2020

Assignment 2 WHI

ANALYTICS
DIVINE OKEY-IKERI

LAKEHEAD U |

Problem definition & Objectives

The purpose of this assignment was to work on any supervised analytics specially one on classification on any publicly available dataset. On the dataset we are meant to apply decision tree and random forests with fold cross validation to calculate performance measures such as accuracy, sensitivity etc. Then employ one feature selection method to identify important features and re-run the classification experiments with selected features using any analytics/ library. For this assignment I chose to do it on python. The significance was to help us understand the analytics techniques we were taught in class and help us implement them.

Data

I got my data from the UCI Machine Learning repo the link is given below. The name of the data set was processed.clevaland.data under heart disease

https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/

https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

Description

This data set was preprocessed. It has 13 attributes all together. Before it was preprocessed it had 8 symbolic values and 6 numeric values. After preprocessing all data types where float64 except Ca and thal which were objects. The last attribute HD (heart disease) was an int64.

Methods & Tools

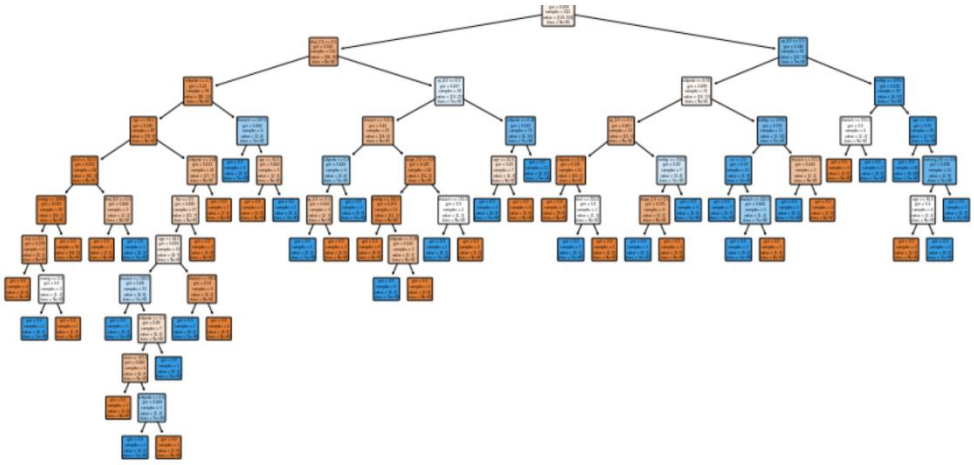
Preprocessing steps.

Even though this data set was preprocessed I still had to rename the columns to make the dataset usable. Then I had to check for the data type of each column. In which 2 of them were objects which suggested to me that something was off. So, I had to search the CA and thal column to see if there were missing values in which I found a combined total of 6 missing values. As this was only 1.98% of the total data set it would not have made a difference. So, I decided to remove them form the dataset taking the total number from 303 to 297. After this I decided to make and X and Y graph in which the X axis would comprise of attributes which would predict this Y axis which was the HD attributes. Then since its not possible to work on categorical values using SKlearn package. I had to hot encode the attributes which had categorical values. These attributes were the cp, restecg, slope, thal attributes.

Analytics, Validation, Tools.

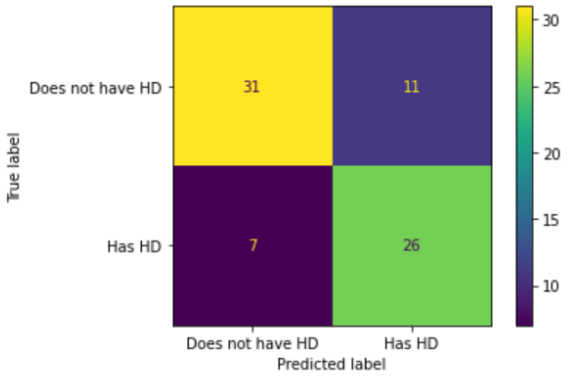
The first part of the analytics method was making a simple decision tree after the data set was ready. For this step I had to make the value of the hd attributes read between 0 and 1. Then I created a decision tree. After creating the decision tree I made a confusion matric to analyze how to the decision tree performs. Once that was complete I had to prune the model to make it more accurate using cross validation (5 fold) to make it as optimal as possible. After that I stored the mean and SD of the scores for each call. Then I Drew a graph for the mean and STD in order to find the new alpha value. Since pythons thinks it's a float value. We need to convert the values to a float64. After getting this value I build a new decision true using the optimal value. Then I drew another confusion matrix to compare it to the iniatial one and was more accurate. Then I plotted a new graph which was significantly smaller than the first one.

Results



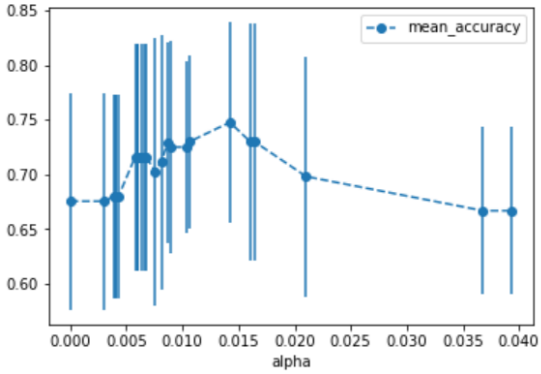
DECISION TREE

```
Out[39]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x179bcb3e040>
```



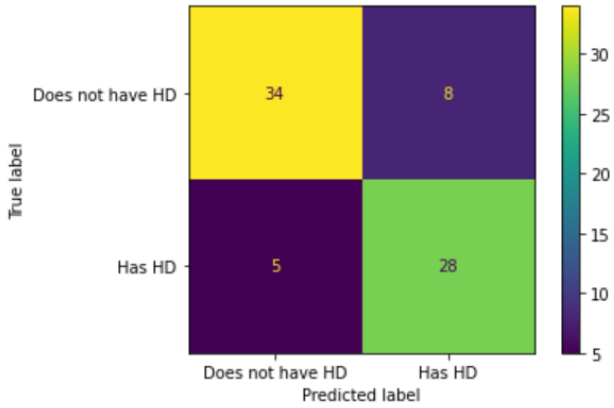
CONFUSION MATRIX 1

Out[63]: <AxesSubplot:xlabel='alpha'>

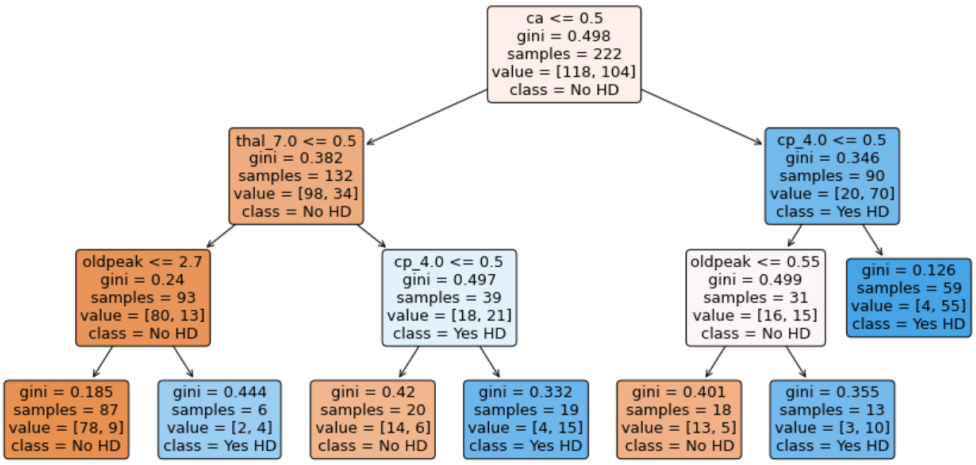


GRAPH TO SHOW MEAN AND STD ACCURACY PREDICTING ALPHA

ut[75]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x179bcf4d2e0>



CONFUSION MATRIX AFTER GETTING ALPHA



FINAL DECISION TREE

DISCUSSION

Basically, for this assignment I was able to preprocesses the data set so that it could be in a condition were and I can do analytics on it. For the analysis and used decision tree and 5 fold cross validation to make the initial model more accurate.

So, After the fist decision the and confusion matrix the we can see that 42 people didn't have heart disease and only 31 were properly classified. We can also that 33 people have heart disease and only 26 were properly predicted. This suggests that there is an overfit somewhere in the training set. This when I decided to prune using the fivefold cross validation by find out the mean and standard deviation. Then I test for alpha using the Training and testing data sets in which I got 0.014. Then I set alpha to 0.14 to build the final decision tree. Which showed a more accurate model as 34 people who didn't have heart disease were properly classified and 28 people with heart disease were properly classified.

The result shoe that the cross validation makes the model more accurate. However, the decision tree method give a lot of room for the model to overfit and even thought the accuracy for heart disease went up after pruning it was still accurate enough to capture most cases.