**Machine Learning: Assignment 2 – Build A Classifier (Jennifer Nolan C16517636)**

The data used to build this classifier is related to a marketing campaign for a banking institution based on previous phone calls made to its customers. There are 24,300 rows available in the dataset. The goal of the classifier is to predict if a customer will subscribe to a term deposit or not after being contacted.

The dataset is split into continuous, or numeric, features and categorical features. The target feature of this dataset is whether the customer will subscribe to a term deposit or not. This value is a binary value and therefore categorical. In total there were seven continuous features and nine categorical features. The continuous features consist of id, age, balance (the customers yearly balance), day (the day the customer was last contacted), campaign (the number of times the customer was contacted during the campaign), pdays (the number of days since the customer was last contacted) and previous (the number of times the customer has been contacted prior to this campaign). The categorical features consist of job, martial, education, default (whether the customer has credit in default), housing (whether the customer has a housing loan), loan, contact (how the customer prefers to be contacted), month and poutcome (what was the outcome from previous campaigns with this customer).

To build the classifier the information-based learning decision tree model was used. This classifier was chosen because it predicts a target feature using the descriptive features the model has been trained on. The decision tree model creates the shallowest tree possible using the given dataset. This model builds trees in a recursive, depth-first way which begins at the root node (the starting point) and works down the tree towards the leaf nodes (the finishing points).

The decision tree classifier consists of a root, or starting node, an interior and leaf, or finishing node. A decision tree begins with choosing the best descriptive feature, using information gain, to test with. The root node is then added to the tree and labelled. From there the dataset is partitioned into testing and training datasets. This same process is followed for the other branches of the tree. There are three situations in which recursion can stop and a leaf node is made. Firstly, if the instances in the dataset all have the same classification. In this case return a leaf node with the classification label. Secondly, if the set of features is empty then a leaf node is returned. And lastly, if the dataset is an empty dataset then the decision tree is also empty.

The decision tree classifier was chosen for this dataset for the following reasons. This classifier works the same way on more complicated datasets. There are a lot of extensions and variations available with a decision tree, for example impurity measures and pruning to avoid overfitting. There are also many advantages to the decision tree model. Firstly, the decision tree mode handles both categorical and continuous descriptive features, both of which are in the dataset being used. This allows for the modelling of the interactions between features. As well as that, the decision tree model is strong against noise, when pruning is used, and the curse of dimensionality.

However, like most models the decision tree model does have its disadvantages. Decision trees can become larger with the more continuous features used. They are also prone to overfitting if there are a lot of features available in the dataset. Also, decision trees are considered an eager learner which means they are prone to concept drift. However, this model was chosen as, in my opinion, the advantages outweigh the disadvantages in this case with this dataset.

To test the accuracy of the model the following was done. Firstly, the classifier was built using the entropy criterion and a second with the Gini criterion. The dataset was then split between training data and testing data. The classifier was then trained using the split training data, followed by predictions made using the test instances. Once the predictions were made for each model the accuracy of these predictions was determined. These predictions were as follows: Entropy – 0.87, Gini – 0.88. After comparing these accuracies, it was found that the Gini criterion for the decision tree was the most accurate and was used to make predictions on the query dataset.