DV2542 – Machine Learning
Assignment 2 – Decision Tree

Chaitanya Malladi
P No – 930808-0150
chma15@student.bth.se

1. Aim:

The aim of the assignment is to implement a tree model that generates a decision tree for a given ARFF data file.

2. Assumptions:

The following assumptions were made in this implementation:

- The ARFF file inputted follows that standards of making the last attribute as class attribute.
- The ARFF file contains only nominal attributes.
- The missing valued instances are ignored to build the tree. It is assumed that they do not affect the tree.

3. Implementation:

- Implementation is based on Algorithm 5.1 and 5.2 from "Machine Learning – The Art and Science of Algorithms that Make Sense of Data" by Peter Flach
- The implementation is done on Java 1.8 in the Eclipse IDE for Windows
- The Java library of Weka 3.6 is used
- A Tree class is written to support the features of a tree structure for the decision tree
- The class DecisionTree extends the class Classifier from Weka library. In this class, the functions buildClassifier and classifyInstance are inherited and implemented to enable the use if Evaluation features
- The functions growTree, bestSplit from algorithm 5.1 and 5.2 including their supporting functions of homogeneous, label, and impurity are also implemented in DecisionTree class
- Impurity is calculated during identification of bestSplit attribute. The Gini Index is taken for calculation of impurity. This is given by:
    - Gini Index: $2p(1-p)$
      where p is empirical probability
- The 'main' function is part of the DecisionTree class. This function initializes the data from given ARFF file, builds decision tree, prints it to the console, uses Evaluation class to perform evaluation and print the results to the console

4. Evaluation:

- Evaluation is done using the functionality provided in the Weka's Evaluation class
- 10-fold cross validation is performed on the ARFF data
- The results of this evaluation are printed to the console

5. Results:

Input: ARFF file with nominal attributes

Output: The corresponding decision tree and the results of 10-fold cross-validation on the tree generated

The sample data file of "weather.nominal.arff" was input to get the following output:

```
Decision Tree
=============
outlook
      :sunny
            humidity
                  :high
                        1.0
                  :normal
                        0.0
      :overcast
            0.0
      :rainy
            windy
                  :TRUE
                        1.0
                  :FALSE
                        0.0


Evaluation
==========

Correctly Classified Instances          12               85.7143 %
Incorrectly Classified Instances         2               14.2857 %
Kappa statistic                          0.6889
Mean absolute error                      0.1429
Root mean squared error                  0.378 
Relative absolute error                 30      %
Root relative squared error             76.6097 %
Total Number of Instances               14     
```