MACHINE LEARNING (DV 2542)
ASSIGNMENT - 2

K.S.SUSHEEL SAGAR
susheelsagar4@gmail.com
P.No. 9303177837

1. AIM:

The aim of this assignment is to build a decision tree model generator, that generates a decision tree for an inputted ARFF classification dataset. The model generated must be printed to the console and a predictive accuracy obtained by the 10-fold cross validation of the model must also be obtained.

2. ASSUMPTIONS:

- As per the standard ARFF format, an assumption is made that the last attribute in the inputted ARFF file is the class Attribute.
- The implementation model assumes that the input ARFF file contains the NOMINAL Attributes ONLY.
- The implemented model does of handle the missing values.

3. IMPLEMENTATION:

i) Tools : Open Java Development Kit (Open JDK 1.7), Weka 3.7

ii) Software Details :

- The Model is build using the Java Integrated Development Environment “Eclipse - Mars”.
- The entire source code is written in JAVA language only.
- Weka 3.7 software package is used to evaluate our model.

iii) Algorithmic Details :

- The decision Tree model generator is build totally based on the algorithms 5.1 and 5.2, presented in the course textbook - “ Machine Learning - The Art and Science of Algorithms that Make Sense of Data” by Peter Flach.

- The algorithm is designed to handle the binary classification problems effectively.
- Initially the dataset is loaded, using the Instances class in weka.
- The BuildClassifier function in Weka's "Classifier" class was extended to implement the decision tree model. The motivation for this choice is that, weka allows the extension of the Classification algorithms through this class[weka site].
- All the data handling is done using the weka's ARFF format handling classes such as Instances, Attribute,etc.,.
- Weka's "Evaluation.class" file is used to perform a 10-fold cross validation of our model.
- The function crossValidate(Classifier , Instances , number of folds, Random) performs the cross validation of the implemented model.
- The Classifier object in the crossValidate function is the object of our implemented decision tree class.
- The Instances corresponds to the instances of the dataset to be given as the input for training the model.
- The number of folds is set to 10 in-order to perform the 10-fold cross validation.
- The random function is the function that randomly selects the instances for the 10-fold cross validation.
- **Impurity Function : <u>Information Entropy</u>** method is used to calculate the impurity of a data set.
- Based on the information entropy method the impurity of a set of instances is calculated using the formula :

  $$- p\log(p) - (1-p)\log(1-p)$$
- In the above formula 'p' is the empirical probability of the set of instances.

4. Results :

The decision tree model is tested using two datasets provided by weka 3-7-13. The Predictive accuracy obtained on each of the data sets is presented below in Table-I.

Dataset	**Average Predictive Accuracy**
weather.nominal.arff	85.7143 %
mushroom.arff (without the attribute stalk-root)	99.1137 %

TABLE - I