INFO

Libraries:

The external libararies used in Assignment 1 include scikit-learn, nltk.

For each method used in the code, comments have been put which depict what the function used does.

Here are some of the important methods used:

1) DecisionTreeClassifier():
   It is the implementation of the decision tree classifier in scikit-learn module.
   The default criterion used is "Gini" and it automatically splits taking the best attribute based on Gini index.
2) word_tokenize():
   It's a part of nltk which tokenizes the line of strings into words.
3) Predict():
   This is used for testing the classifier we have developed.
4) metrics.accuracy_score(predicted, groundtruth):
   This gives us the Accuracy of the testing dataset that predicts using a particular classifier.

Challenging Parts:

The challenging part of the assignment is the preprocessing of the String data in Amazon Review dataset.

1) Initially, I have used the LabelEncoder() of scikit-learn which encodes the string into a particular label.
   Using this, the accuracy I obtained was 39% on the Testing dataset and 99% on the training dataset.
   So, this approach was wrong. So we have changed the algorithm to use NLTK.
   Along with my partner, we have learned various methods in NLTK, which we have implemented and used a sample code reference to form our algorithm.

Easy Parts:

The easy part is Digit Recognition dataset. This was one of the easy part used to predict the labels.