Part 1 : Digit Recognition

a) Status and stopping point-

In the first dataset for optical recognition of digits, the maximum accuracy percentage is around ~85% which can be considered as an acceptable value. Decision Tree Classifier is used in this case with and without pruning. While pruning the Decision tree to a certain depth, the accuracy decreases drastically. For instance when the tree is pruned till the depth of '9', the accuracy is pretty good. However, when the tree is pruned till the depth of '5', the accuracy drops. This could be the stopping point in perception of the depth of the tree.

b) Additional functions and analysis:

Decision tree can be considered as a proper classifier to classify this type of dataset. The code seemed quite efficient in runtime perspective as the data size seemed quite fine. The time taken by the data for loading and manipulation is not very high. So, that is the good part. Few experiments and analysis are done which leads to the following inferences:

- Pruning of the Decision tree leads the results to vary.
- When the maximum depth of the tree is taken as 20 and more then the graph of the correct predictions gets into a monotonous form i.e. the accuracy reaches till a saturation point.
- On a similar note if the maximum depth of the tree is taken as 9 or less the accuracy started dropping from 79% and low respectively.

c) The most challenging part is the data formation. Understanding and Importing the data and the proper use of libraries took time. So code must be efficient to run over those data. That is what I learnt; to handle data and manipulating the data using efficient code methodologies.

Part 2: Amazon Data Set

a) Status and stopping point -

With Amazon Data Set, the maximum accuracy percentage that I get is around ~60%. I tried using Decision tree classifier with and without pruning. Without pruning the accuracy is around ~54% but with pruning to depth 25 it bumps up to ~60%. I tried tackling same problem by dividing ratings in two categories namely good and bad rating. But the accuracy that I got was ~60% which really had no big impact on my result. Also I give it try by using ForestTreeClassifier which gave me accuracy of ~54%. Hence I settled for Decision tree with pruning. If the depth is 25 and below and I getting maximum accuracy and above 25 it is less.

b) Additional functions and analysis -

After analysing data I realized that it's necessary to clean the data as data set had rows with no reviews. To get rid of those rows I used dropna() method on data frame.

Then review itself had some html markup words. To counter that I used Beautiful soup. And then in order to get rid of stop words such as ‘is’,’and’ I used nltk package.
If I don’t perform all these operations the accuracy of the code is lower than 54%.

c) Challenges -

Amazon data set is really huge and playing with such huge data for the first time was challenging for me. It took me while to understand that there are certain scenarios which might affect the output such as review column being blank. Then trying to get rid of those columns. Also know why and how to clean data was important as well. It took me some time to understand that as well. The term Bag of Words was new to me as well.