Burrell, David. COMP307
300209541 Assignment One Report

Part One: Nearest Neighbour Method.

1.
```
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-setosa	Actual: Iris-setosa
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-virginica		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-versicolor		Actual: Iris-versicolor
Class: Iris-virginica		Actual: Iris-virginica
Class: Iris-versicolor		Actual: Iris-virginica
Class: Iris-versicolor		Actual: Iris-virginica
Class: Iris-virginica		Actual: Iris-virginica
Class: Iris-virginica		Actual: Iris-virginica
```

```
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-versicolor    Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-versicolor    Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-versicolor    Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Class: Iris-virginica     Actual: Iris-virginica
Accuracy: 92.0%
```

2. Accuracy: 94.66666666666667%
   K = 1 Isn't quite accurate enough, with K = 3 there are at least some comparisons being made and will definitely have a better result.

3. The main disadvantage of the K-nearest neighbour method is that you do have to store the entire training set for classification of the test set, needing to look at the whole training set for every test case is not ideal for performance.
   An advantage to this is that the implementation is not particularly difficult.

4. Before the Classify() method is run I would set up a Validate() method in which the I would take the training set and using a loop make k = 5 subsets as the new testset and run them through either a new classify(ArrayList<Iris>) method or edit the classify() method to be able to deal with the validation sets.

5. I would have to implement the K-Means clustering algorithm instead, using random means and the Euclidean distance measures to group the data.

Part Two: Decision Tree Learning Method.

1. My classifier: 23/27 correctly classified.
   Baseline classifier: 23/27 correctly classified.
   Something may be going wrong with my classifier as there is no difference between mine and the baseline, and I would expect there to be at least a minor difference but not exactly the same.
2. Using the provided split data files rather than splitting my own with the script.
   'run01': 33/37
   'run02': 32/37
   'run03': 29/37
   'run04': 29/37
   'run05': 32/37
   'run06': 29/37

‘run07’: 33/37
‘run08’: 32/37
‘run09’: 30/37
‘run10’: 30/37
Average accuracy = 31/37

3. a) to prune leaves from a DT you would look at pairs of child nodes and calculate the increase of impurity that would occur if you removed them both, and if it is sufficiently small increase then you remove the child nodes and make the parent the leaf.
   b) this would reduce overfitting of the training set making test set accuracy potentially smaller.

4. Because the impurity measure is a function that fits between 0 and 1, it would be difficult to calculate a way for it to split the function between 3 or more classes with any viable accuracy.

Part Three: Perceptron.

1. I know something is clearly wrong with my perceptron as it seems to be 100% accurate going at the start of the first epoch. Forcing it to complete 100 epochs will occasionally drop the accuracy down by 1 data point for one single image object. I would like to say that my perceptron does find a correct set of weights, however I doubt that is true purely because it shouldn’t be perfect before doing any calculations.
   I don’t really know where it goes wrong. If somebody identifies the problem I would appreciate the info.

2. It is not a good idea because of the usual overfitting problems, I would want to use a validation set to validate at set intervals or before moving on to testing. Ideally a few validations before moving on.