AI Induvidual Assignment

MugangaCharles A96447

February 2023

1 Question

You have been provided with a data set called "Golf Gaming.csv" which has information on whether golf will be played at the Namulonge Golf Course depending on various weather parameters.

2 Approach

i. Importing the libraries and modules.
ii. Reading the data set.
iii. Exploring the data set

3 Dealing with the data

Through the process of describing the data set, our variables have unique repeating values and since our data set is full of categorical data, I attached labels to the data. Labels were encoded automatically using 'sklearn', numerical values starting from zero(0).
Then Two columns 'Play Golf' and 'Day' are Dropped. We are dropping 'Play Golf' because it is what we are predicting and 'Day' because I am not going to use it when training my model.

4 Data Cleaning

To be on the safe side, I removed any missing values if any and the value with a question mark(?), and also removed the outliers based on the labels set that is to say any field contains a label greater than two.

Then separating the x input and y target variable.
Due to the number of x inputs in our data set, various testing points are made that is to say, for my approach, I considered xdata1,xdata2,xdata3 and xdata4.

- ytarget variabe still 'play Golf'

- xdata1 containing two variables ’Outlook and Temperature’
- xdata2 containing two variables ’Humidity and Wind’
- xdata3 containing two variables ’Outlook and Humudity’
- xdata4 containing two variables ’Temperature and Wind’

5 Building the Model

From a variety of modules to build the models, I used Logistic regression.
- First by splitting the data into train and test data.

The training size is 80 and the test size is 20.
Results of splitting are as follows;
Xtrain (13, 2)
Xtest (4, 2)
Ytrain (13,)
Ytest (4,)

- Traing the model using the first set of input points

The score is 1.0
The prediction is ’Yes’
- Training the model using the second set of input points

The score is 0.5
The prediction is ’Yes’
- Training the model using the third set of input points

The score is 0.75
The prediction is ’Yes’
- Training the model using the fourth set of input points

The score is 0.75
The prediction is ’Yes’
- Training the model using the data set

The prediction is ’Yes’
The trained models are saved as ”.joblib”.

6 Conclusion

From the predictions that are provided at different testing points, I can conclude that golf can be played on D17.
Then I replaced ’Yes’ with ’?’ given in the data set.
Also considering the performance/score of some given variables, there is a need to check and see how the different variables correlate with one another.

7 Alternatives

I tried out some alternative approaches.
- Using the Decision Tree approach.

- Using the Pipeline Approach.