Heart Failure Prediction

PRML Course Project Report

PROBLEM STATEMENT

Cardiovascular disorders(CVDs) are a common cause of heart failure. People who have cardiovascular disease or are at high cardiovascular risk due to one or more risk factors such as hypertension, diabetes, hyperlipidemia, or previously existing disease require early detection and management, which a machine learning model can provide.

DATA SET DESCRIPTION

The classification goal is to predict whether the patient has Cardiovascular disorders(CVDs) or not. The data set provides the patients' information. It includes 918 records and 11 attributes. Each attribute is a potential risk factor. There are both demographic and medical risk factors.

Attributes:

1. Demographic:
    a. Age: age of the patient [years]
    b. Sex: sex of the patient [M: Male, F: Female]
2. Information on medical records:
    a. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
    b. RestingBP: resting blood pressure [mm Hg]
    c. Cholesterol: serum cholesterol [mm/dl]
    d. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
    e. 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]
    f. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
    g. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
    h. Oldpeak: oldpeak = ST [Numeric value measured in depression]
    i. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

Target variable to predict:

Predicting Cardiovascular disorders (CVDs) — (binary: "1", means "There is a heart failure", "0" means "Normal")

EXPERIMENTATION

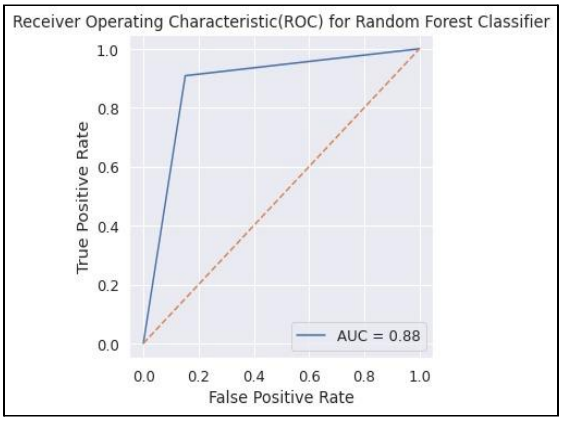
The full code is implemented in Python and different classification algorithms are used. Below is a brief description of the general approach employed:

1. Data cleaning and pre-processing: Here we checked and dealt with missing and duplicate variables from the data set as these can grossly affect the performance of different machine learning algorithms.
2. Exploratory Data Analysis: Here we wanted to gain important statistical insights from the data and the things that we checked for were the distributions of the different attributes, correlations of the attributes with each other, and the target variable and we calculated important odds and proportions for the categorical attributes.
3. Model development and comparison: We used five classification models, i.e., Random Forest Classifier, K-Neighbors, Decision Trees, XGboost Classifier, and LightGBM Classifier, After which we compared the performance of the models using their accuracy and F1 score.

Model Development And Comparison:

Using the training set, we trained the above-mentioned five classifiers, after training each model and tuning their hyper-parameters using random search, we evaluated and compared their performance using the Accuracy score, F1 Score, Precision Score, and ROC, as follows:

1. Random Forest Classifier



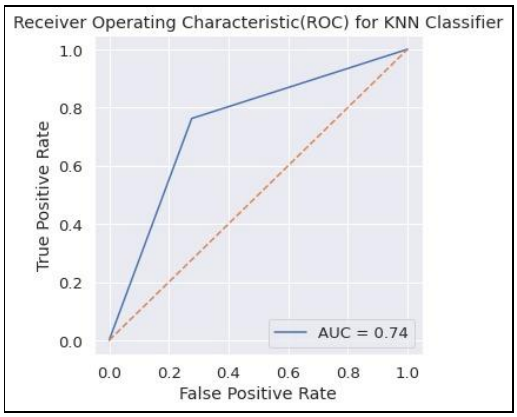
RandomForestClassifier

	precision	recall	f1-score	support
0	0.85	0.86	0.86	110
1	0.91	0.90	0.90	166
accuracy			0.88	276
macro avg	0.88	0.88	0.88	276
weighted avg	0.88	0.88	0.88	276

Hyper-parameters:

{'n_estimators': 275, 'min_samples_split': 5, 'min_samples_leaf': 4, 'max_features': 'log2', 'max_depth': 15, 'criterion': 'entropy'}

2. K-Neighbors Classifier



KNeighborsClassifier

	precision	recall	f1-score	support
0	0.72	0.68	0.70	120
1	0.76	0.80	0.78	156
accuracy			0.75	276
macro avg	0.74	0.74	0.74	276
weighted avg	0.75	0.75	0.75	276

Hyper-parameters:

{'weights' : 'distance', 'p': 1, 'n_neighbors' : 7, 'leaf_size' : 36}

3. Decision Trees Classifier

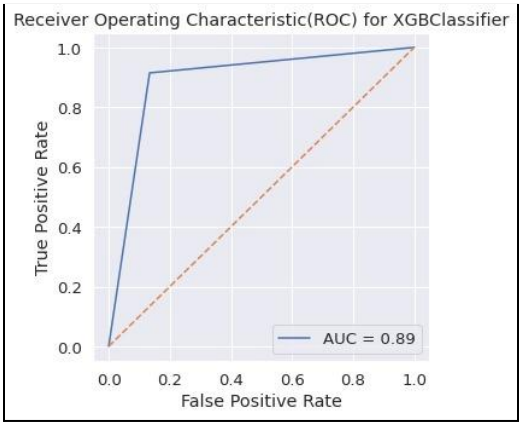
DecisionTreeClassifier

	precision	recall	f1-score	support
0	0.85	0.80	0.82	119
1	0.85	0.89	0.87	157
accuracy			0.85	276
macro avg	0.85	0.85	0.85	276
weighted avg	0.85	0.85	0.85	276

Hyper-parameters:

{'splitter': 'best', 'min_samples_split': 7, 'min_samples_leaf': 7, 'max_features': 'auto', 'max_depth': 7, 'criterion': 'entropy'}

4. XGboost Classifier



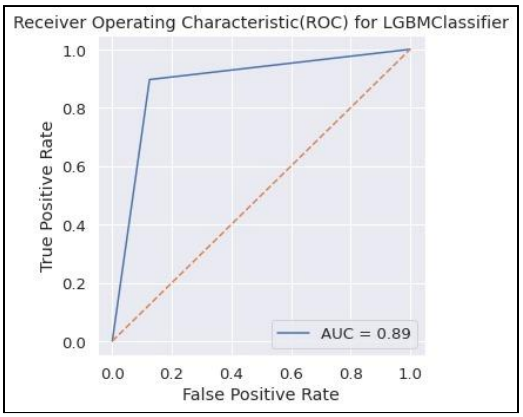
XGBClassifier

	precision	recall	f1-score	support
0	0.87	0.87	0.87	111
1	0.91	0.91	0.91	165
accuracy			0.89	276
macro avg	0.89	0.89	0.89	276
weighted avg	0.90	0.89	0.90	276

Hyper-parameters:

{'n_estimators': 39, 'max_depth': 3}

5. LightBGM Classifier



LGBMClassifier

	precision	recall	f1-score	support
0	0.88	0.85	0.86	115
1	0.90	0.91	0.90	161
accuracy			0.89	276
macro avg	0.89	0.88	0.88	276
weighted avg	0.89	0.89	0.89	276

Hyper-parameters:

{'num_leaves': 75, 'min_split_gain': 0.01, 'min_data_in_leaf': 11, 'max_depth': 4, 'learning_rate': 0.04, 'colsample_bytree': 0.4}

RESULT

The XGboost Classifier was the best performing model across all metrics. Its high Accuracy and F1 score also show that the model has a high true positive rate and is thus sensitive to predict if one has a CVD or not.

Team Members and Contribution:

1. Anirudh Bajaj (B20CS005)
    - Implementation of models, Random Forest, and LightBGM classifier
    - Pre-processing of dataset
    - Comparison of models by Accuracy and Precision
    - Commenting in colab file
    - Debugging
2. Harshita Gupta (B20CS018)
    - Implementation of model XGboost
    - Exploratory Data Analysis
    - Comparison of models by ROC
    - Report
    - Debugging
3. Khobragade Atul Yashwant (B20CS027)
    - Implementation of models, Decision Tree Classifier, and K-Neighbours
    - Visualization of dataset
    - Comparison of models by F1 score
    - Commenting in colab file
    - Debugging