**Report on Data Mining Assignment**

*By Radeeb Bashir and JunHyeong Lee*

**Notes about testing the program:**

Our program is tested with two different datasets**:**

**training\_data1** takes in the form of pandas dataFrame that is extracted from .csv file called ‘iris.csv’. The labels are manually typed in. Program can be tested with a different csv file for training by changing line 8.

**training\_data2** mimics the example in homework instruction and program is tested for correct classification.

Program uses two different functions for creating data table according to the format of training data provided: one for reading from csv files and the other from the list provided inside the program.

**Program Summary/ Report for Real Data Set:**

For the real data set we have extracted the very popular **Iris** data set from the UC Irvine Machine Learning Repository. As mentioned earlier, we have used pandas library for the manipulating the csv file into a DataFrame.

The Iris data set has three class and the goal of the program is to accurately predict one of the three subtypes of the iris flower. The training data consists of four different physical features namely sepal length in cm, sepal width in cm, petal length in cm, and petal width in cm. Our algorithm performs just like it performs on the smaller data set. It calculates the entropy of every attribute, splits the data set into subsets where the information gain is maximum. It makes a ‘decision tree node’ containing that attribute. Then the program runs recursively on subsets using remaining attributes.