Ashiq Imran

Student ID: 1001454594

CSE 5335

Programming assignment

Data Mining

Model 1

For 75%/25% data division,

a) I have used Random Forest Classifier. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. I have used Random Forest Classifier with the following feature columns from Wine datasets:

From the features of the datasets,

['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol']

I have dropped following features after calculating feature importance.

[‘residual sugar', 'total sulfur dioxide', 'citric acid', 'fixed acidity’]

Now, my dataset contains following features

['volatile acidity', 'chlorides', 'free sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol']

b) Accuracy is printed in the code

c) Confusion Matrix is printed in the code

Model 2

For cross validation, I used the following approach.

I have used the following feature columns:

['volatile acidity', 'chlorides', 'free sulfur dioxide','density', 'pH', 'sulphates', 'alcohol']

I have decided to drop [‘residual sugar','total sulfur dioxide','citric acid','fixed acidity’]

a) I have used Random Forest Classifier for cross validation. I have used 10-fold stratified cross-validation on a GridSearchCV using Standard Scaling on a pipeline.

b) Accuracy for each fold is printed in the code. Lastly, average accuracy is printed.