CAPSTONE PROJECT - 3

Cardiovascular Risk Prediction

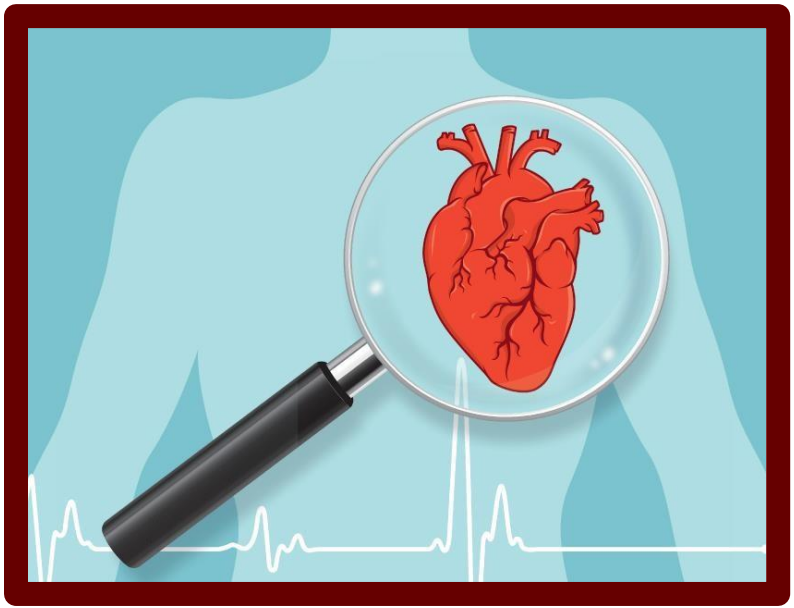
Created By : Sakshi Dhyani

Project Details

In this project, the dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The target is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).

Steps performed

- Data cleaning
- Data visualizations
- Data preprocessing
- Model Implementation
- Evaluation metrics

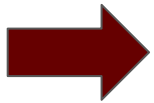


Data Summary

Independent Features

Sex :	Male or Female
Age :	Age of the patient
Is_smoking :	whether patient smokes or not
Cigs Per Day :	average no of cigarettes that person smoked on one day
BP Meds:	Whether patient is on blood pressure medication or not
Prevalent Stroke:	Whether patient previously had a stroke
Prevalent Hyp:	Whether patient was hypertensive or not
Diabetes:	Whether or not patient has diabetes
Tot Chol :	Total Cholesterol level
Sys BP :	Systolic Blood pressure
Dia BP :	Diastolic Blood pressure
BMI :	Body Mass Index
Heart Rate	
Glucose :	Glucose level

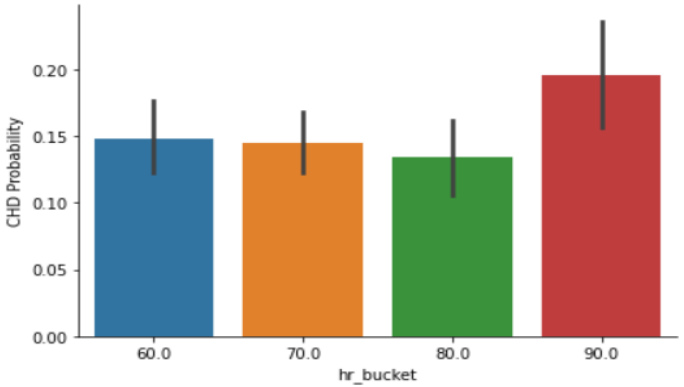
Dependent Feature



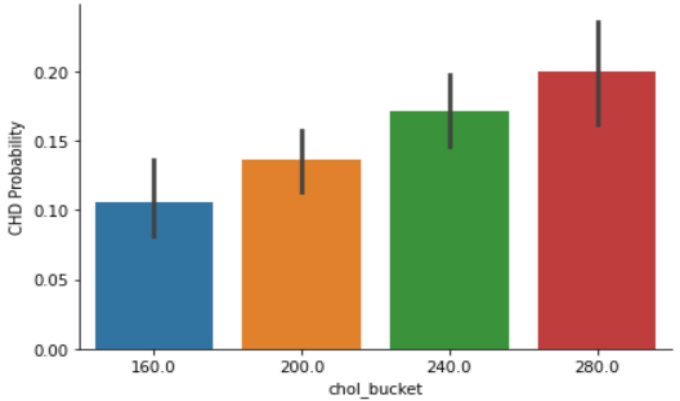
10- year risk of coronary heart disease (CHD)

Feature Engineering

Impact of heart rate on the target variable: People with high heart rates have a high risk of having CHD in the next 10 years.

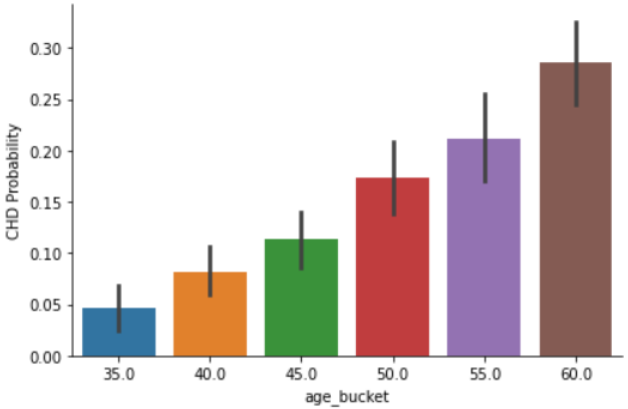


Impact of cholesterol level on the target variable: People with high cholesterol levels have a high risk of having CHD in the next 10 years.

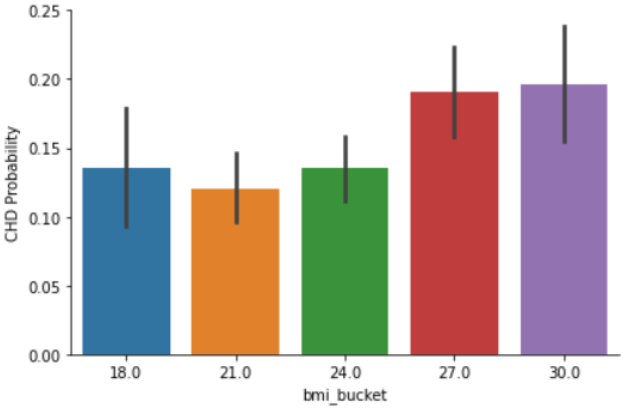


Feature Engineering

Impact of age on the target variable: Older people have a high risk of having CHD in the coming 10 years.

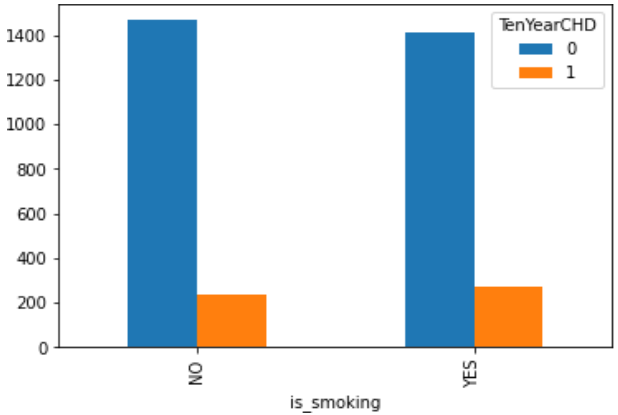


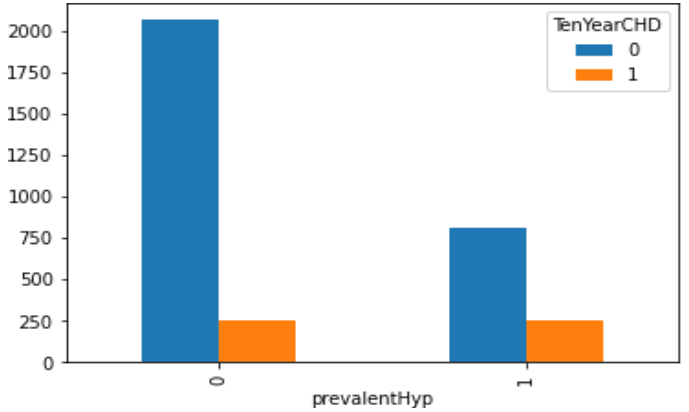
Impact of Body Mass Index on the target variable: People with high body mass index have a high risk of having CHD in the next 10 years.

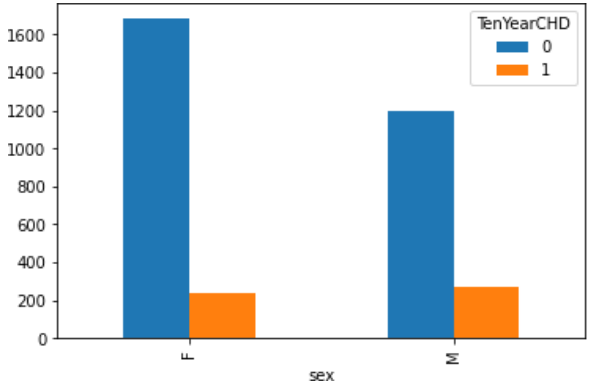


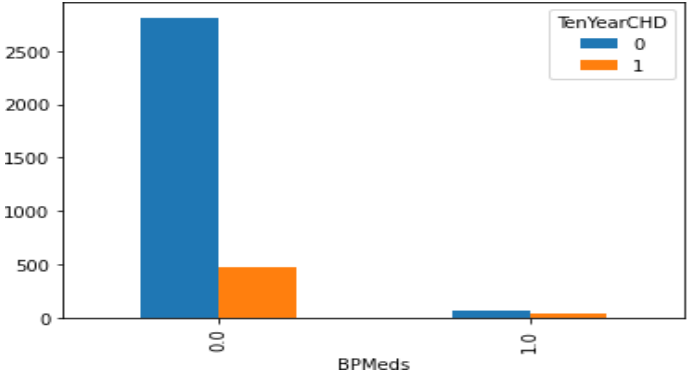
Exploratory Data Analysis

Bar Plots to study the impact of predictors on Coronary Heart Disease risk value



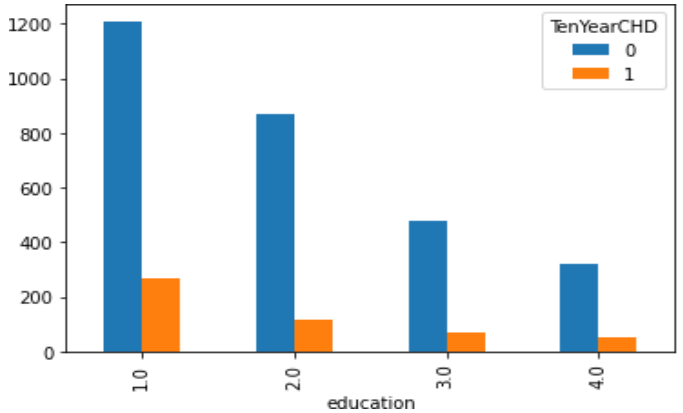


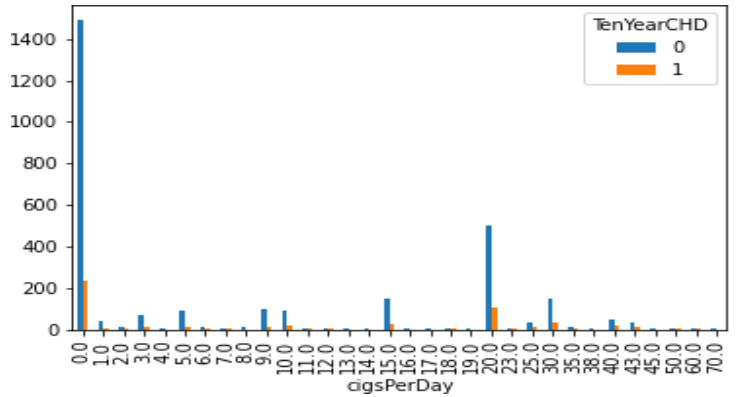


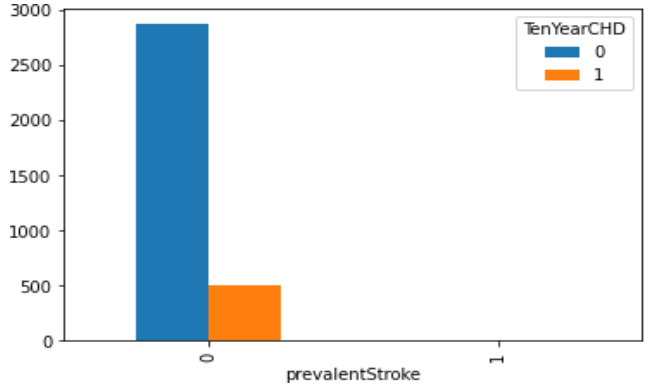


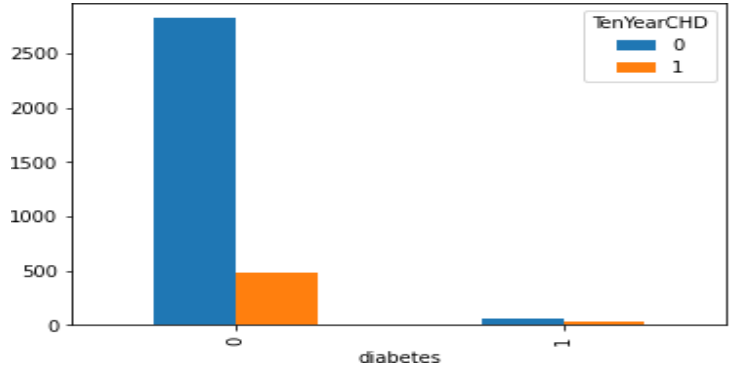
Exploratory Data Analysis

Bar Plots to study the impact of predictors on Coronary Heart Disease risk value









Data Cleaning

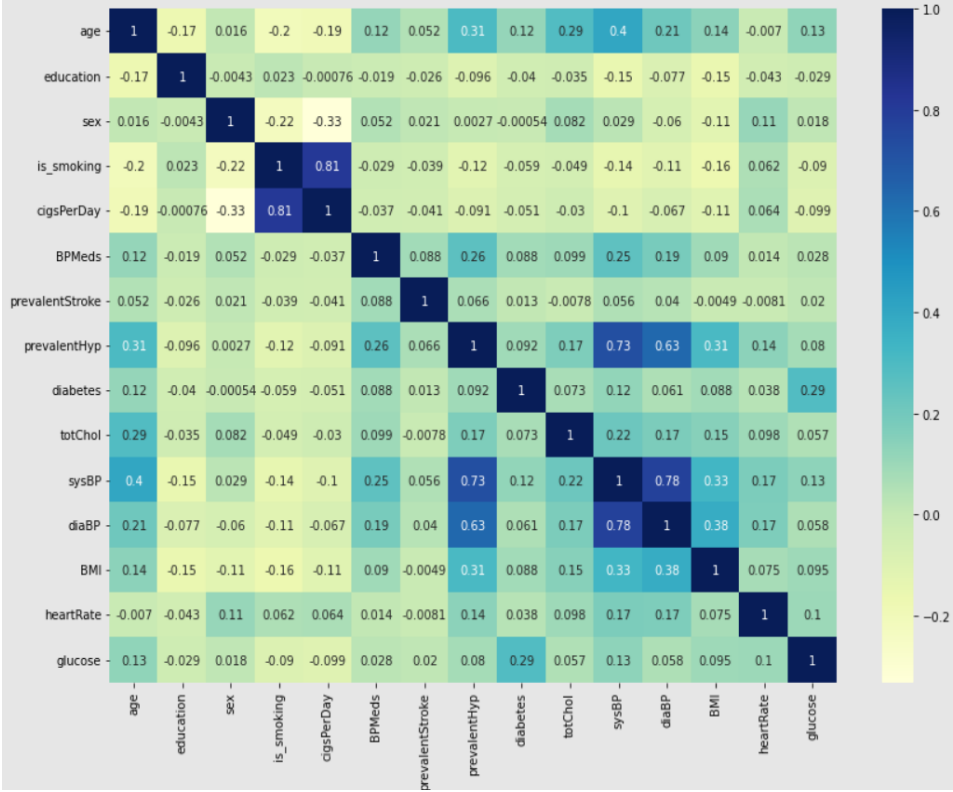
Dealing with missing values for categorical and numerical predictors:

Categorical Variables: To fill up the missing values of data in the case of categorical variables I have used a simple imputer that imputes the null values with a label that is most frequent in that feature column.

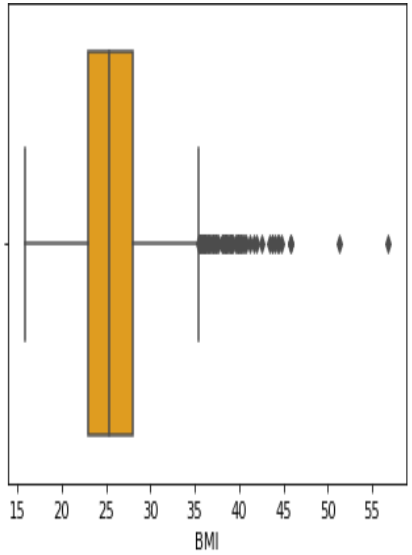
Continuous Variables: To treat the null values in continuous variables, we use KNN imputer which uses an unsupervised clustering algorithm to come up with values of the features.

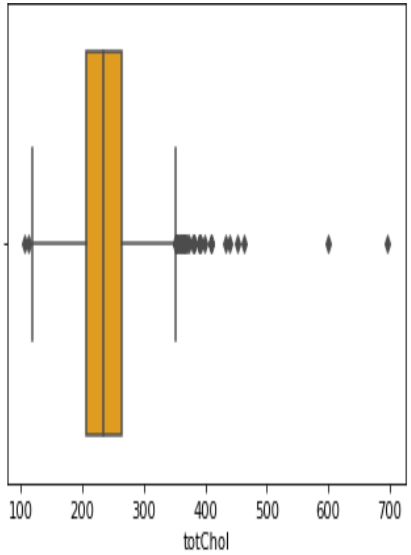
Feature Selection

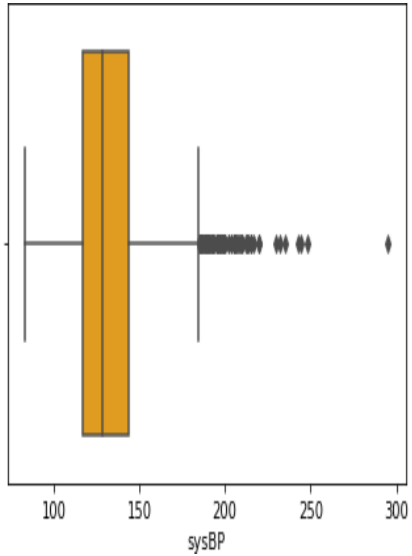
- **Cigs Per Day and is_smoking are highly correlated.**
- **SysBP and Prevalent Hyp are highly correlated.**
- **DiaBP and SysBP are highly correlated.**
- **Combined SysBp and DiaBP to denote a new feature pulse rate.**

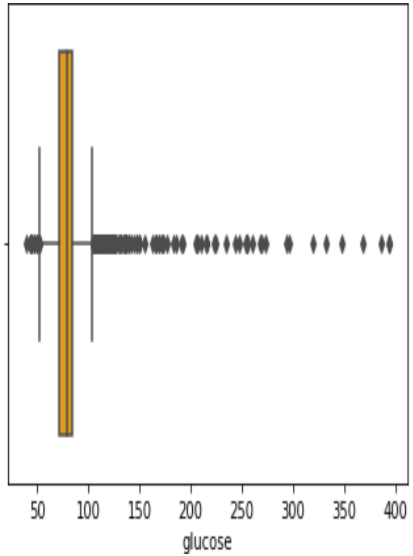


Outlier Detection using Box Plot

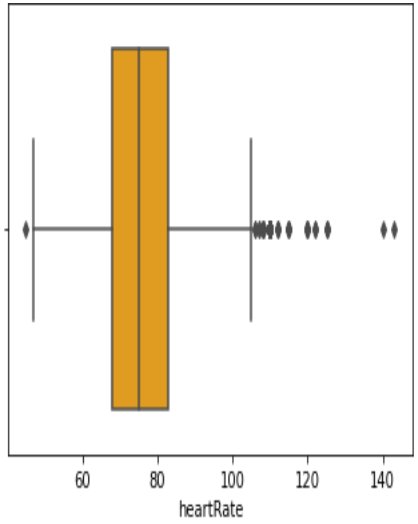


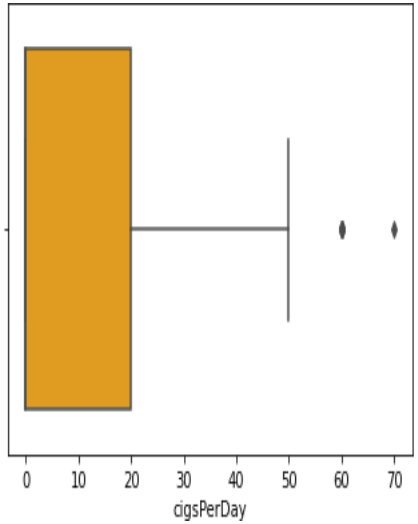


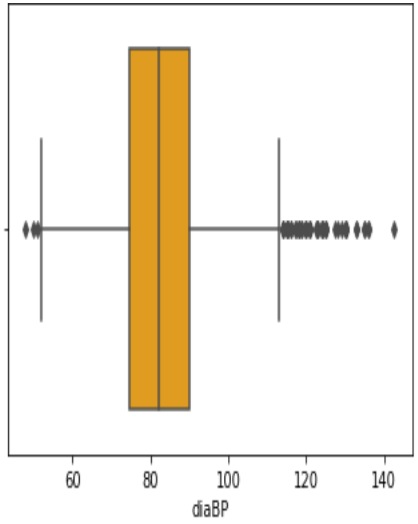


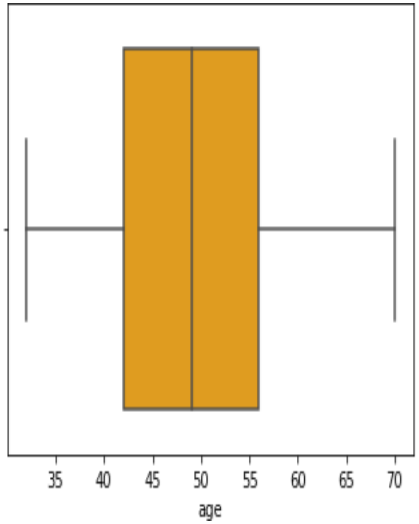


Outlier Detection using Box Plot









Data preparation

Initial Training Data Shape - 2712 Rows and 17 Columns (1 column is dependent)

After resampling dataset using SMOTE technique:

Training Data – 3328 Rows and 17 Columns (1 column dependent)

Test Data – 678 Rows and 17 Columns (1 column dependent)

Dependent column will be predicted as that is the target variable named “Ten Year CHD.

Random Forest Classifier

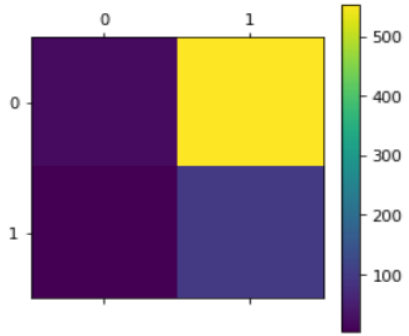
Evaluation Metrics for test data: **Choosing Recall as an evaluation metric as the goal is to reduce the false negatives.**

Recall Score: 96 %

Used Randomized Search CV for hyperparameter tuning which increased recall score from 65% to 96%

Confusion Matrix

Random Forest is not an interpretable model. Let us see the recall score of other interpretable models in further slides.



Logistic Regression

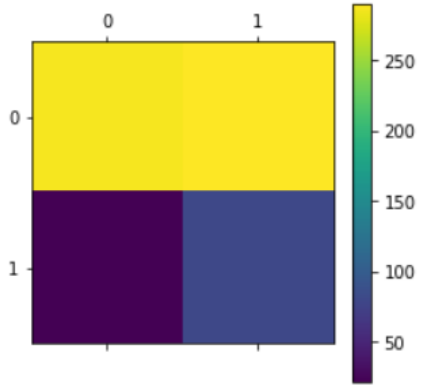
Evaluation Metrics for test data: **Choosing Recall as an evaluation metric as the goal is to reduce the false negatives.**

Recall Score: 78 %

Used class weight as balanced.

Logistic Regression is an interpretable model. Let us see the recall score of other interpretable models in further slides.

Confusion Matrix



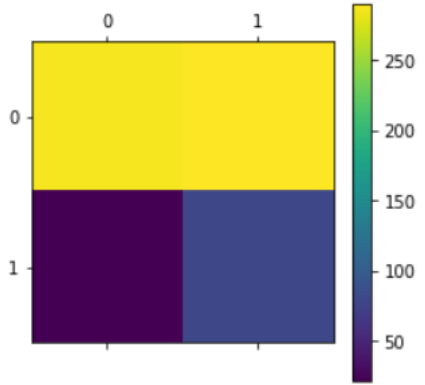
Decision Tree Classifier

Evaluation Metrics for test data: **Choosing Recall as an evaluation metric as the goal is to reduce the false negatives.**

Recall Score: 75 %

Logistic Regression is an interpretable model. Let us see the recall score of other models in further slides.

Confusion Matrix



Support Vector Classifier

Evaluation Metrics for test data: **Choosing Recall as an evaluation metric as the goal is to reduce the false negatives.**

Recall Score: 78 %

The recall score is the same as the Logistic Regression Model but Support Vector Classifier is not an interpretable model. It depends on our business case whether interpretability should be considered to select the best model.

Conclusion

I considered Recall score as the evaluation matrix as our task was to reduce the false-negative value so that patients do not get detected incorrectly and are proved safe. This can cause a huge impact on the patient health.

Resampling of data was performed as the data was not balanced. Imbalance data can give high accuracy but recall, precision, and F1 score need to be taken care of in such cases.

Performed missing value imputation using KNN-imputer and carried out data treatment for outliers.

Implemented SMOTE boosting to oversample the minority class observations to address the class imbalance issue.

Used the insight from EDA engineered newer features like - pulse pressure, age bucket & BMI bucket that helped to explain the separation in the Risk.

Owing to the parametric relationship in the data - Implemented a Logistic Regression model and was able to achieve a Recall of 78%. For SVM, also the recall score was the same but SVM is not an interpretable model and as per this scenario, I decided to select an interpretable model.

THANK YOU