Heart Failure Prediction

By Akshat Jain and Sumit Kumar Prajapati
B20AI054 and B20CS074

Deployed Website: https://cardium.herokuapp.com/
Github Repository: https://github.com/akshatjain1004/heart-stroke-prediction

Abstract: This paper reports our experience with building a Heart Failure predictor i.e. a classifier based on past training data provided to us. Our dataset consists of 12 columns: 11 Features and 1 label with a mix of continuous and categorical data. We implemented a machine learning pipeline with data importing, encoding of categorical data, standard scaling and predicting using a trained model. We also performed data analysis and evaluated our chosen model against a number of evaluation metrics.

1. Introduction

Cardium: Heart Disease Predictor:-

- Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.
- People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

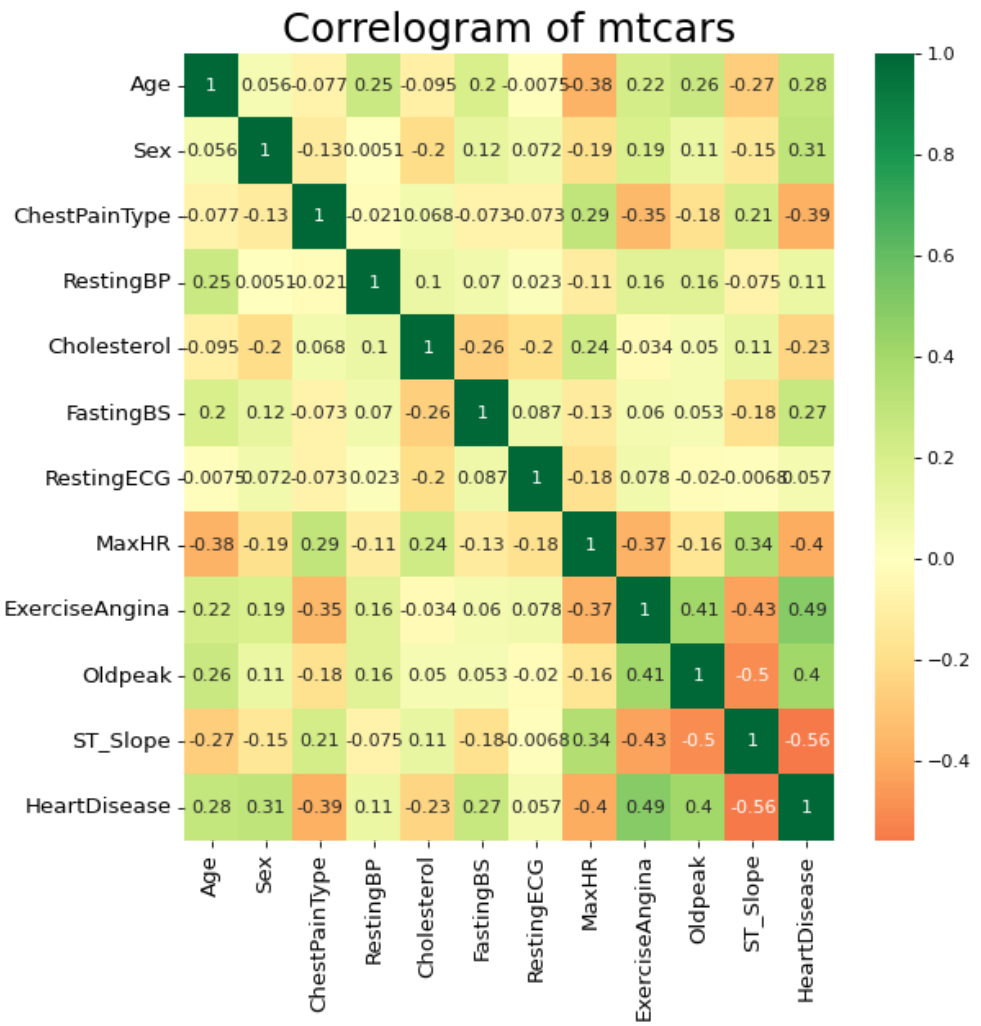
Dataset Overview:-

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]

9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

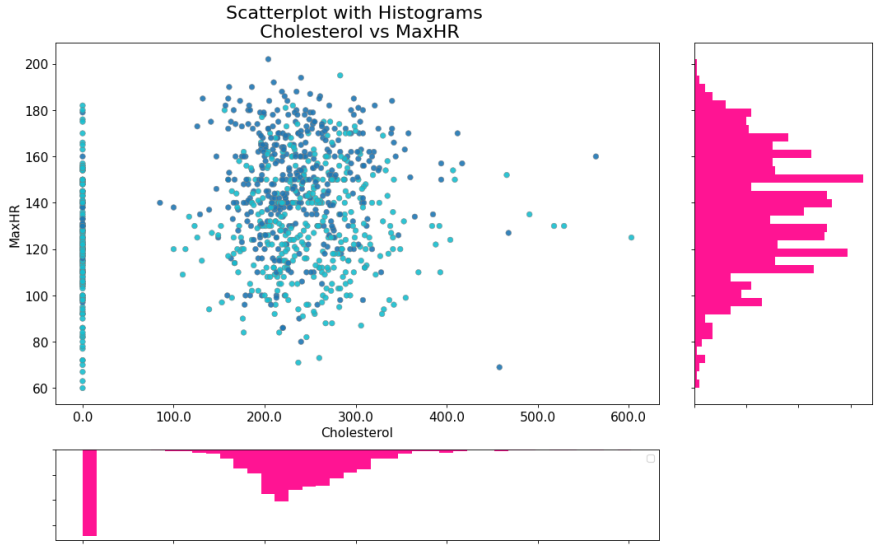
2. Data Analysis and Exploration

A. Correlogram:



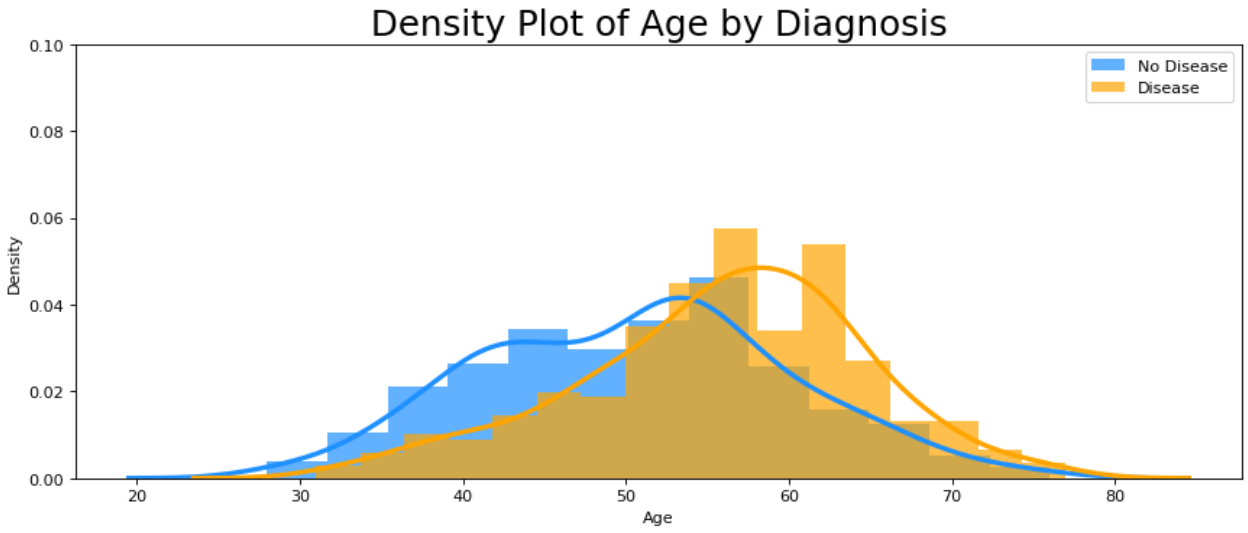
This analysis shows the correlation values between different features. In layman's terms, a good positive correlation value (close to 1) suggests that on increasing 1 feature/column the other will also increase similarly, a negative correlation suggests inverse relation. For our study, last row is very important as it tells us how different features affect probability of Heart Disease

B. Histogram and Scatter Plot:



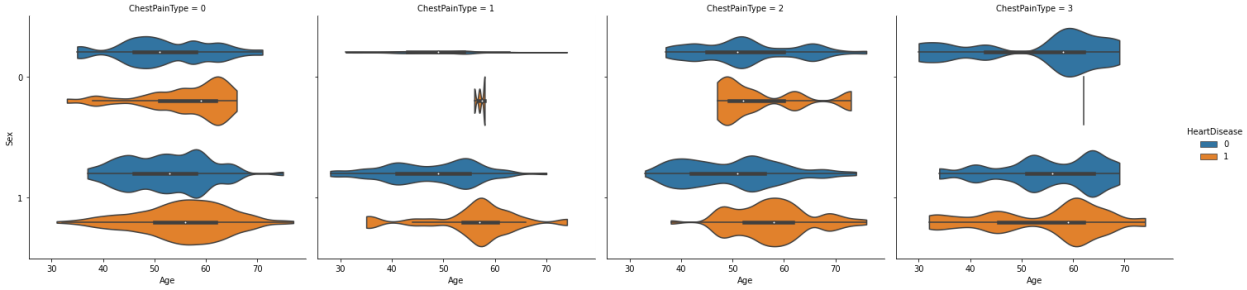
This plot shows the relation between Cholesterol and Max. Heart Rate for the two classes (disease and no-disease) through the scatter plot. And their respective distributions through the histogram.
Dark Blue Dots: No Disease
Light Blue Dots: Disease

C. Density Plot of Age by Diagnosis



This plot shows the analaysis of 'Age' column of the dataset and shows it's density distribution for the two classes.

D. Categorical Violin Plot



Above is a violin plot to check the distribution of positive and negative samples across different 'Sex', 'ChestPain' and 'Age'. The higher width/amplitude of violin suggests more number of samples at that particular parameters.

E. Population Waffle Chart

Above is a waffle chart and shows the distribution of population in terms of categorical variables:

1. Heart Disease
2. Chest PainType and
3. Resting ECG

3. Methodology

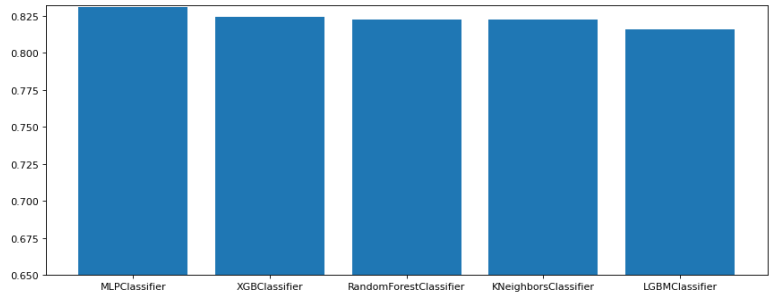
- Next we scaled the continuous data values using Standard Scalar library from sklearn.
- After which we trained a set of classification models and analysed their cross-validation scores:

Models used:-

1) MLP Classifier
2) XGB Classifier
3) RandomForest Classifier
4) LightGBM classifier
5) K-Neighbour Classifier

- **Results:**

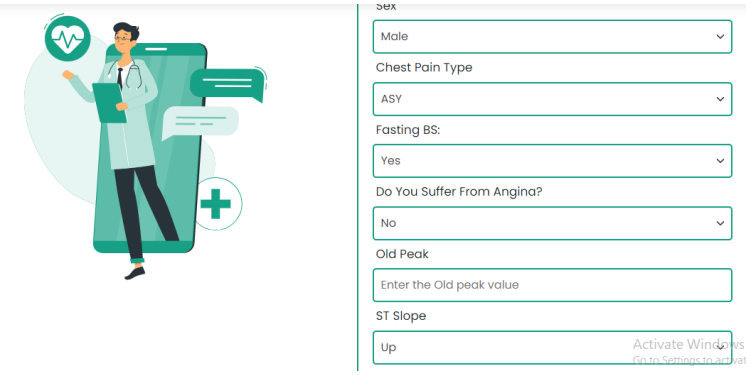
	Model	CV Score
0	MLPClassifier	0.831082
1	XGBClassifier	0.824543
2	RandomForestClassifier	0.822369
3	KNeighborsClassifier	0.822375
4	LGBMClassifier	0.815823



Based on the cross-validation we conclude that Multi-Layer Perceptron is the best performing model

- **Sequential Feature Selection:** We picked the 7 top most features using Sequential Feature Selection, the usage of this step is two-fold:
    1. Improve the Accuracy of the Model
    2. Reduce the number of Required user input

  We have also kept in mind the second point while deploying our engine on **website** and kept certain default values for parameters which are not usually known to the user. These values have been chosen from the data analysis done above.

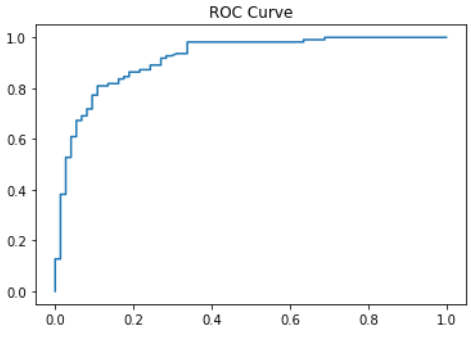


CV Score after SFS= 0.85068

- **Random Search CV:** To get the best set of hyper-parameters for our best performing classifier i.e. MLP.
  CV Score after Random Search CV= 0.8899782135076253

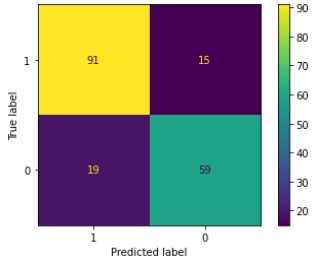
3. Results and Model Performance:

A. ROC-AUC Curve:



AUC Score= 0.8919533

B. Confusion Matrix:

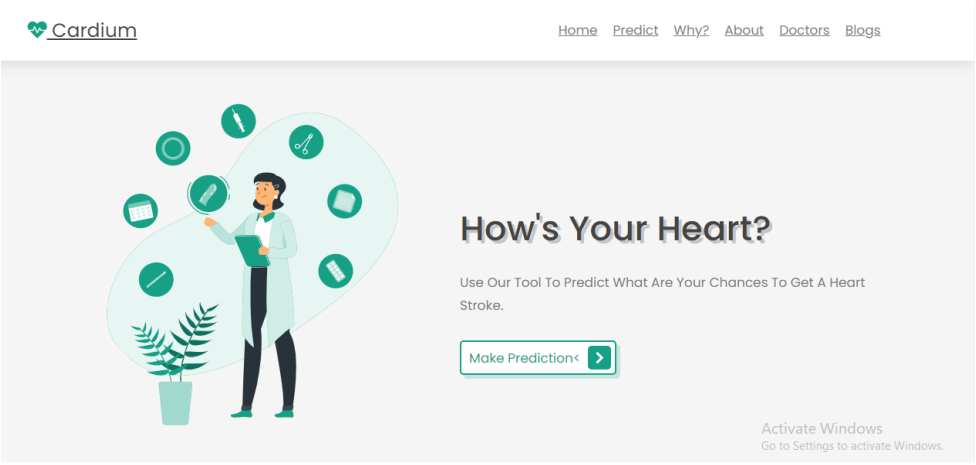


C. Precision, Recall and F1-Score:

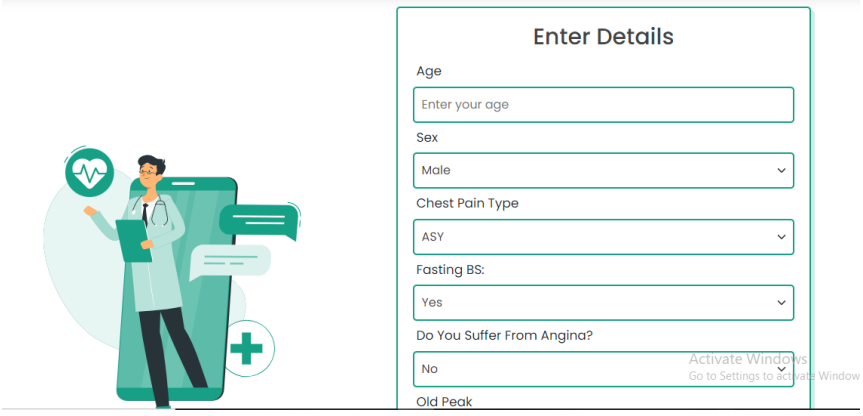
Metric	Value
Precision	**0.88**
Recall	**0.8**
F1-Score	**0.838095**

Overall Accuracy= 0.8152173913043478

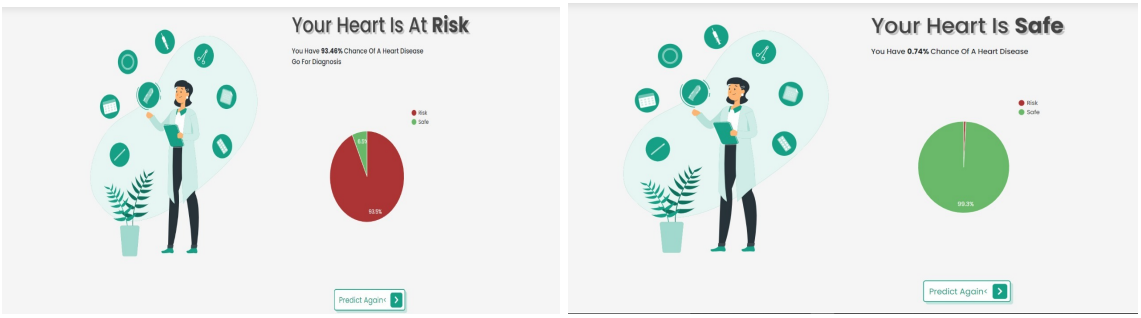
4. Website and Functionalities



Fully Responsive UI, with other information such as relevant stats, famous cardiologists and related Blogs



Simple Form with default values for uncommon parameters



Once we submit our relevant details, it renders a pie chart, predicts the probability and gives advice based on the result.

Contributions:-

Akshat Jain (B20AI054): Performed data Analysis and made the plots, Trained and checked LGBM Classifier, Performed cross-validation analysis of all the models, Performed Sequential Feature Selection and Random Search CV to improve the model performance. Performed Model Performance analysis using metrics like AUC-ROC curve, Confusion Matrix, Precision, Recall and F1 Score. Designed and created a fully responsive frontend with functionalities like Nav-bar, footer, quick links, blogs, About, relevant stats and famous cardiologists. Wrote the final report for the project.

Sumit Kumar Prajapati (B20CS074): Performed pre-processing of data i.e. importing, encoding of categorical variables and standard scaling the continuous data, Trained the classification models like MLP, XGB, RandomForest, and KNN, Synced frontend with backend using Flask framework by requesting appropriate data from the form and displaying a pie-chart and probability of heart disease, Created ML Pipeline for saving the model as a serialised binary file which can be used later, Deployed the model on Heroku.

References

[1] Feature Selection | Machine Learning Mastery
[2] XGB Regressor | XGBoost Documentation
[3] Random Forest Regressor | Level Up Coding
[4] LightGBM | Analytics Vidhya
[5] KNN Regressor | Analytics Vidhya
[6] RandomSearchCV and GridSearchCV | Machine Learning Mastery
[7] Matplotlib Plots for Data Analysis | Machine Learning +