5/6/2019

Heart Health Prediction

Data Science Project Report

Name: Syed Ammar Mahmood

Roll Number: 16K-3629

Section: GR1

Table of Contents

1. Abstract

Machine learning is an adaptive and continuous enhancement with respect to algorithms in which the growth is referred by prediction within traditional, common and manual approaches. These methods can be used in the health industry for healthier predictions of the diseases. The major reason for the death worldwide is the heart disease, which is a tough job for medical specialists to identify the heart attack without prior knowledge and understanding. This project aims to provide the acquaintance of the algorithms with better accuracy rate for predicting heart disease. Different machine learning algorithms such as: Random Forest, SVM (Support Vector Machine), Decision Tree and Logistics Regression have been implemented for the prediction of heart disease with Heart health dataset.

2. Introduction

The major reason for the death in worldwide is the heart disease in high and low developed countries. The data scientist uses distinctive machine learning techniques for modeling health diseases by using authentic dataset efficiently and accurately. The medical analysts are needy for the models or systems to predict the disease in patients before the strike. High cholesterol, unhealthy diet, harmful use of alcohol, high sugar levels, high blood pressure, and smoking are the main symptoms of chances of the heart attack in humans.

Data Science is an advanced and enhanced method for the analysis and encapsulation of useful information. The attributes and variable in the dataset discover an unknown and future state of the model using prediction in machine learning. Chest pain, blood pressure, cholesterol, blood sugar, family history of heart disease, obesity, and physical inactivity are the chances that influence the possibility of heart diseases. This project emphasizes to evaluate different algorithms for the diagnosis of heart disease with better accuracies by using the patient's dataset because predictions and descriptions are fundamental objectives of machine learning. Each procedure has unique perspective for the modeling objectives. Algorithms have been implemented for the prediction of heart disease with our Heart patient dataset.

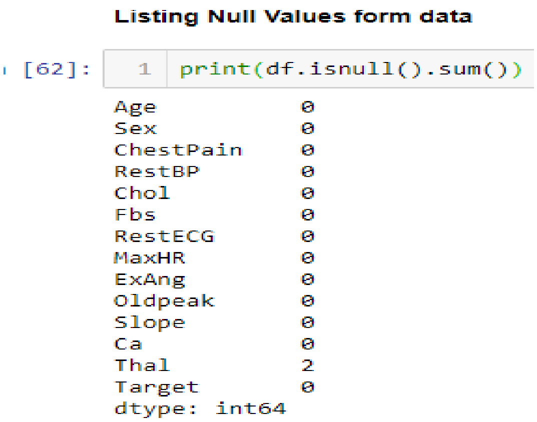
3. Dataset Understanding Details:

3.1. Dataset Information:

The dataset consists of 303 rows and 14 columns with label Target. Data is categorical as well as continuous.

3.2. Data Cleansing:

There is only one column that contains null value in which only two rows have null value that can easily be dropped by removing NA function. The data have two objects that are Chest Pain and thalassemia that are categorical data we have to covert this data into numeric type by using dummies values and Slope and Rest ECG have numerical data but data is divided in category by digits.



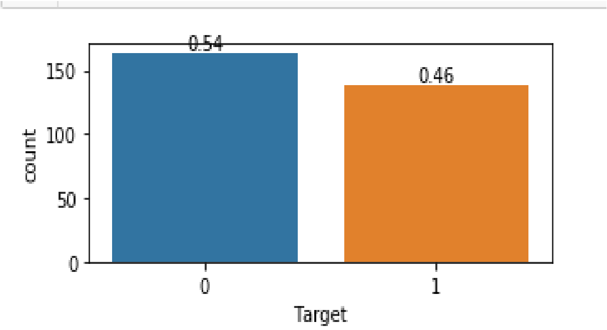
Listing Null Values form data

```
[62]:   1  print(df.isnull().sum())
```

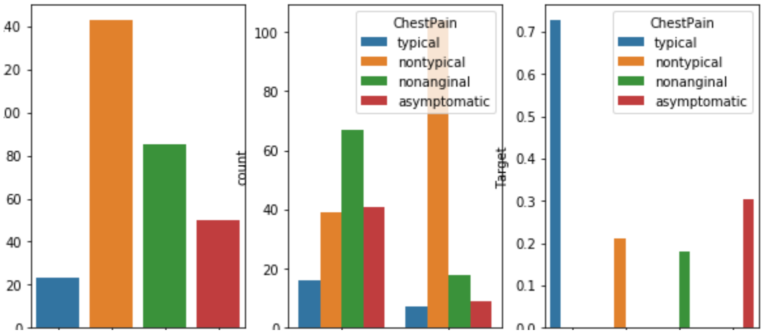
```
Age          0
Sex          0
ChestPain    0
RestBP       0
Chol         0
Fbs          0
RestECG      0
MaxHR        0
ExAng        0
Oldpeak      0
Slope        0
Ca           0
Thal         2
Target       0
dtype: int64
```

3.3. Data Visualization:

Data Visualization is done by step by step process with critical analysis I use correlation matrix to find most dependent variable to the label which is Age. I plot graph of label (Target) to show the ratio of heart Disease



I plot the graph for Categorical Data for Chest pain and Thalassemia and for Continuous data

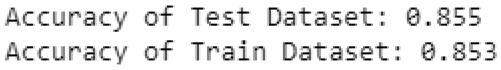


4. Methodology

To explain and identify the problem and resolve medical objectives, different data Science technique, which interpret the medical goals, have been implemented to diagnose the heart disease and to improve the success standards of the algorithms for prediction. Suitable machine learning algorithms, like: Random Forest, SVM (Support Vector Machine), Decision Tree and Logistics Regression were preferred for the training and implementation in python for developing and evolving the predictive model. These algorithms executed on the model will help medical experts to predict and diagnose heart attacks in the patient dataset. The main goal is to identify which machine-learning algorithm has the best accuracy for the prediction of heart disease from the patient dataset.

4.1. Logistics Regression:

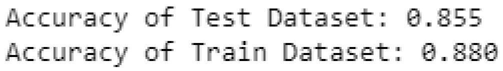
Logistic regression technique is for the prediction of the results of the categorical as well as numeric data. It is widely used in the medical industry for the prediction models. They are based on dependent and independent variables where dependent variable is binary. The probability of success and the estimated prediction can be evaluated through logistic regression that is widely used. For the prediction of heart disease on the data set, logistic regression function has used built-in in python language and implements the functions for the better accuracy. By fitting the data to a logistic curve, the probability of the heart diseases that occur or not have been predicted by using patient dataset.



```
Accuracy of Test Dataset: 0.855
Accuracy of Train Dataset: 0.853
```

4.2. Support Vector Machine

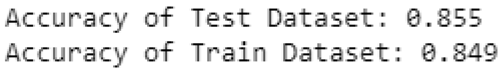
For prediction SVM is one of the appropriate and most used algorithms for the better and desired results. For fundamental risk reduction, Support vector machine gives the estimated implementation. The method requires decent and suitable generalized functions. It works when the dataset has two classes to predict the better results and it finds the hyper plane that splits the data points between classes. SVM is a mathematical model to resolve real world tough problems. They use kernel spaces like: linear quadratic and polynomial for training the data. This method describes the heart disease problem in effective manner and can be more enhanced in the future for the better and perfect results.



```
Accuracy of Test Dataset: 0.855
Accuracy of Train Dataset: 0.880
```

4.3. Decision Tree

For the prediction of data decision tree plays a major role that uses classification and regression. When the data is continuous we prefer regression but when we have a grouped data we use classification. The decision tree is the important and leading machine learning technique. We use a path from root node to leaf node for the evaluation of the data. One of the decision tree algorithm has been used in this paper for the estimation and prediction of the heart disease, which gives the classifiers as a tree is used to construct smaller trees with a fast speed and better accuracies than the algorithms used before in the decision trees. Constructing a tree and operating it on the dataset are the steps we have used to predict the possibilities of heart attacks.

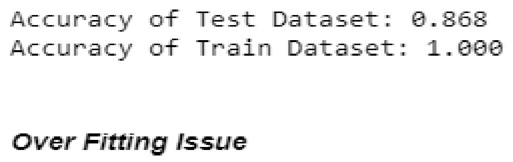


```
Accuracy of Test Dataset: 0.855
Accuracy of Train Dataset: 0.849
```

4.4. Random Forest:

A technique which is used to create predictions and gathers their result is known as Random forest (RF). It splits on the basis of selection of input variables that are trained on the original training data. CART is a binary decision tree, which is used for splitting data from the root node to the leave nodes.

It manages the variables well that are used for prediction. This method has the great performance and better impact on the dataset used in this paper and it is one of the well-used algorithms for the prediction of heart disease. The random forest technique consists of some constraints, like: Base Classifier, Split Measure, Number of Passes, Combine Strategy and Number of attributes used for base prediction.



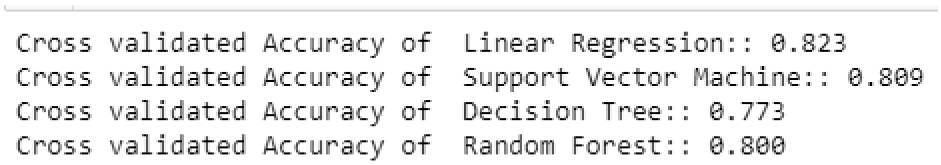
```
Accuracy of Test Dataset: 0.868
Accuracy of Train Dataset: 1.000
```

Over Fitting Issue

5. Results

5.1. Cross Validation:

Cross Validation is also done for all the models. The results are same but have some variance in accuracy. After Cross Validation the result become clear that Logistic regression is good for this case



```
Cross validated Accuracy of  Linear Regression:: 0.823
Cross validated Accuracy of  Support Vector Machine:: 0.809
Cross validated Accuracy of  Decision Tree:: 0.773
Cross validated Accuracy of  Random Forest:: 0.800
```

5.2. Model Selection

Hence, it is concluded that Logistics Regression is the best suitable model for this problem, although more accuracy can be achieved in future.

6. Conclusion

The aim of this project is to predict the heart disease in patient dataset using the four different machine algorithms, i.e. Random Forest, Support Vector Machine (SVM), Decision Tree and Logistic Regression with higher and estimated accuracy. Moreover, computer world has improved and enhanced the medical industry to a leading level. The dataset used in this project has been divided in the training and testing dataset for the better results. In today’s world medical practitioners are in need for the models that can solve complicated problems and can predict and estimate the better accuracy and results. Machine learning algorithms used in this project are more efficient and have great impact on the systems that predict the heart diseases in the patients. Later for the more proficient and effective systems of heart disease prediction, this proposed work could be further improved, enhanced and extended.

7. References

https://www.kaggle.com/zhaoyingzhu/heartcsv