Dialog Data Science Academy

Machine Learning Foundations Course - Capstone Project

MACHINE LEARNING MODEL DEVELOPMENT FOR HEART FAILURE PREDICTION

Prepared By: Crishmi Costa

Date of Submission: 21 - Nov – 2021

TABLE OF CONTENTS

1. INTRODUCTION

1.1BACKGROUND

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age.

The most important behavioral risk factors of heart disease and stroke are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioral risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. These “intermediate risks factors” can be measured in primary care facilities and indicate an increased risk of heart attack, stroke, heart failure and other complications.

Heart failure is a common event caused by CVDs and the dataset used in this project comprises 11 different features that can be used to predict the possibility of an individuals at a high cardiovascular risk. The presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established diseases make individuals more prone to CVDs.

Hence, the machine learning model developed in this project will be of great advantage to people with cardiovascular diseases or people at a high cardiovascular risk as this can be used to provide early detection and management of the disease.

1.2OBJECTIVES

The main objectives of this project are,

i. To build a variety of Classification models and compare the models giving the best prediction on Heart Disease.
ii. To use the best model and make a prediction on the target variable ‘Heart Disease’

2. DATA

This project aims to determine whether a person is likely to get a Heart Disease or not, based on several input parameters. Hence, the data set used for this project consists 11 different features that can be used to predict the same.

The data set consists of 12 different columns where 11 of them are considered as the input parameters which are in turn used to predict a single output parameter, which is the target variable.

The input parameters include the following,

i. **Age** : Age of the patient in years
ii. **Sex** : Sex of the patient | M: Male, F: Female
iii. **ChestPainType** : Chest Pain Type | TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic
iv. **RestingBP** : Resting Blood Pressure in mm Hg
v. **Cholesterol** : Serum Cholesterol in mm/dl
vi. **FastingBS** : Fasting Blood Sugar | 1: if FastingBS > 120 mg/dl, 0: if otherwise
vii. **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
viii. **MaxHR** : Maximum heart rate achieved | Numeric value between 60 and 202
ix. **ExerciseAngina** : Exercise-induced Angina | Y: Yes, N: No
x. **Oldpeak** : Old peak = ST (Numeric value measured in depression)
xi. **ST_Slope** : The slope of the peak exercise ST segment (Up: upsloping, Flat: flat, Down: downsloping)

The output parameter/target variable include the following,

i. **HeartDisease** : HeartDisease | 1: With Heart Disease, 0: Normal

The data set used in this model prediction was obtained from Kaggle.com and the link is as follows.

Link - https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

3. METHODOLOGY

The focus of this project is to determine the risk of Heart Disease to an individual. As the output of the prediction should be either 1 or 0, we consider this to be a binary classification problem where 1 would mean that the individual is at a risk of heart disease and 0 would mean that the individual is normal/healthy.

In this study, we analyzed a random dataset of 918 individuals where 508 of them were reportedly positive cases and 410 of them were healthy. During the analysis, we applied several machine learning classifiers including Logistic Regression, Decision Tree, Random Forest and KNN and examined various aspects to determine and identify the model that best predicts the target variable, 'Heart Disease'.

In order to achieve this, a several steps were followed.

3.1PRE-PROCESSING, EXPLORATORY DATA ANALYSIS & VISUALIZATION

If the data set that we use to develop our machine learning model has not been screened well, the results obtained could be quite misleading. Hence, the representation and quality of data is first and foremost before running any analysis. Similarly, we too have taken our data set through a several pre-processing stages to cleanse the data before training and modeling stages.

i. isna() function to identify cells with N/A errors.
ii. Descriptive statistics like count, mean, std, and a range of quartiles were checked to identify the distribution of the data set.
iii. The data set was also split into numerical and categorical variables and were further analyzed based upon that.
iv. Different methods of data visualization like box plots, pair plots and heat maps were used to analyze various features and the correlation among the variables.
v. We also checked how the output variable 'Heart Disease' varies with each individual input parameter.
vi. We also checked for any missing values in the data set

3.2MODEL DEVELOPMENT

Once the initial analysis of the data set is completed, we split the data for training and testing where the size of the test data set was 15% and the observations were selected at random.

```
[ ]  X = data.drop(["HeartDisease"], axis=1)
     y = data["HeartDisease"]
```

```
[ ]  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15, stratify = y, random_state = 101)

     print(F"Train sample size = {len(X_train)}")
     print(F"Test sample size  = {len(X_test)}")

     Train sample size = 780
     Test sample size  = 138
```

Feature scaling (Normalization) is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. For machine learning, in general, it is necessary to normalize features so that no features are arbitrarily large (centering) and all features are on the same scale (scaling). Hence, we have performed a feature scaling operation to our data set as well in order to rescale and standardize our data for modeling.

```
[ ]  from sklearn.preprocessing import MinMaxScaler
```

```
[ ]  scaler = MinMaxScaler()
     scaler

     MinMaxScaler()
```

```
[ ]  X_train_scaled = scaler.fit_transform(X_train)
```

```
[ ]  X_test_scaled = scaler.transform(X_test)
```

With that, various types of ML models were used to evaluate the data.

Types of Models used	Logistic Regression, Decision Tree, Random Forest, KNN
Independent variables	Age, Sex, Chest Pain Type, Resting BP, Cholesterol, Fasting BS, Resting ECG, Max HR, Exercise Angina, Old peak, ST_Slope
Dependent variables	Heart Disease

3.3COMPARISON OF MODELS

For the four model types that were used to analyze the data, the F1 score, accuracy, recall and the ROC values were used to evaluate and identify the model that provides the best prediction of the target variable, ‘Heart Disease’.

3.4MODEL SAVING & TESTING

Once the model was finalized, two methods of model saving using ‘Pickle’ and ‘Joblib’ were carried out.

4. RESULTS

The following observations were made with regards to the F1 score, accuracy, recall and ROC values of the four model types that were used.

	Model	F1_score	Recall	Accuracy	ROC_AUC
0	Logistic Regression	0.891720	0.921053	0.876812	0.871817
1	Decision Tree	0.866242	0.894737	0.847826	0.842530
2	Random Forest	0.906832	0.960526	0.891304	0.883489
3	KNN	0.857143	0.868421	0.840580	0.837436

5. CONCLUSION

As observed, it can be concluded that the best model to predict the risk of ‘Heart Disease’ of an individual would be the Random Forest Classifier.

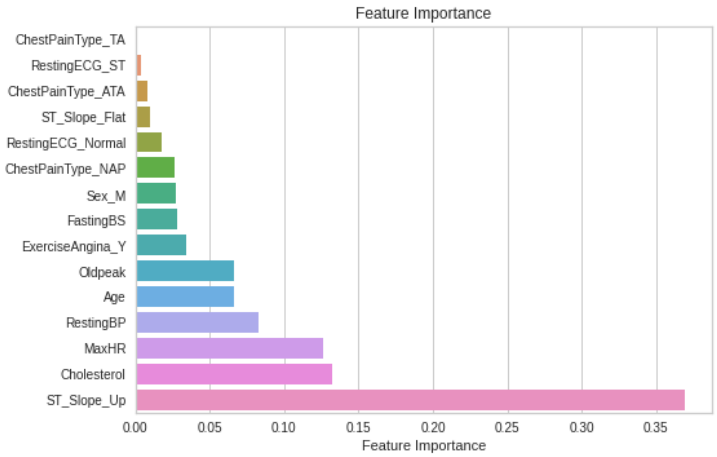
The model shows an approximate accuracy of about 89%.

This ML model has the potential to impact on clinical practice, becoming a new supporting tool for physicians when predicting if a heart failure patient will survive or not.

According to the observations made, the medical doctors aiming to understand if a patient will survive after heart failure may focus mainly on the Cholesterol level and the ST_Slope_Up.

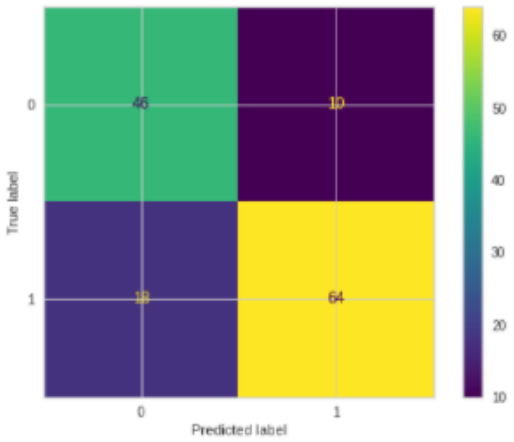
6. DISCUSSION

The data set we used in this analysis was already screened pretty well, and therefore we did not identify any columns or cells with null or missing values.



According to the above observations, it can be seen that the ST_Slope_Up and the Cholesterol levels are the main features that mainly contributes to increasing the risk of a 'Heart Disease' in an individual. If these features are not considered when developing our model, the accuracy of the model will vastly decrease (also has the true positive predictions & false negative predictions has vastly decreased).

	train_set	test_set
Accuracy	1.0	0.797101
Precision	1.0	0.864865
Recall	1.0	0.780488
f1	1.0	0.820513



Before modeling, we also checked if there is any effect of skewness in the end results and it was identified that there isn't and hence, we proceeded to the next stages without handling with skewness as it did not have much of an impact to the final results.

During the modeling process, it was observed that there wasn't much of a difference between the train and test results of F1 score, accuracy, recall, ROC values of the models.

And according to the final observations, it was seen that the Random Forest Classifier showed the highest values for F1 score, accuracy, recall, ROC.

7. REFERENCES

[1] https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2] https://www.kaggle.com/andrewmvd/heart-failure-clinical-data