CAPSTONE PROJECT - III

Cardiovascular Risk Prediction

By:

Shrinidhi Choragi

Data Science Trainee

Almabetter

Contents

- *Introduction*
- *Problem statement*
- *Dataset*
- *Exploratory Data Analysis*
- *Feature Engineering*
- *Data Preparation*
- *Hyperparameter Tuning*
- *Evaluation Metrics*
- *Data Modeling*
- *Conclusion*

Introduction

- Currently, cardiovascular diseases (CVDs) account for two-thirds of the total non-communicable disease burden in India.
- Cardiovascular disease risk reduction revolves around the major risk factors, including hypertension, diabetes, heredity etc. Although some risk factors, such as age and hereditary factors cannot be modified, lifestyle modification is key to preventing cardiovascular disease.
- Risk prediction models are the mathematical functions that predict the occurrence of an event of interest based on certain predictors, such as patient demographics, medical history, medication use, physical examination, disease characteristics, and heart laboratory values.
- The CVD risk approach is a cost-effective way to identify those at high risk, especially in a low resource setting.

Problem Statement

To objective is to understand the rationale for using cardiovascular risk prediction methods to make effective and appropriate risk factor treatment decisions in clinical practice.

The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD) and try to explore the following aspects.

- Understanding the impact of different risk factors.
- Studying the variation in risk due to different habits/medical history.
- Be able to interpret and analyze the prediction outcomes of the cardiovascular prediction model.

Dataset

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

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

The features of the dataset are:

- Sex: male or female("M" or "F")
- Age: patient's age;(Continuous)
- is_smoking: whether or not the patient is a current smoker (YES/ NO).
- Cigs Per Day: the number of cigarettes smoked on average in one day.

Dataset

- 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 has diabetes
- 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)
- Glucose: glucose level (Continuous)
- ***Ten_year_chd***: 10-year risk of coronary heart disease (Target Variable)

Exploratory Data Analysis

- Missing Value Analysis

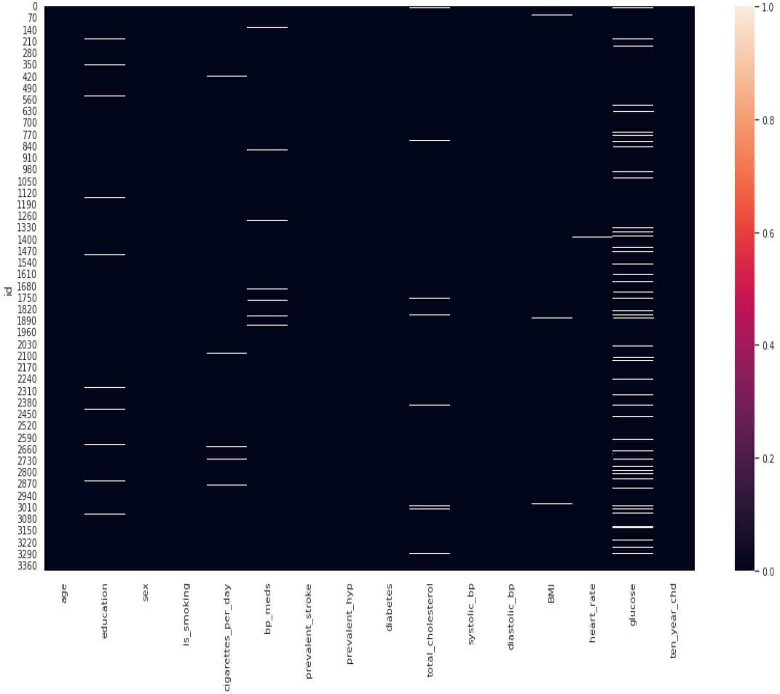
    - Categorical Variables

Imputed with respective mode values for features- education, bp_meds

    - Numerical Variables

Imputed with respective median values for features- cigarettes_per_day, total_cholesterol, BMI, glucose, heart_rate.

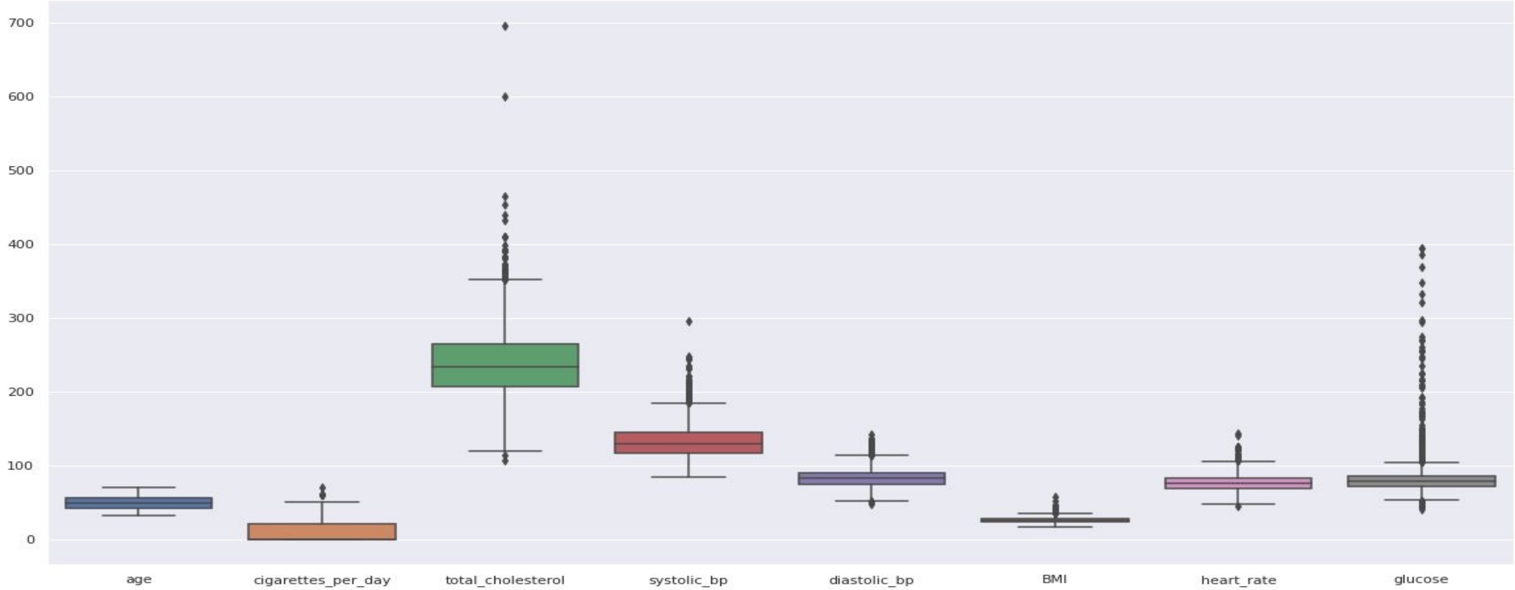
Note: Redundancies are to be taken care of while imputing with median values.



Exploratory Data Analysis

- **Outlier Analysis**

The risk factor outliers may contain important and clinically meaningful information. Therefore outliers are left untreated in this case.

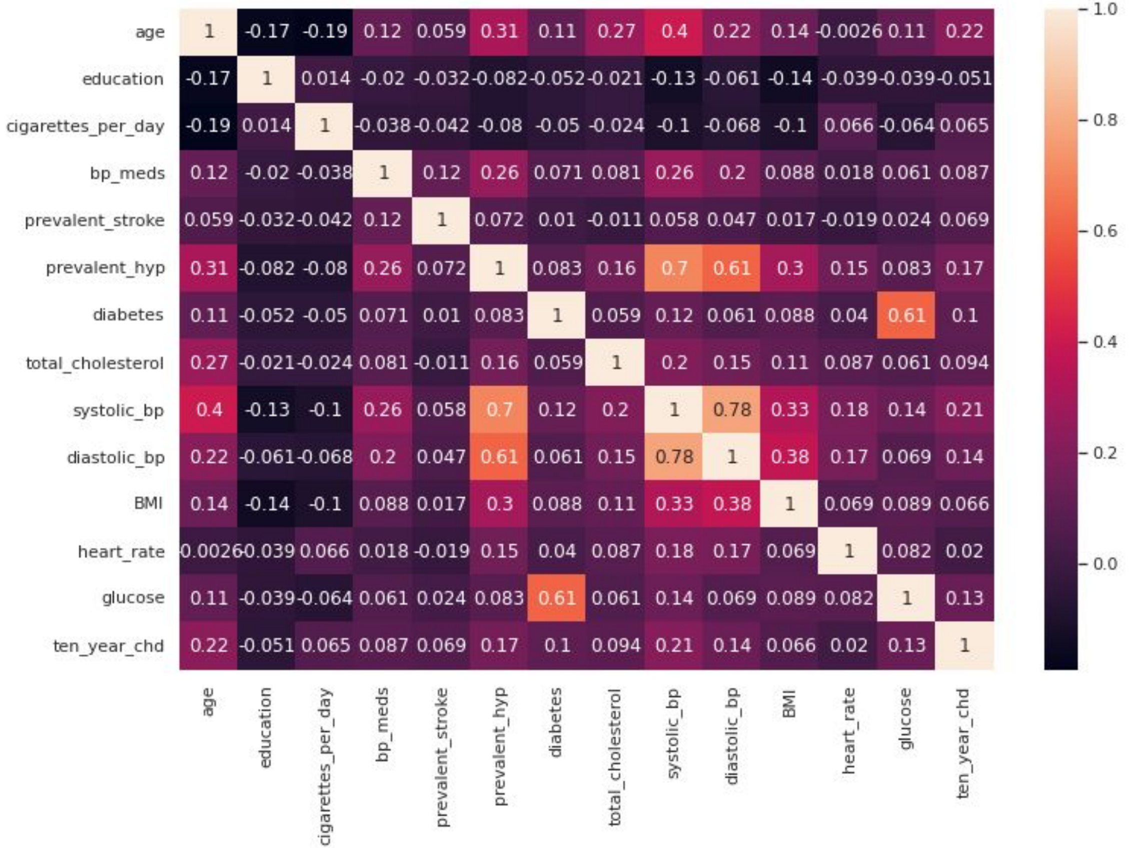


Exploratory Data Analysis

- <u>**Correlation Analysis**</u>

➢ No significant correlation between target variable and features.

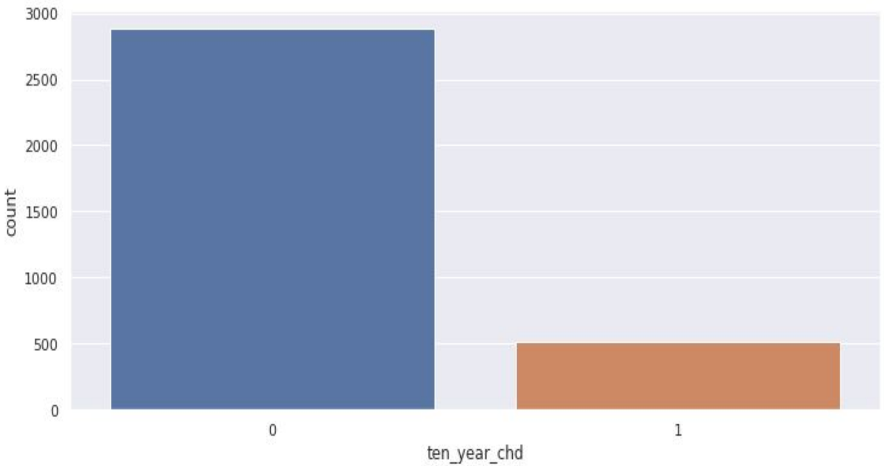
➢ Multicollinearity exists.



Exploratory Data Analysis

- **Dependent Variable**

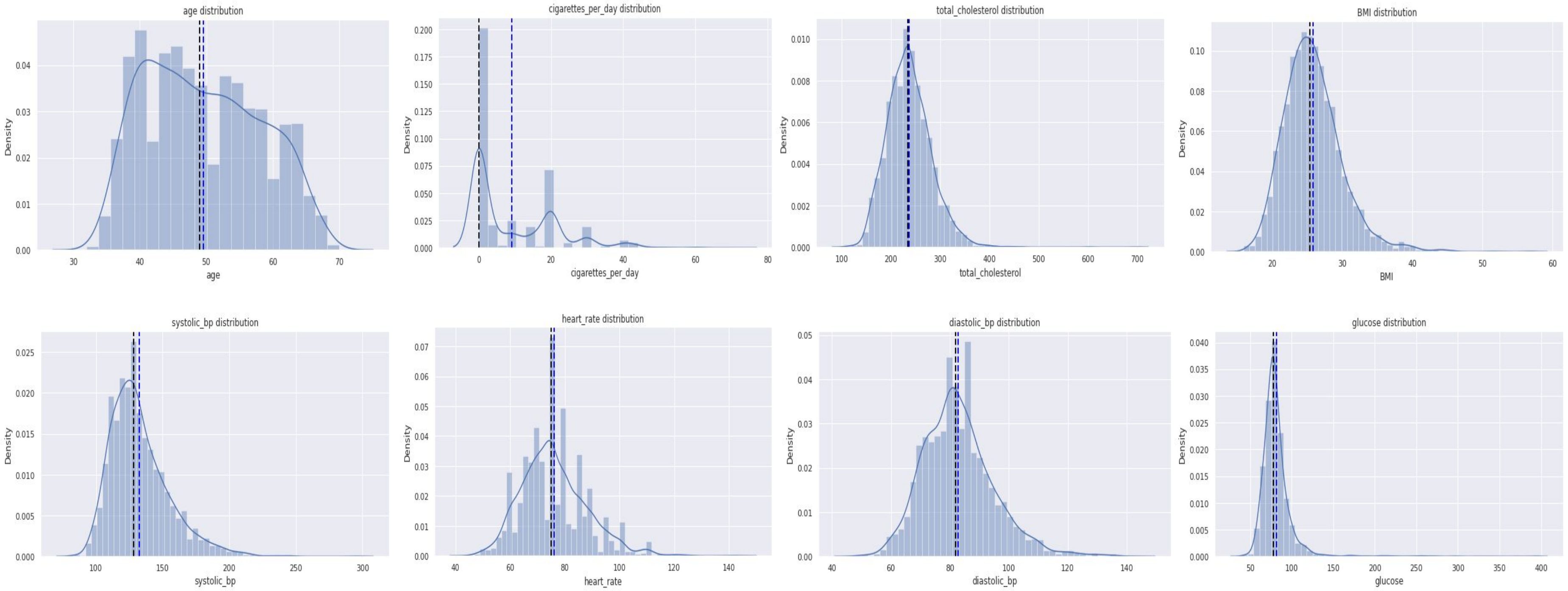
➢ It is binary (categoric)
➢ There exists data imbalance



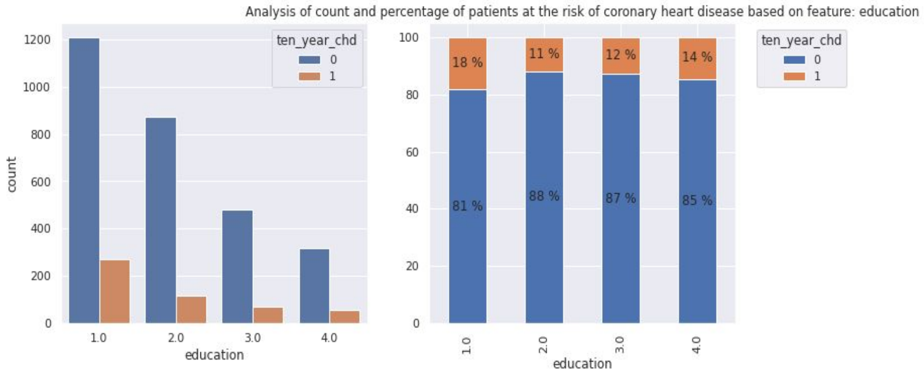
***Imbalanced Classification**: A classification predictive modeling problem where the distribution of examples across the classes is not equal.*

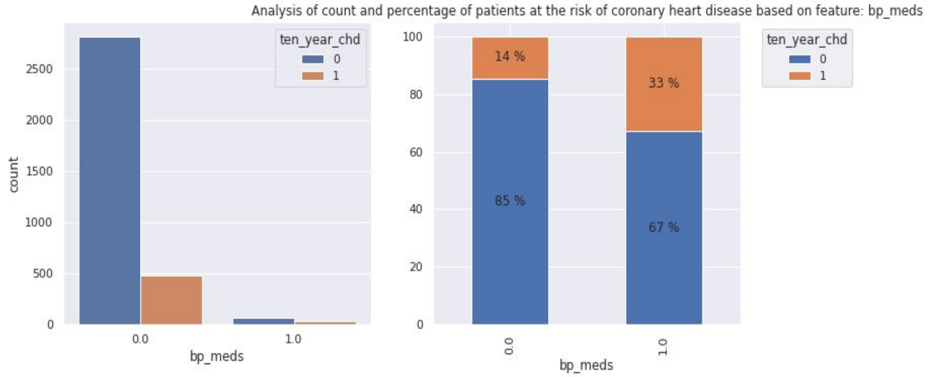
Solutions: Collect more data of minority class, choose appropriate metrics, resampling dataset, generate synthetic samples, threshold moving etc.

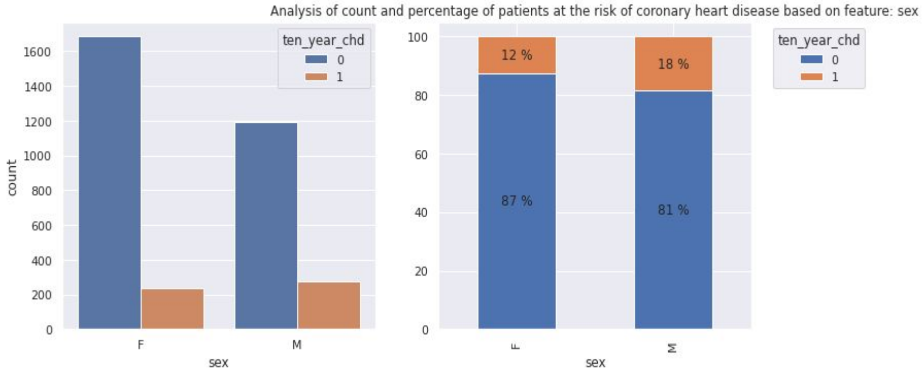
Exploratory Data Analysis

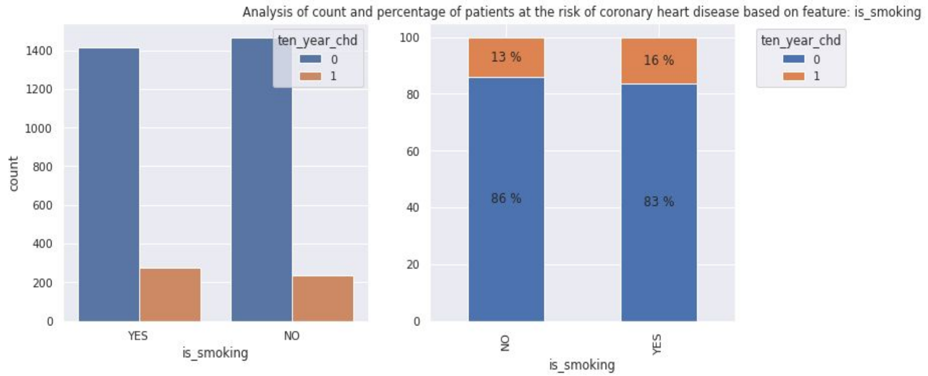


Exploratory Data Analysis

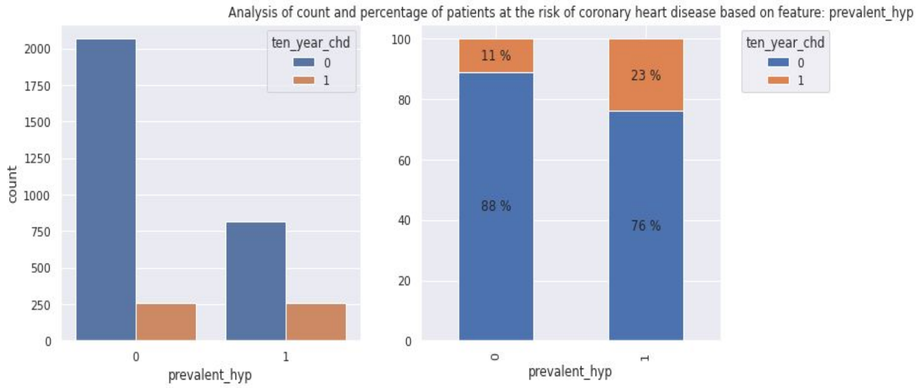


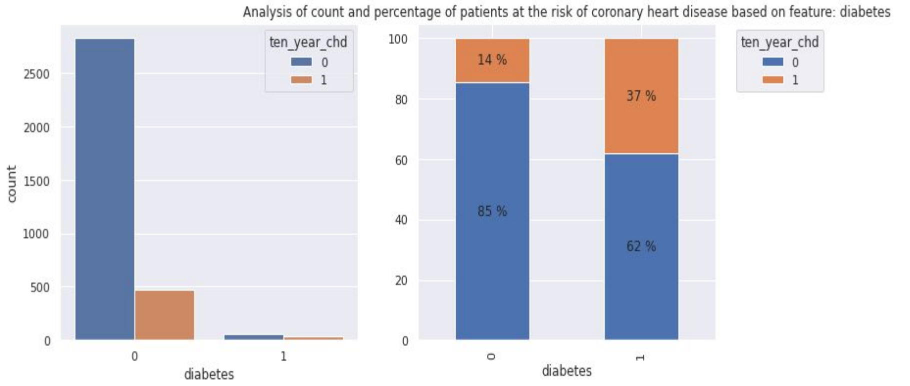


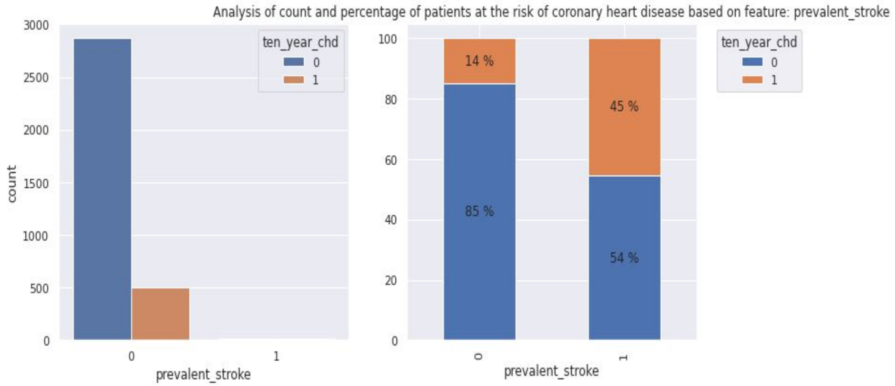


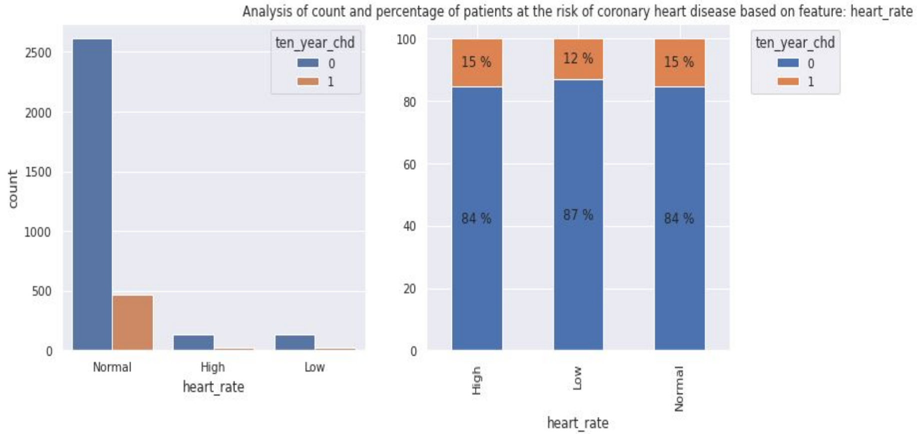


Exploratory Data Analysis

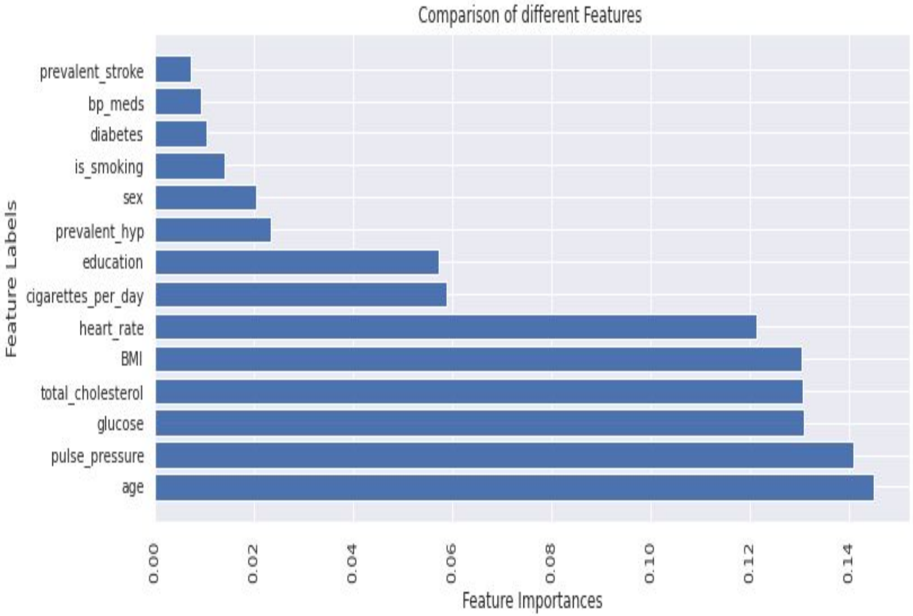




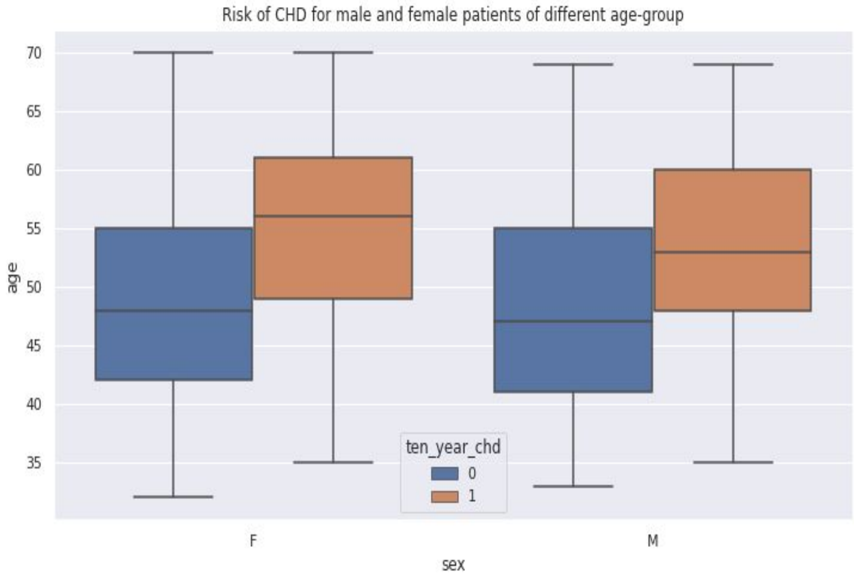




Exploratory Data Analysis



Age is the most important feature



The average age of risk of CHD is higher in female patients.

Feature Engineering

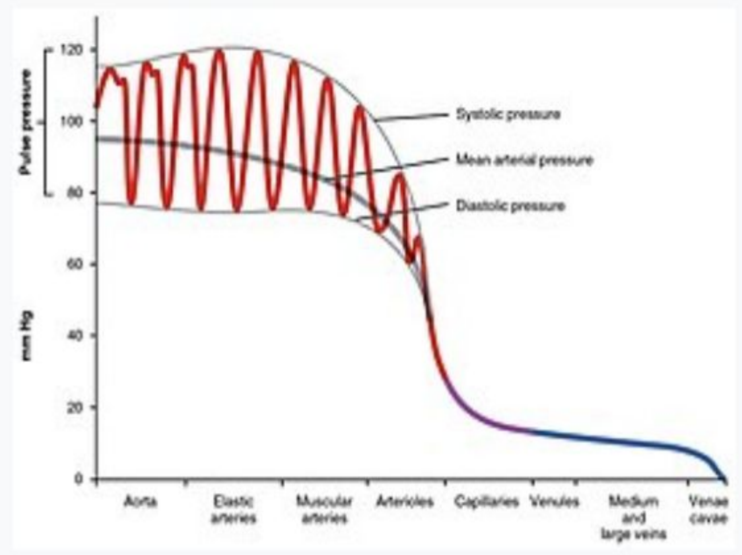
➢ **Feature Imputation**

Pulse pressure is the difference between the systolic and diastolic blood pressure.

It represents the force that the heart generates each time it contracts.

Pulse pressure tends to increase as one gets older, and it can also be an indicator of health problems before the symptoms are developed.

Pulse Pressure



Pulse Pressure = Systolic Blood Pressure - Diastolic Blood Pressure

Feature Engineering

➢ <u>Feature Selection</u>

Null Hypothesis (H0): Features are independent of the target variable.

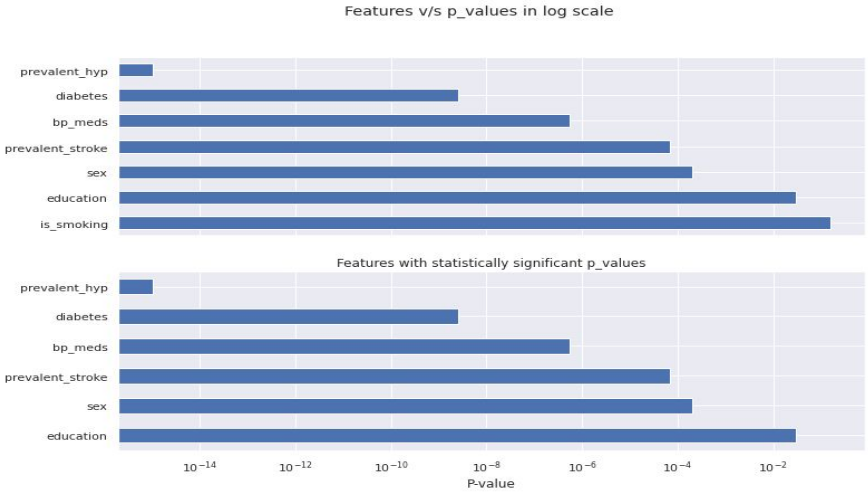
Alternate Hypothesis (H1): Features are dependent on the target variable.

- In feature selection, we aim to select the features which are highly dependent on the response variable.
- Scores based on statistical tests such as **Chi-Square** and **ANOVA F-test** /**F-Statistic**, provide a p-value, that is used to rule out some features.
- A **p-value** measures the probability of obtaining the observed results, assuming that the null hypothesis is true.
- The lower the p-value, the greater the statistical significance of the observed difference.

Feature Engineering

- **Chi-Square Test**

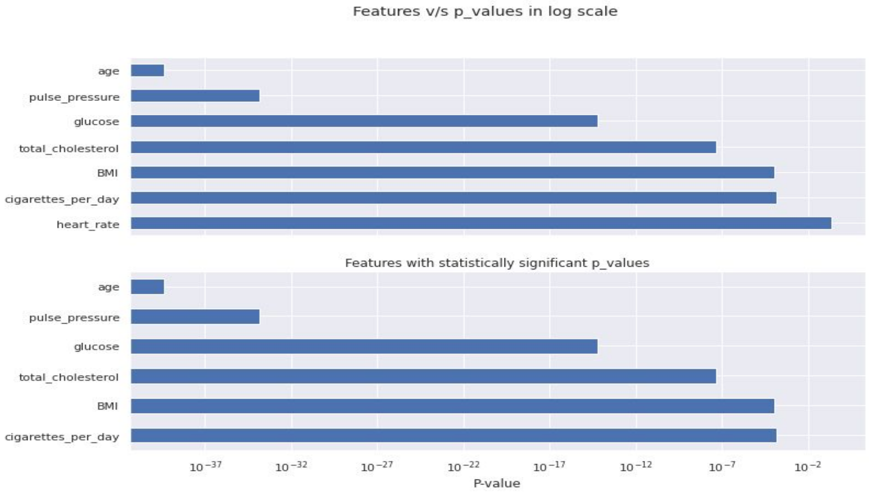
$$\chi^2 = \frac{(Observed\ frequency - Expected\ frequency)^2}{Expected\ frequency}$$



- **ANOVA F-Test**

$$F = (\chi_1^2 / n1 - 1) / (\chi_2^2 / n2 - 1)$$

Where χ_1, χ_2 are Chi distributions and n1,n2 are its respective degrees of freedom.



Data Preparation

➢ **<u>Skew Transformation</u>**

The skewed distributions are converted to a normal distribution using logarithmic and reciprocal transformation.

As a rule of thumb,

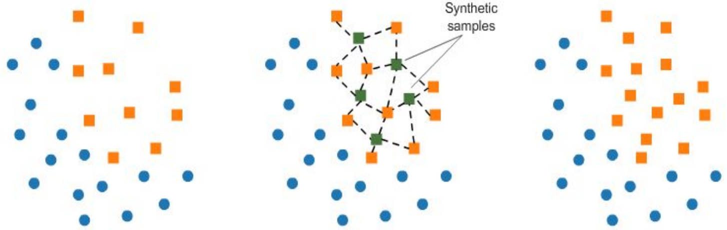
- If $skewness < -1$ or $skewness > 1$: the distribution is highly skewed.
- If $-1 < skewness < -0.5$ or $0.5 < skewness < 1$: the distribution is moderately skewed.
- If $-0.5 < skewness < 0.5$: the distribution is approximately symmetric.

➢ **<u>Data Splitting</u>**

The dataset is split into train and test data in the ratio of 70:30 resp.

Data Preparation

➢ **Handling Class Imbalance: Oversampling - SMOTE**



SMOTE: *Synthetic Minority Oversampling Technique*

- Choose a minority class as the input vector and find its k nearest neighbors.
- Choose one of these neighbors and place a synthetic point anywhere on the line joining the point under consideration and its chosen neighbor.
- Repeat the steps until the data is balanced.

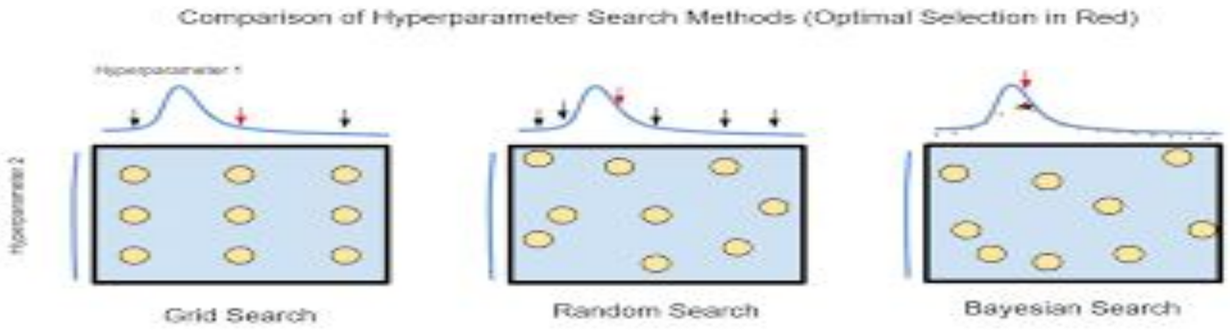
Note: Typically undersampling/oversampling techniques will be done on train split only, this is the correct approach. In order to avoid using synthetic data for testing purposes.

➢ **Feature Scaling**

Standard Scaler	$\frac{x_i - \text{mean}(x)}{\text{stdev}(x)}$

Hyperparameter Tuning

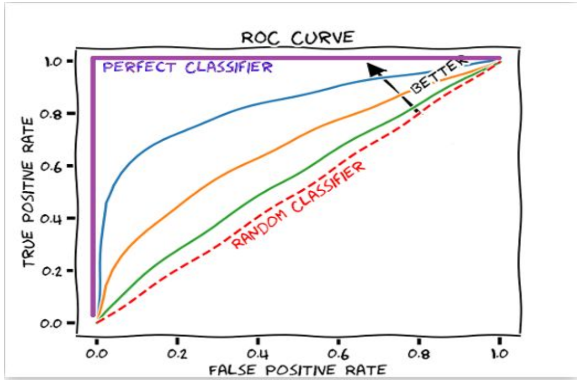
- ➢ ***Grid Search***: Exhaustive and computationally expensive, used when hyperparameter search space is restricted.
- ➢ ***Random Search***: Larger search spaces, improved models compared to Grid Search
- ➢ ***Bayesian Optimization***: A hyperparameter optimization process based on a probabilistic model, often the Gaussian Process.



Evaluation Metrics

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$= \frac{True\ Positive}{Total\ Actual\ Positive}$$



ROC Curve

Confusion Matrix

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Data Modeling

The following models have been studied and implemented on the given dataset:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine
- K-Nearest Neighbor

Model	Hyperparameter_tuning	Train-Recall	Test-Recall
Logistic Regression	GridSearhCV	0.7016	0.6846
Logistic Regression	RandomSearchCV	0.6977	0.6644
Logistic Regression	BayesSearchCV	0.7041	0.6644
Decision Tree	GridSearhCV	0.8468	0.8322
Decision Tree	RandomSearchCV	0.7558	0.6913
Decision Tree	BayesSearchCV	0.8468	0.8322
Random Forest	GridSearhCV	0.7479	0.6846
Random Forest	RandomSearchCV	0.6987	0.6644
Random Forest	BayesSearchCV	0.7618	0.7047
SVM	GridSearhCV	0.7225	0.6779
SVM	RandomSearchCV	0.7131	0.6577
SVM	BayesSearchCV	0.9861	0.2282
K-Nearest Neighbor	Manual	0.8165	0.6107

Logistic Regression

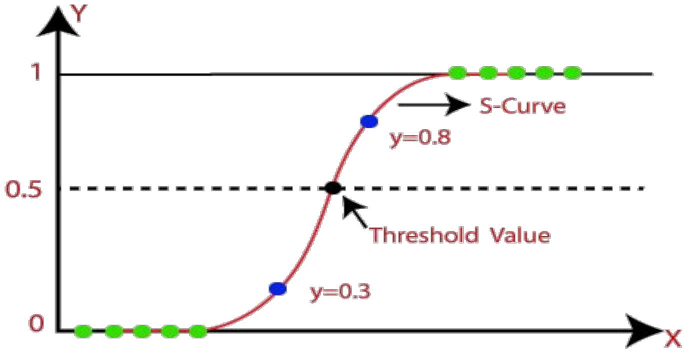
➢ ***Hyperparameter Tuning***: Grid Search CV

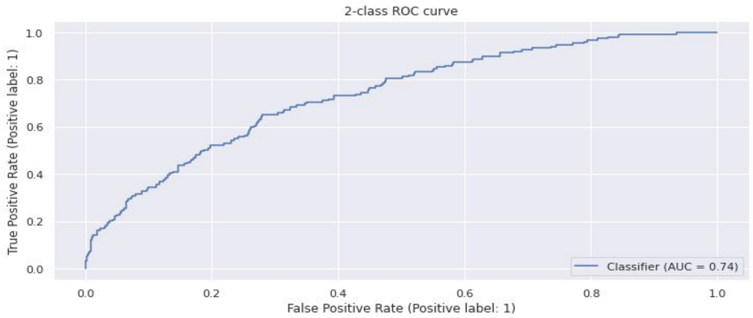
➢ ***Best parameters***

'C': 0.01,
'class_weight': 'balanced',
'max_iter': 10,
'penalty': 'l2'

➢ ***Evaluation results***

- Train Recall : 0.7016
- Test Recall: 0.6845
- ROC AUC : 0.74

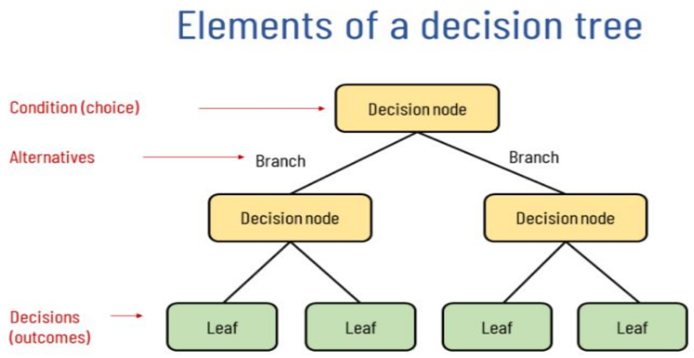




Decision Tree Classifier

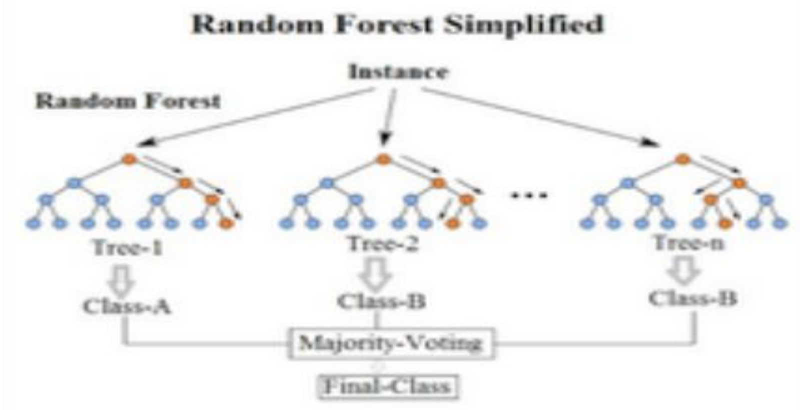
- ➢ ***Hyperparameter Tuning***: Bayes Search CV
- ➢ ***Best parameters***
  Class_weight: 'balanced'
  Criterion: 'entropy'
  Max_depth: 4
  Min_samples_leaf: 0.12
  Min_samples_split: 0.83

- ➢ ***Evaluation results***
  - Train Recall : 0.8468
  - Test Recall: 0.8322
  - ROC AUC : 0.64

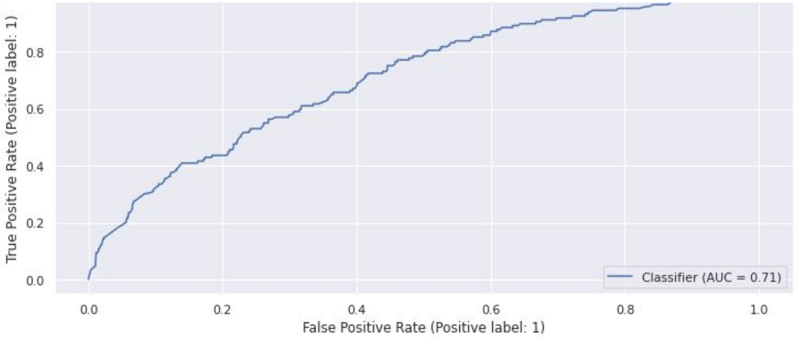


Random Forest Classifier

- ***Hyperparameter Tuning***: Bayes Search CV
- ***Best parameters***
  No. of estimators: 100
  Criterion: 'entropy'
  Max_depth: 1
  Min_samples_leaf: 0.1
  Min_samples_split: 0.1



- ***Evaluation results***
  - Train Recall : 0.7618
  - Test Recall: 0.7046
  - ROC AUC : 0.71



Support Vector Machine

➢ ***Hyperparameter Tuning***: Grid Search CV

➢ ***Best parameters***

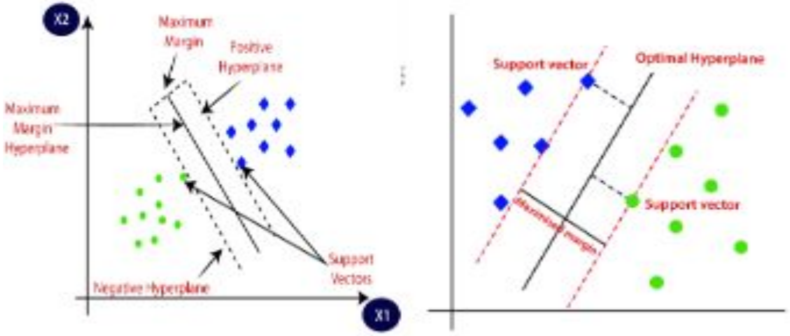
'C': 1

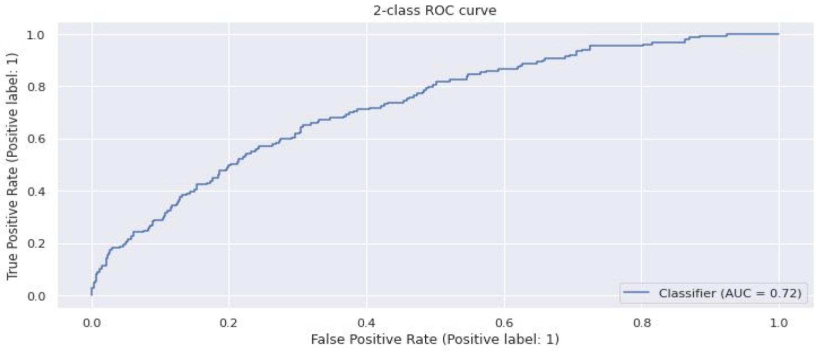
Gamma: 0.001

Kernel: ‘rbf’

➢ ***Evaluation results***

- Train Recall : 0.7299
- Test Recall: 0.6778
- ROC AUC : 0.72





K-Nearest Neighbor

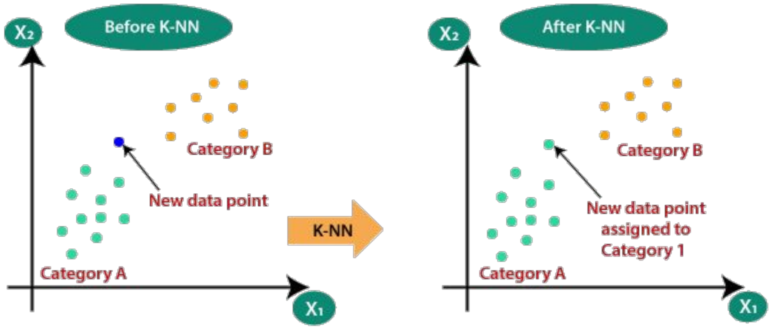
➢ ***Hyperparameter Tuning***: manual

➢ ***Best parameters***

Optimal k =59

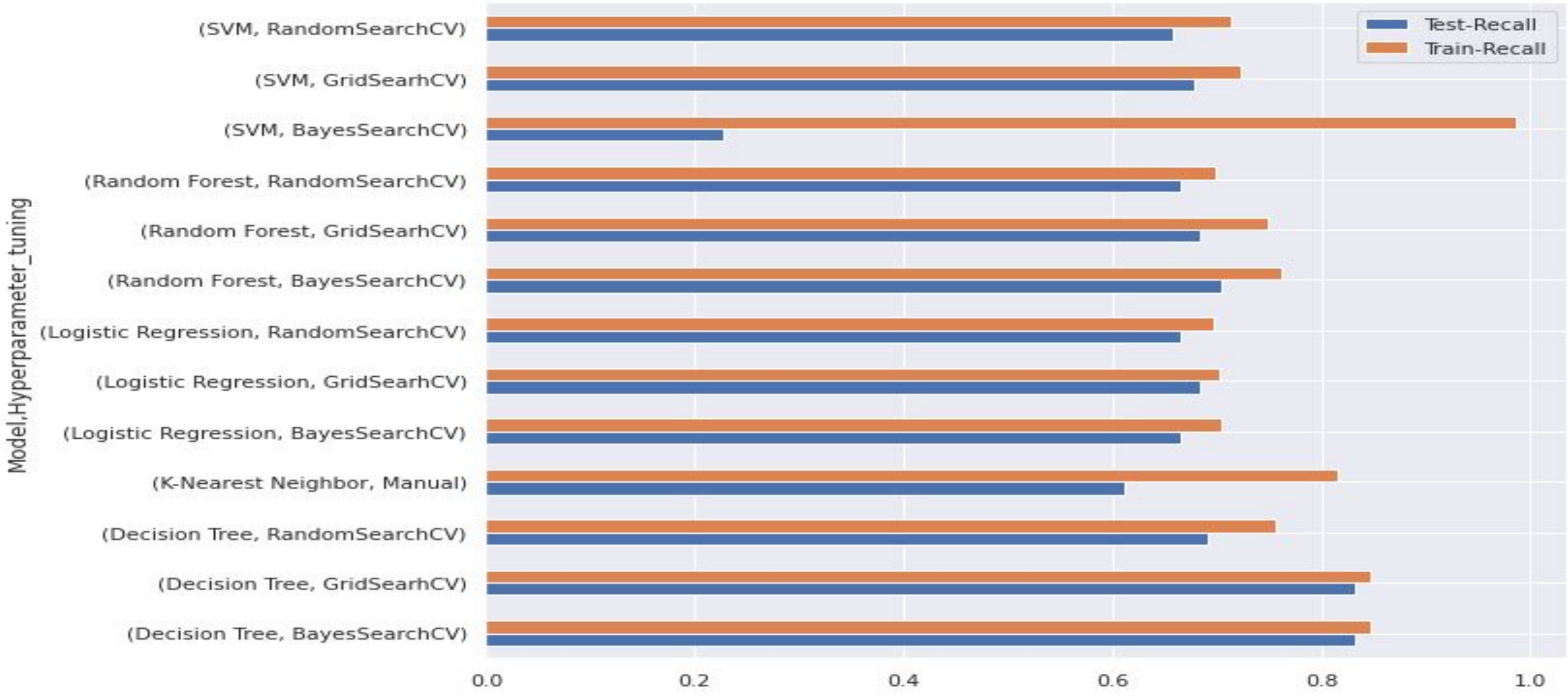
➢ ***Evaluation results***

- Train Recall : 0.8165
- Test Recall: 0.6107
- ROC AUC : 0.67



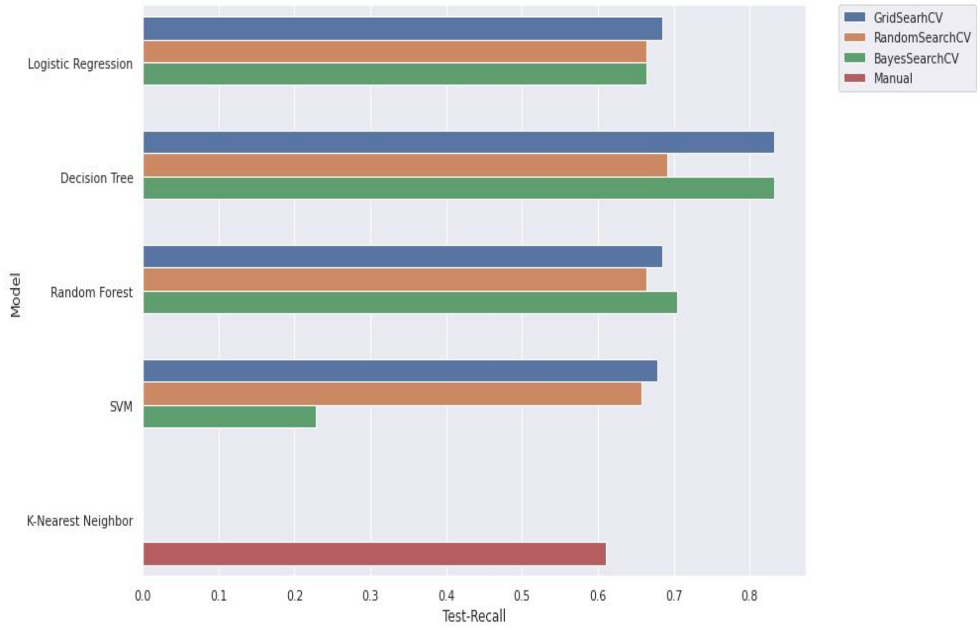
Summary

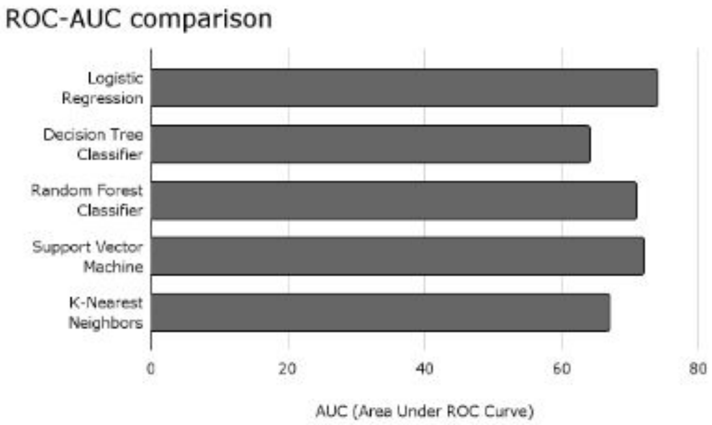
<u>*Recall Scores*</u>



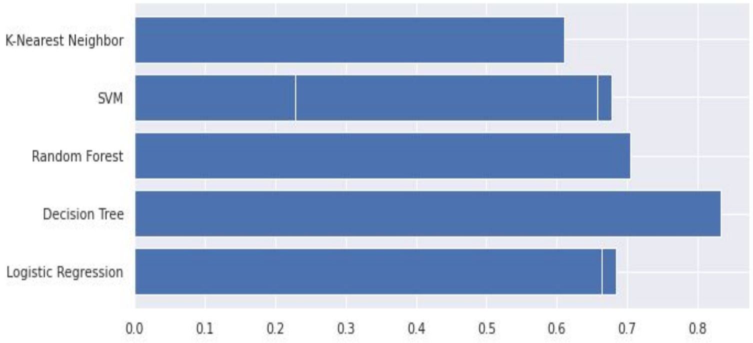
Summary

Test Recall Scores





Test Recall Scores for best models



Conclusion

- ➢ Considered to maximize the recall score while having a reasonable ability to distinguish between the classes as indicated by the ROC_AUC score.
- ➢ The conclusion is that the ***Random Forest Classifier*** model with *Bayesian optimization*, Recall-score (0.7046), (ROC_AUC= 0.71) can be considered the most optimal model.
- ➢ Model calibration ('goodness of fit') is a more clinically relevant measure of model performance as clinician and patient want to know if the predicted risk resembles the actual risk.
- ➢ Since it seems right to apply the oversampling techniques only on train split and go for prediction rather than including the synthesized samples in test data to overestimate the results.
- ➢ However future improvements may include attempting to collect more data on minority class that helps the analysis to be improved further.

Thank you