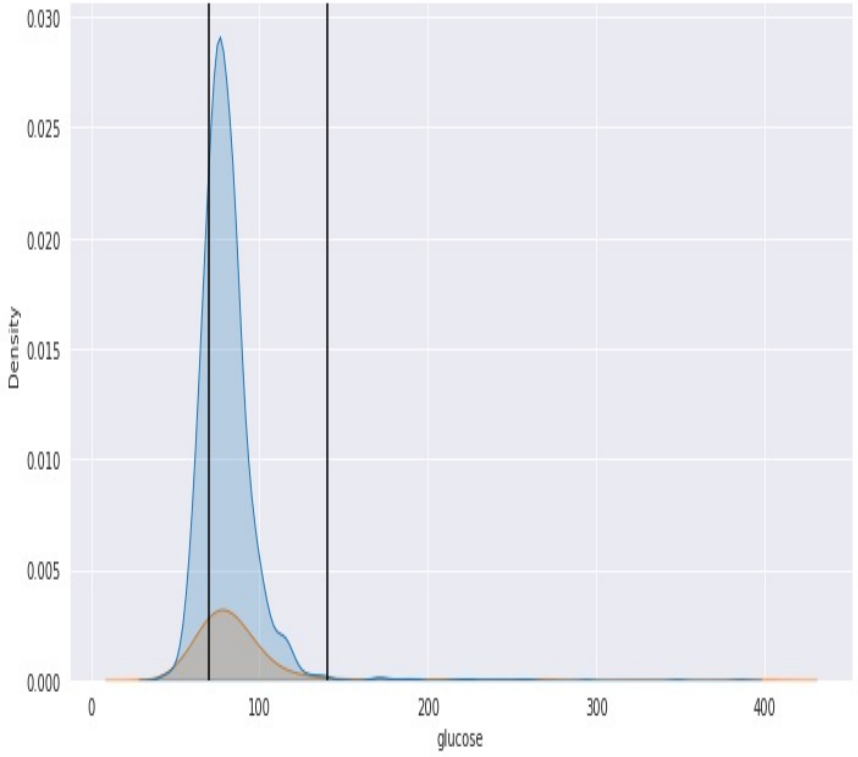
Predict variable (desired target)

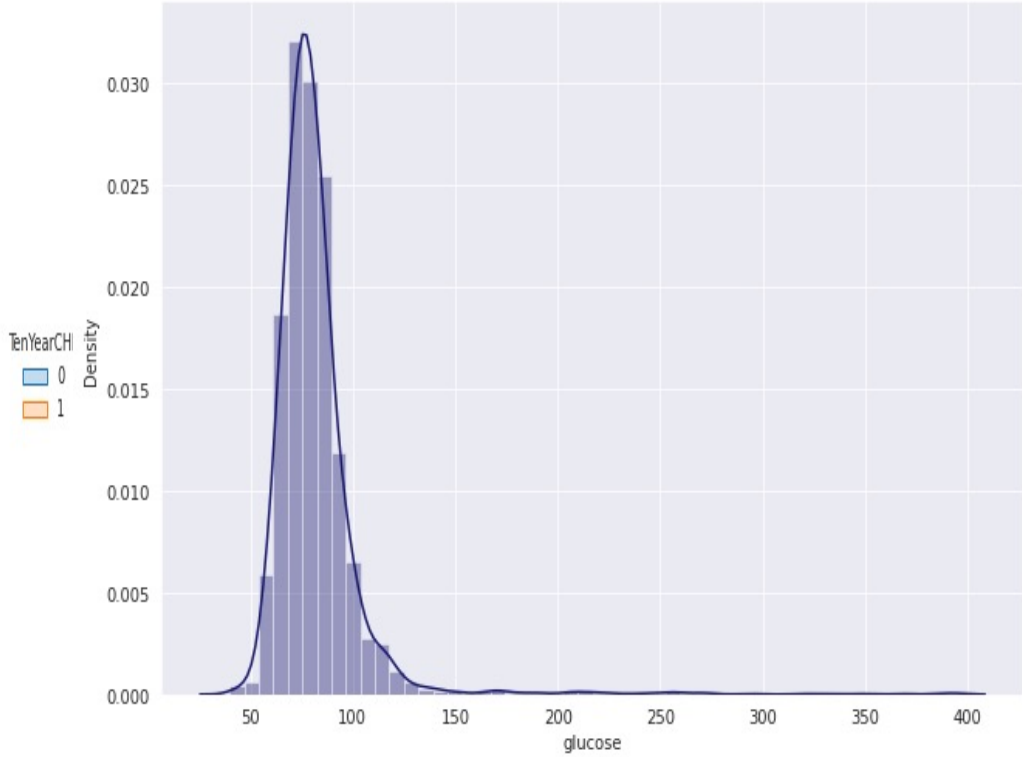
- **TenYearCHD**: 10-year risk of coronary heart disease CHD(binary: “1”, means “Yes”, “0” means “No”)

Dependent Variable(TenYearCHD)

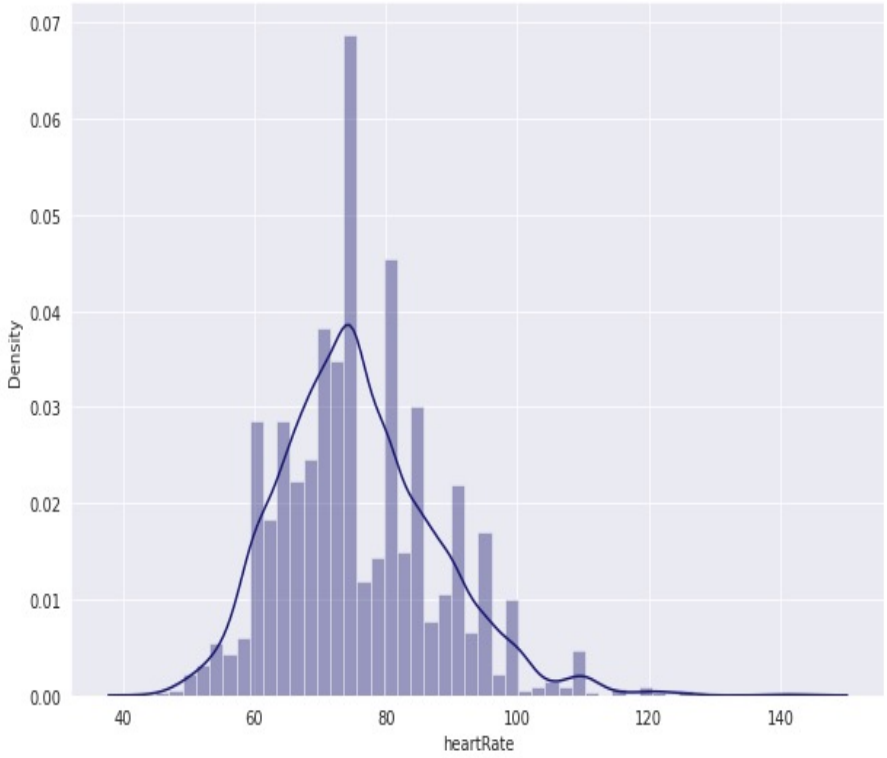


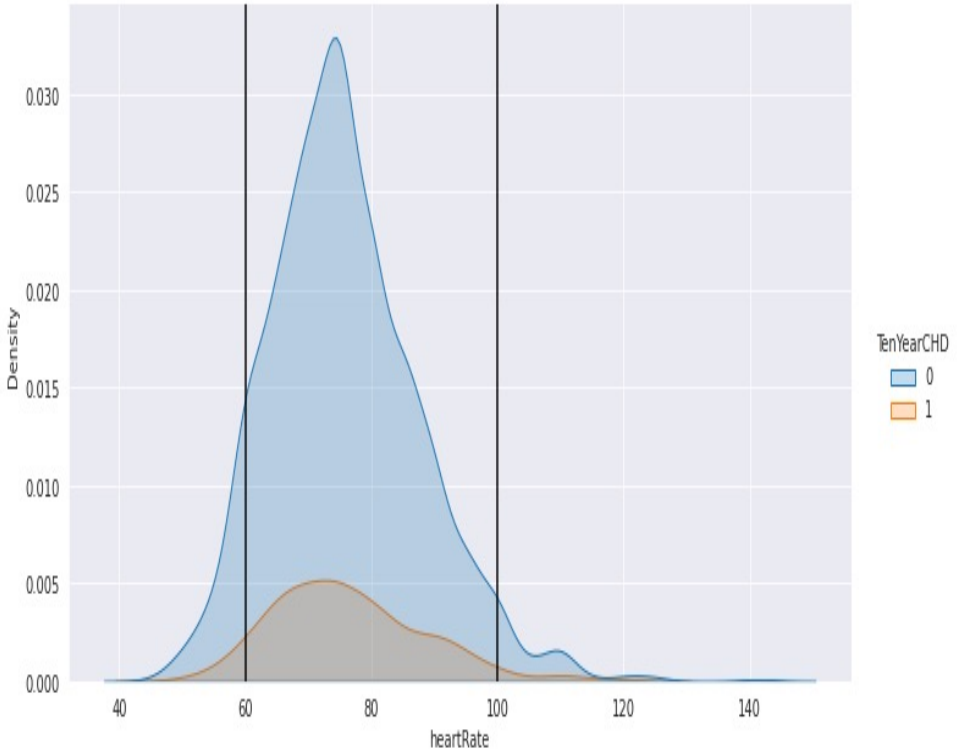
Glucose (distribution)





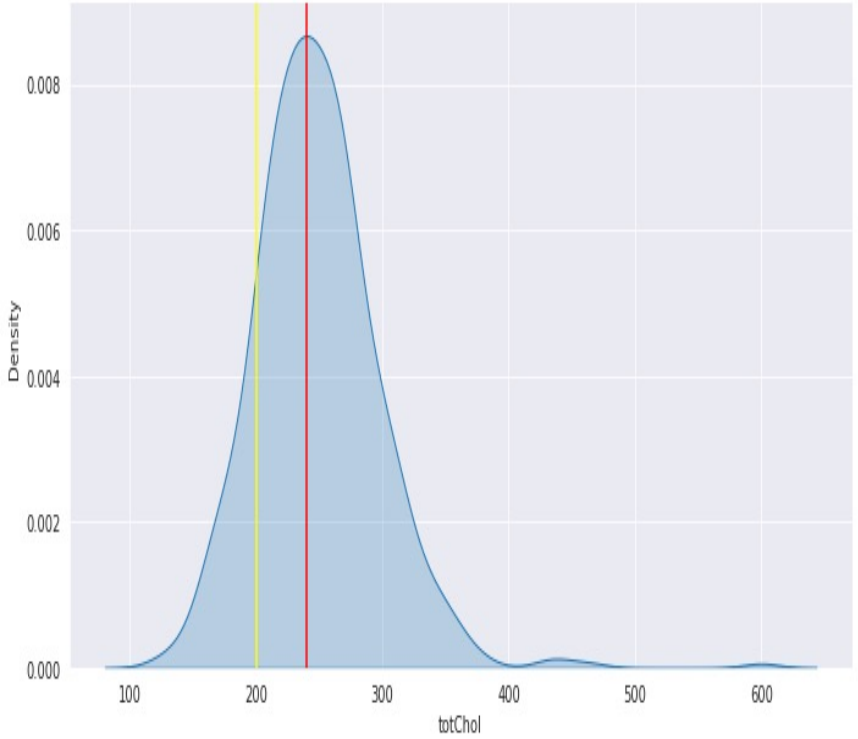
Heart rate (distribution)

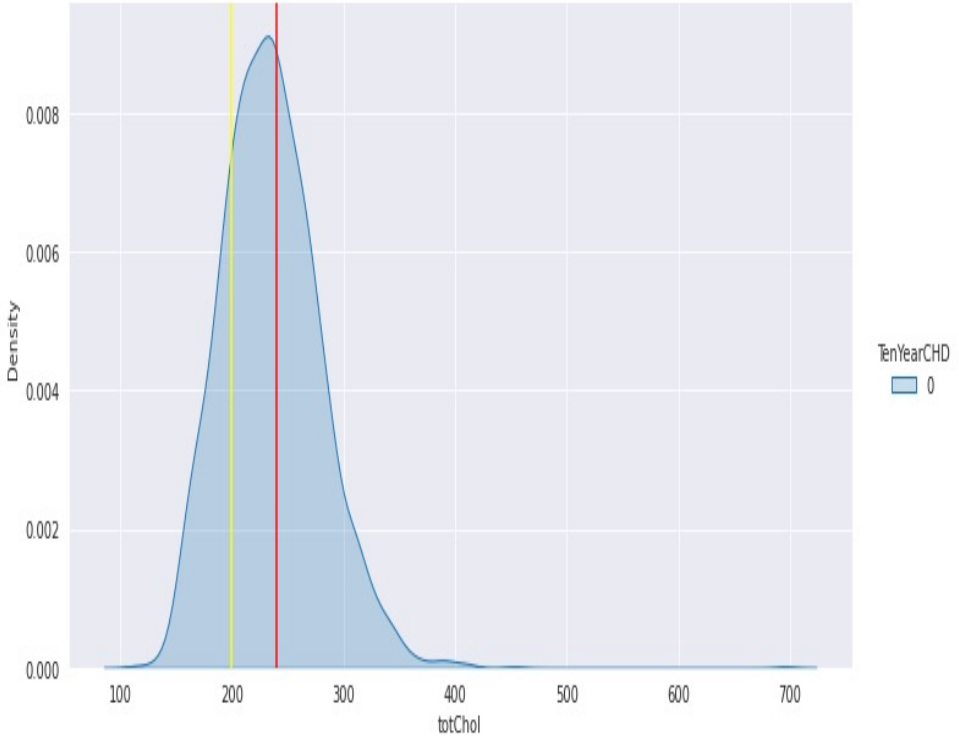




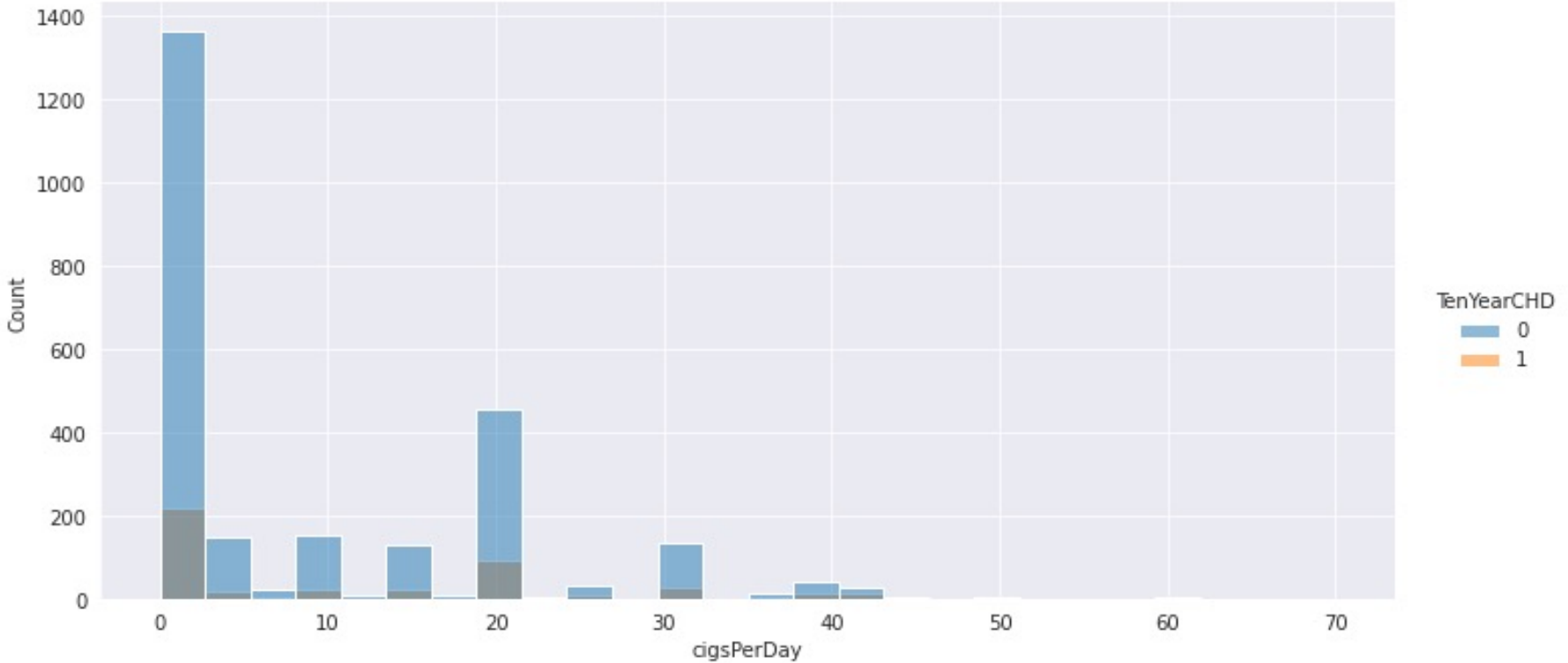
Body Mass Index vs TenYearCHD

Total Cholesterol vs TenYearCHD

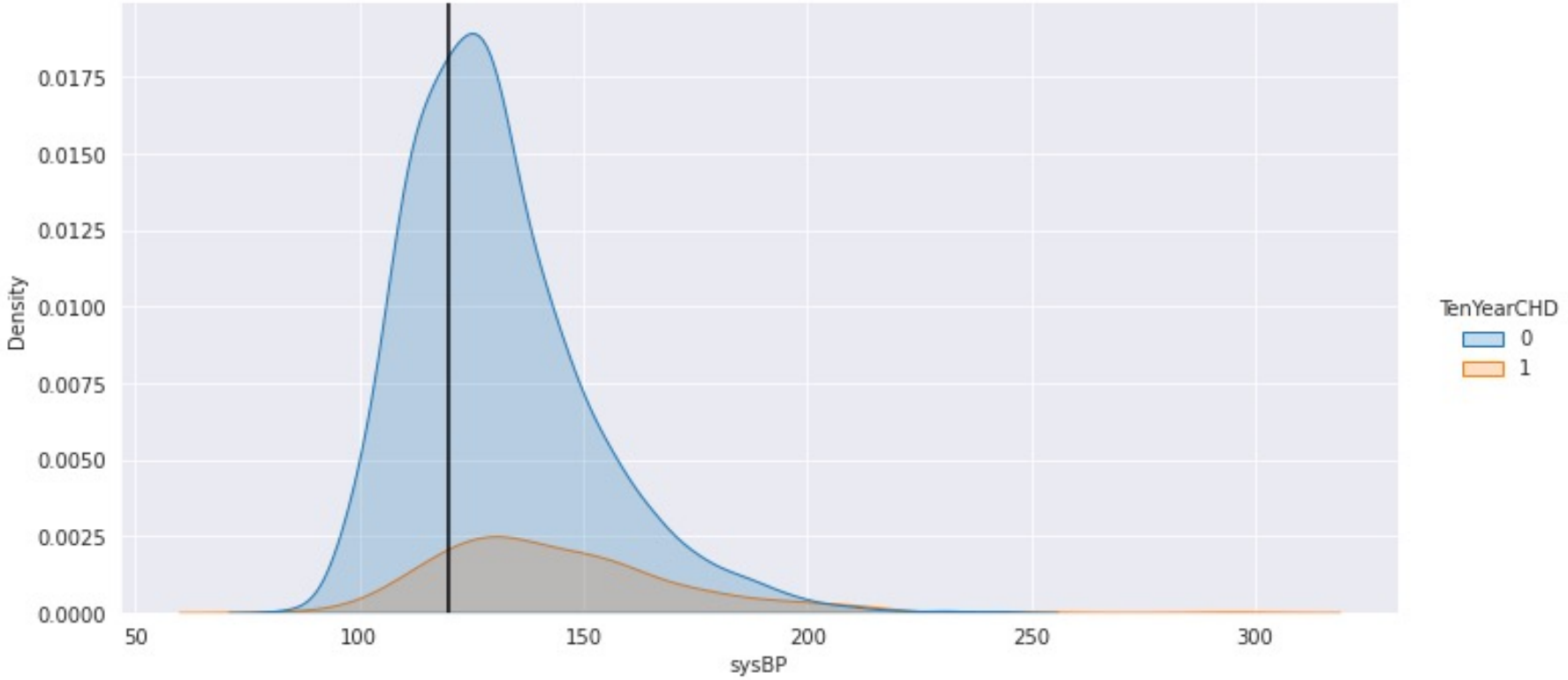




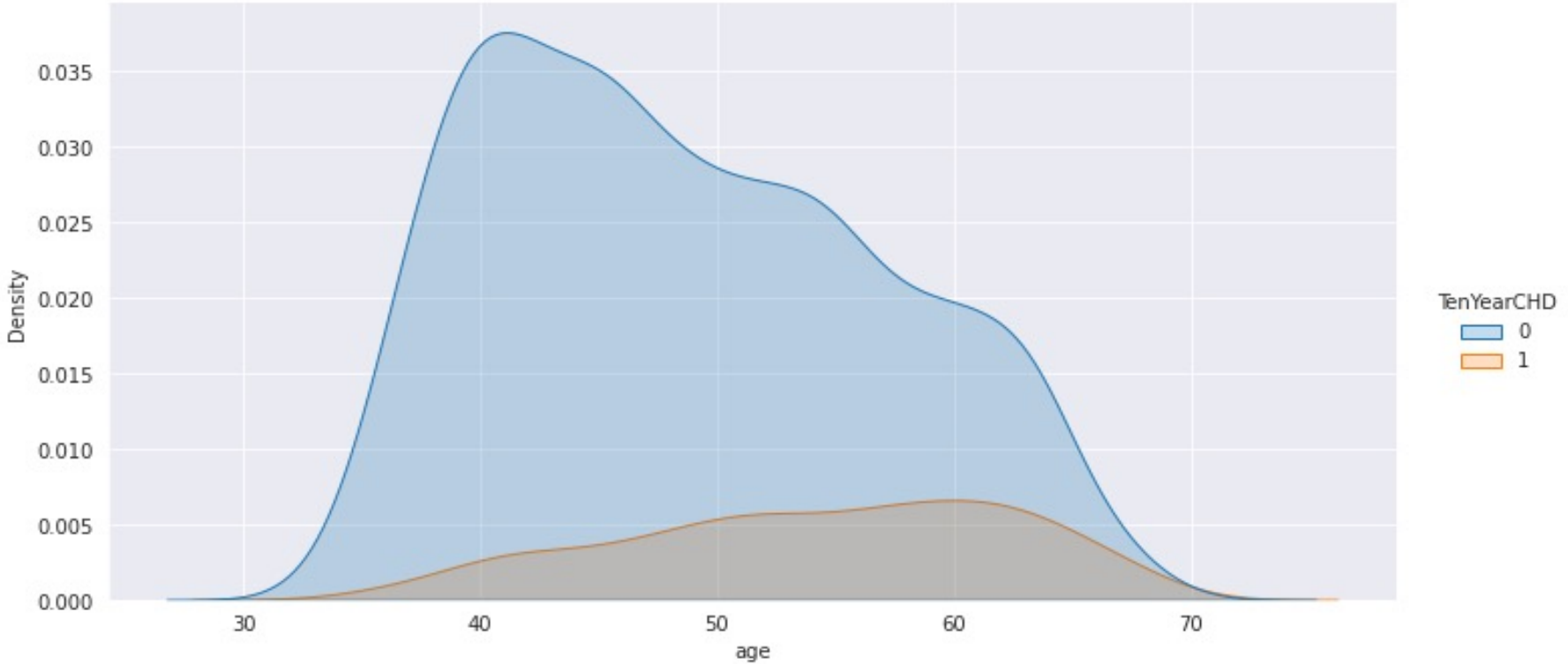
Cigarettes per day vs TenYearCHD



Systolic blood pressure vs TenYearCHD



Age vs TenYearCHD



Supervised Machine Learning
Classification analysis

Supervised machine learning algorithms is defined by its use of labelled datasets to train algorithms to classify data or predict outcomes accurately.

Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

Classification is used when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.

Classification models used :

- Logistic Regression.
- K Nearest Neighbour.
- Decision Tree Classifier.
- Random Forest Classifier.
- XGBoost Classifier.
- Naive Bayes.
- Support Vector Machine.

We will go through the results of the best two i.e. Random forest and Support vector machine algorithms.

SMOTE
(synthetic minority oversampling technique)

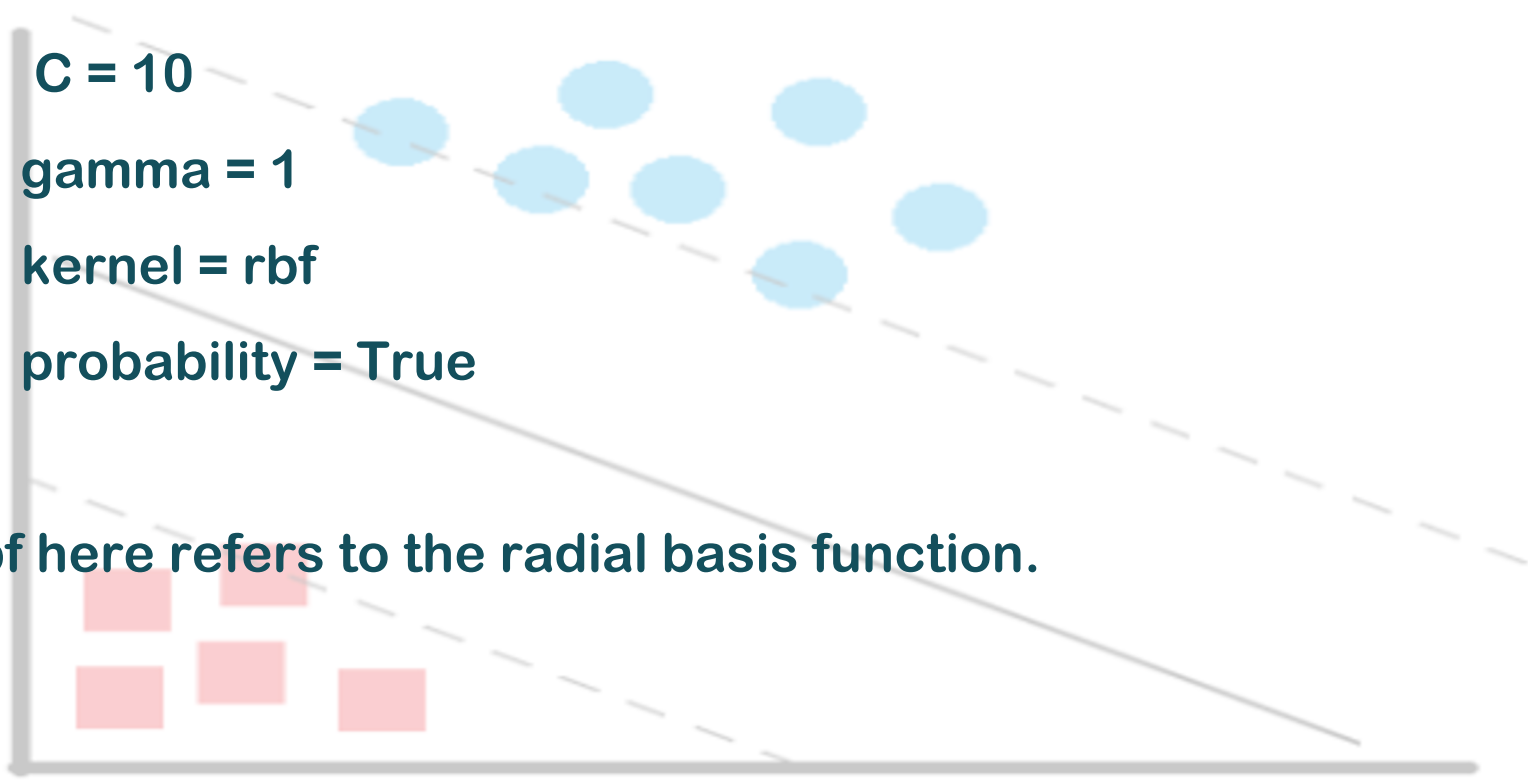
- This was a class imbalanced dataset so we used SMOTE(Synthetic minority oversampling technique) which is a class imbalance handling technique before running our algorithms.

- **SO, WHAT IS SMOTE ?**
- This is a statistical technique for increasing the number of cases in your dataset in a balanced way. The module works by generating new instances from existing minority cases that you supply as input. This implementation of SMOTE does not change the number of majority cases.
- SMOTE takes the entire dataset as an input, but it increases the percentage of only the minority cases.

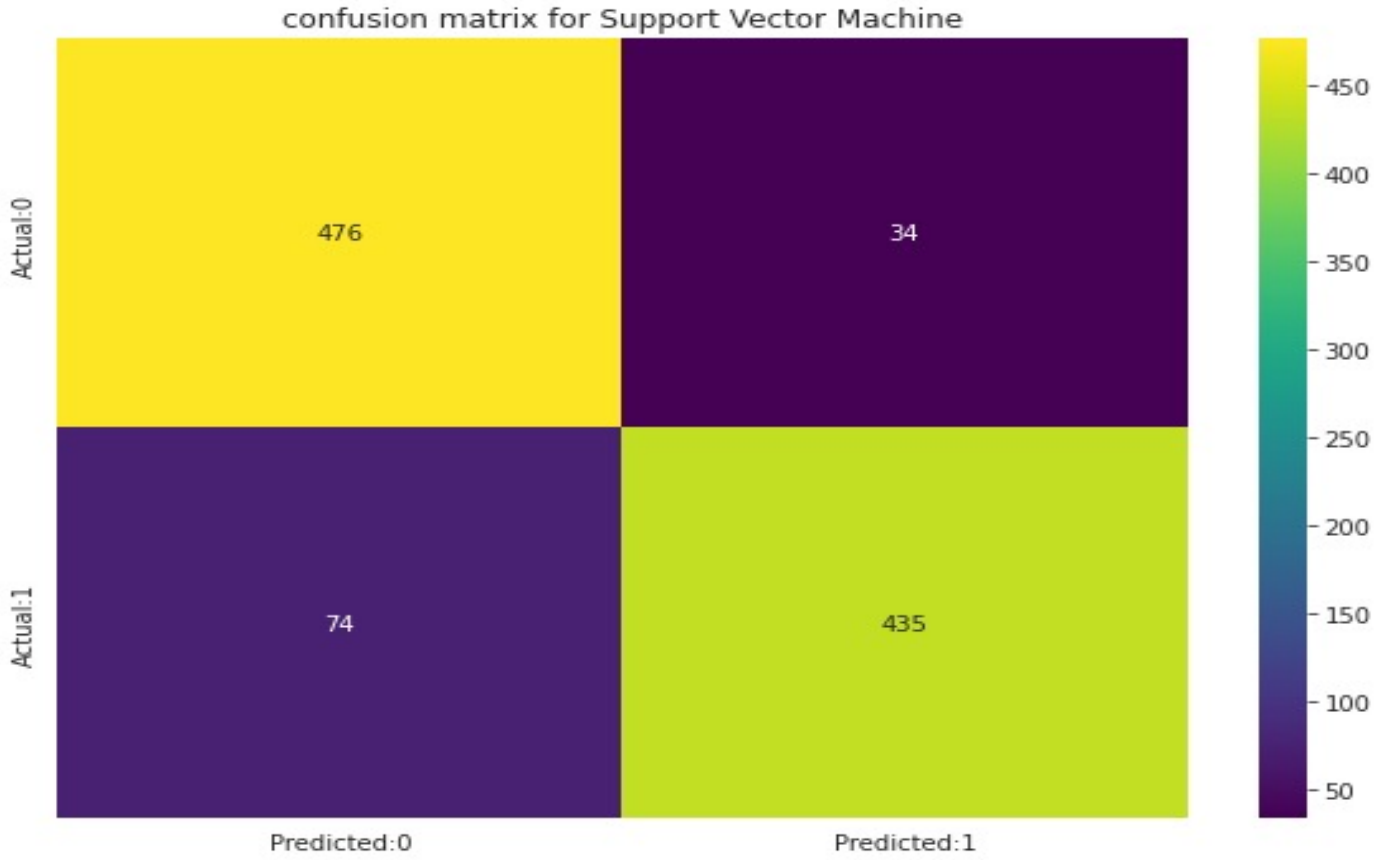
Support Vector Machine Hyperparameters

- C = 10
- gamma = 1
- kernel = rbf
- probability = True

rbf here refers to the radial basis function.



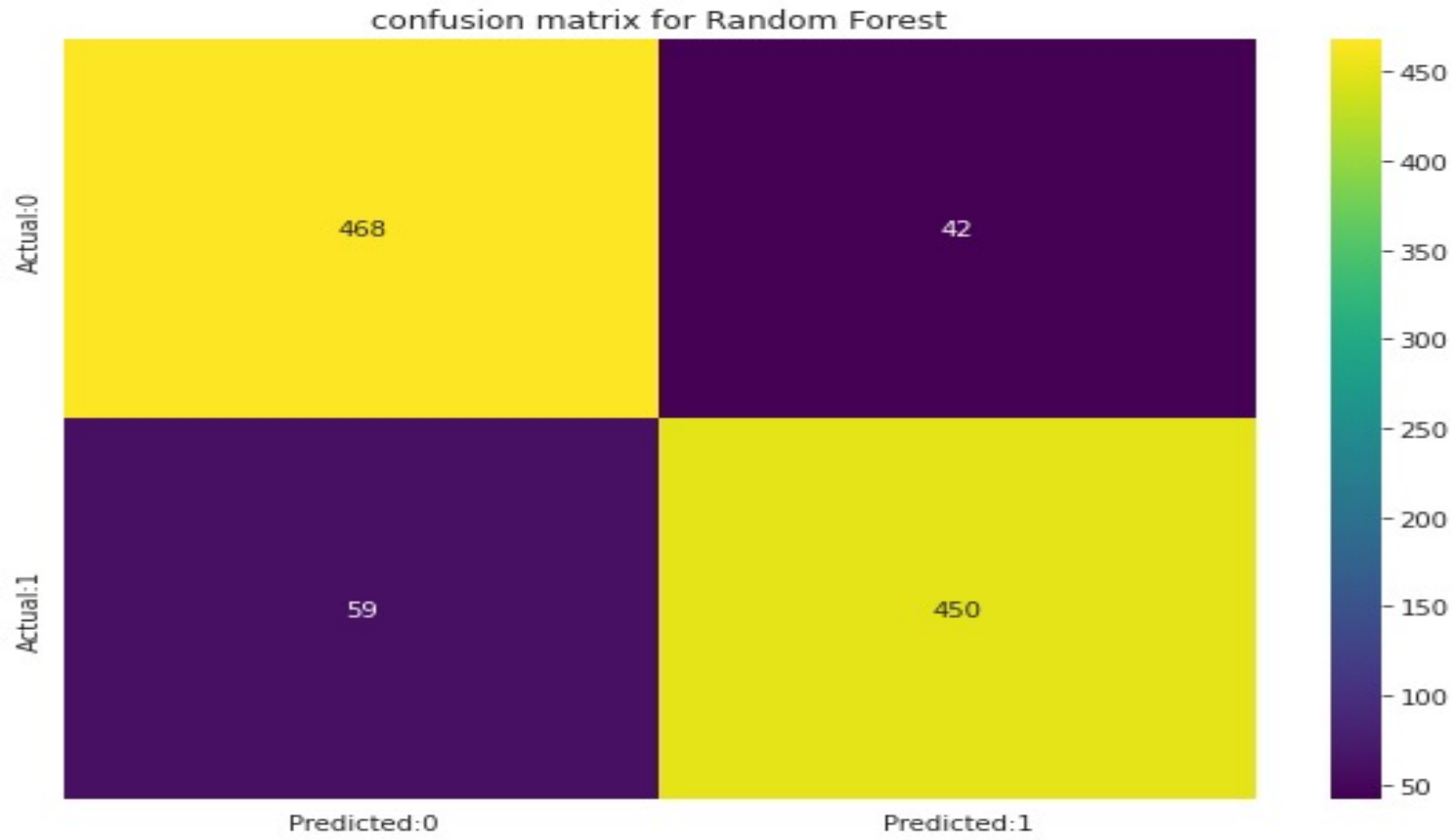
Support Vector Machine – confusion matrix



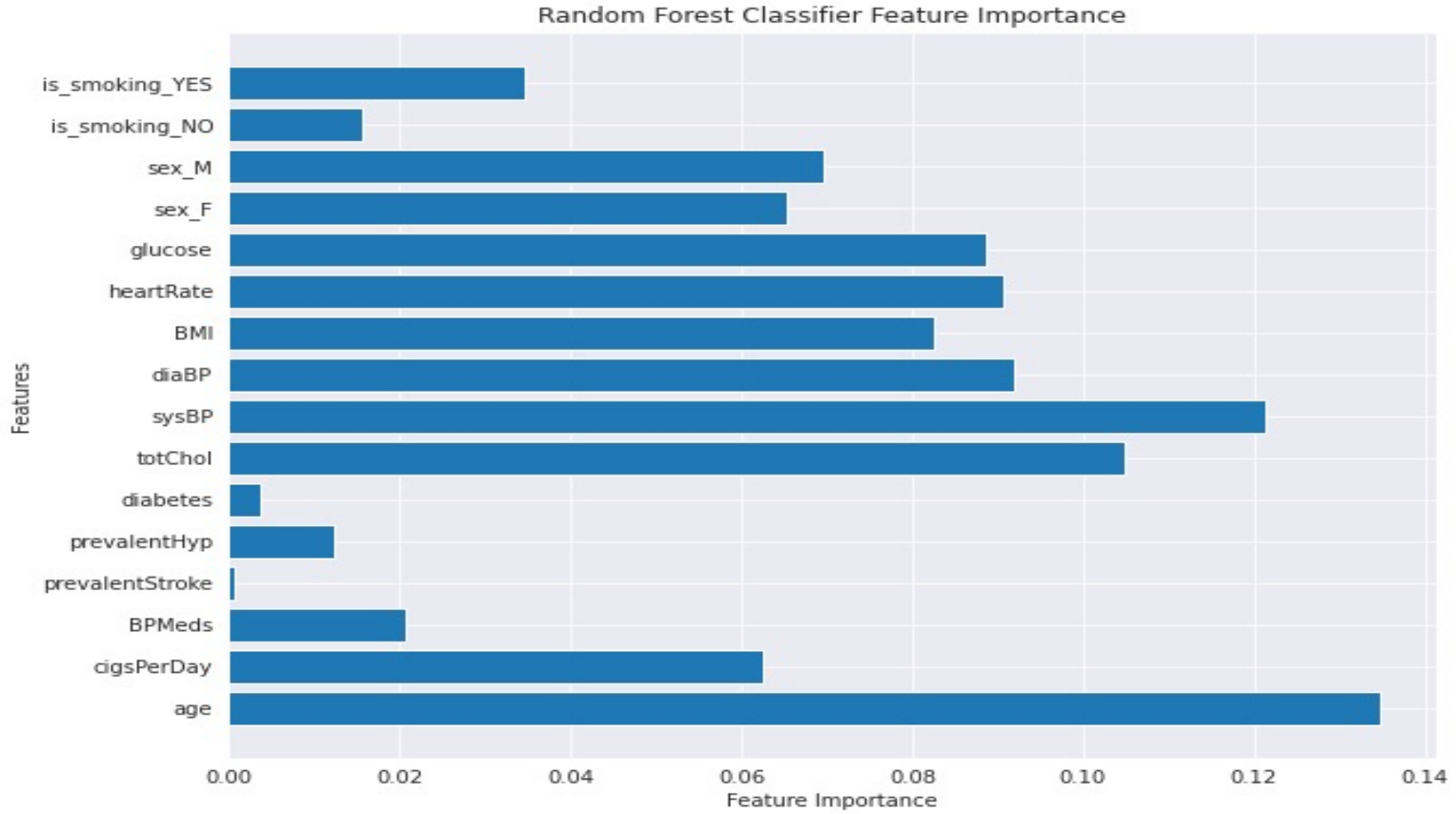
Random Forest Hyperparameters

- bootstrap=True
- max_depth=30
- max_samples=0.8,
- max_features_=auto
- min_samples_split=2
- n_estimators=100
- n_jobs=-1
- random_state = 0

Random Forest – confusion matrix



Random Forest – feature importance



Evaluation metrics

- **Accuracy** - The accuracy of a classification algorithm is one way to measure how often the algorithm classifies a data point correctly. Accuracy is the number of correctly predicted data points out of all the data points.

- **AUC ROC curve** - The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR(True positive rate) against FPR(False positive rate) at various threshold values and essentially separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

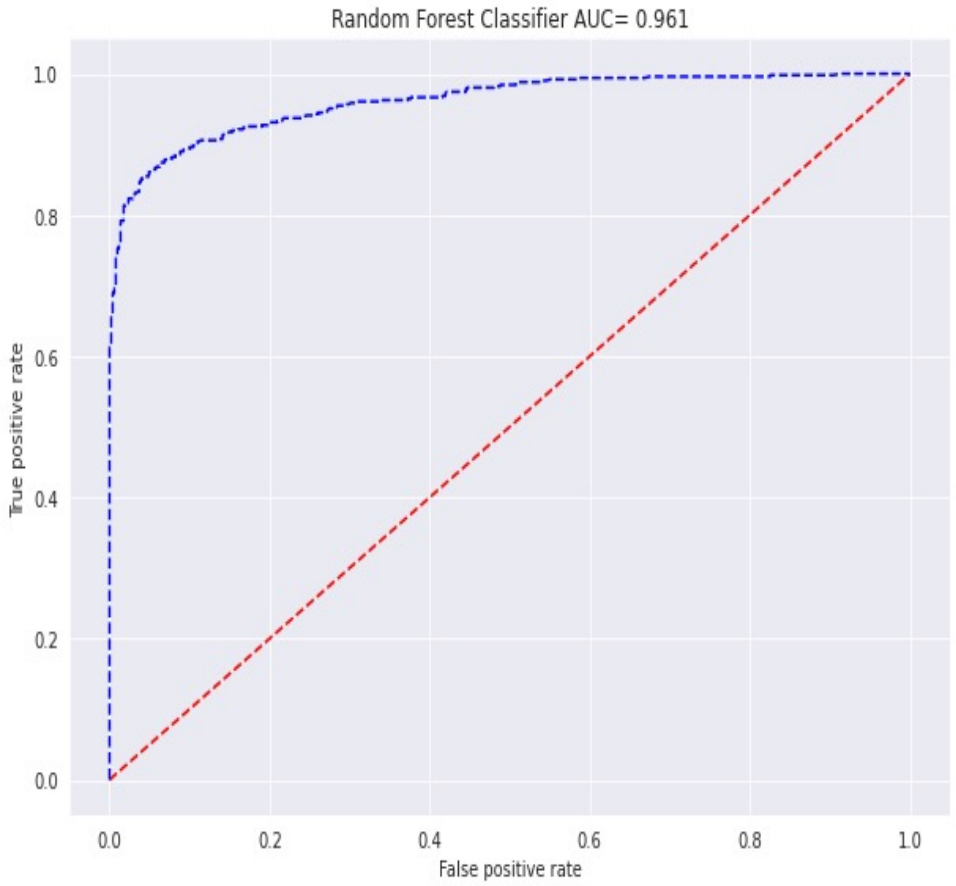
Evaluation metrics

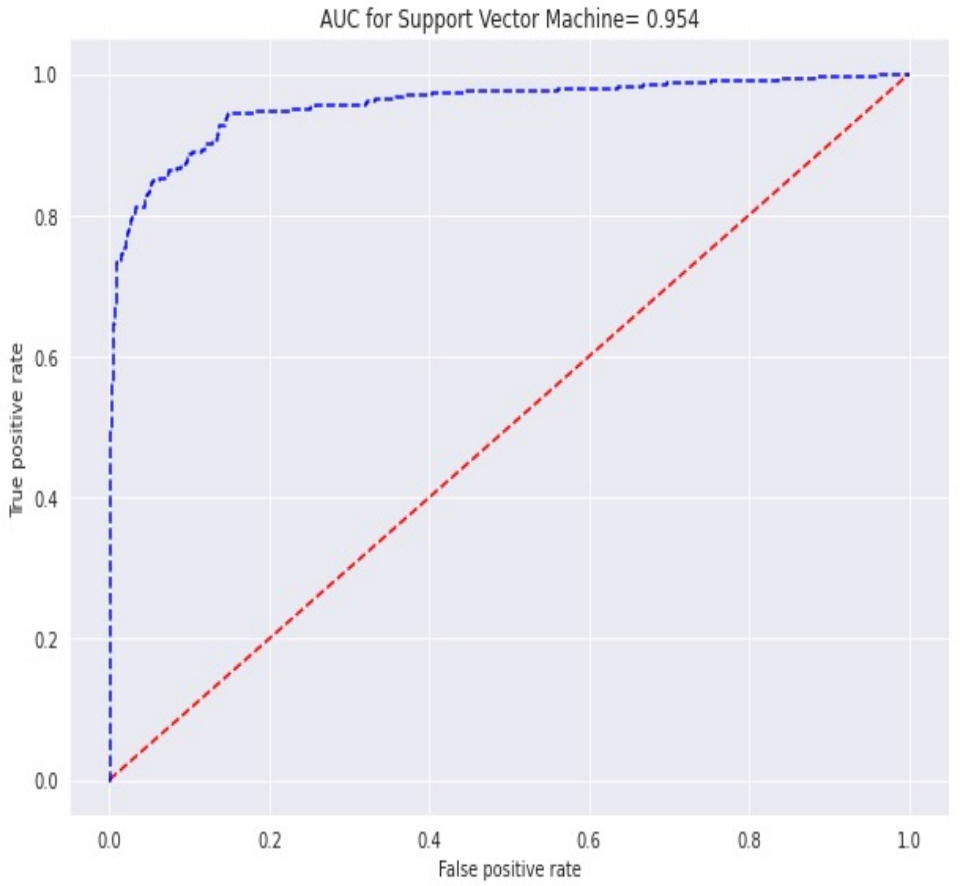
- **Precision** - It is the *number of true positives divided by the number of true positives and false positives*. Simply put, it is the number of positive predictions divided by the total number of positive class values predicted.

- **Recall** - It is *the number of True Positives divided by the number of True Positives and the number of False Negatives*. Simply put, it is the number of positive predictions divided by the number of positive class values in the test data.

- **F1 Score** - The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean.

AUC ROC Curve – Random forest vs SVM





Performance evaluation

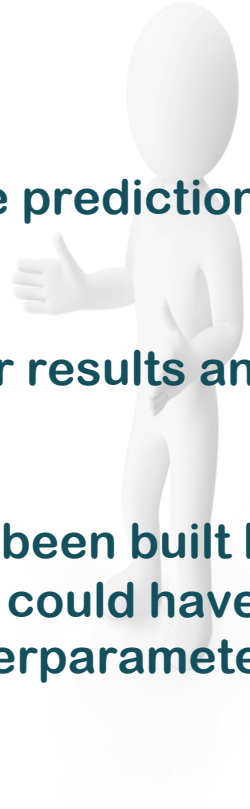
Model Name	Accuracy	AUC	F1 Score
Naive Bayes	0.640	0.752	0.543
K-nearest neighbours	0.872	0.872	0.881
XGBoost	0.879	0.948	0.875
Logistic regression	0.742	0.832	0.689
Decision trees	0.797	0.811	0.800
Support vector machine	0.887	0.954	0.888
Random forest	0.914	0.967	0.913

Conclusion

- We tried to fit various models to our dataset to predict the risk of Coronary Heart Disease for ten years.
- This was a class imbalanced dataset so we used SMOTE(Synthetic minority oversampling technique) which is a class imbalance handling technique before running our algorithms.
- Age, systolic blood pressure, total cholesterol and heart rate were some of the important features in the case of random forest classifier In medical field also they're considered the major contributors to the risk of CHD.
- Sex, age and whether the person smokes or not were the important features in the case of XGBoost.

Conclusion

- Naive Bayes classifier was not able to perform well in the prediction of target variable.
- Support Vector Machine and Random Forest gave similar results and were better than the rest of the models.
- If there were more data points better models could have been built but there's always scope for improvement. In our models we could have also used Principal component analysis, some more hyperparameter tuning and various other methods.



THANK YOU