Lab Exercise 3

Introduction to Data Classification - Decision Trees

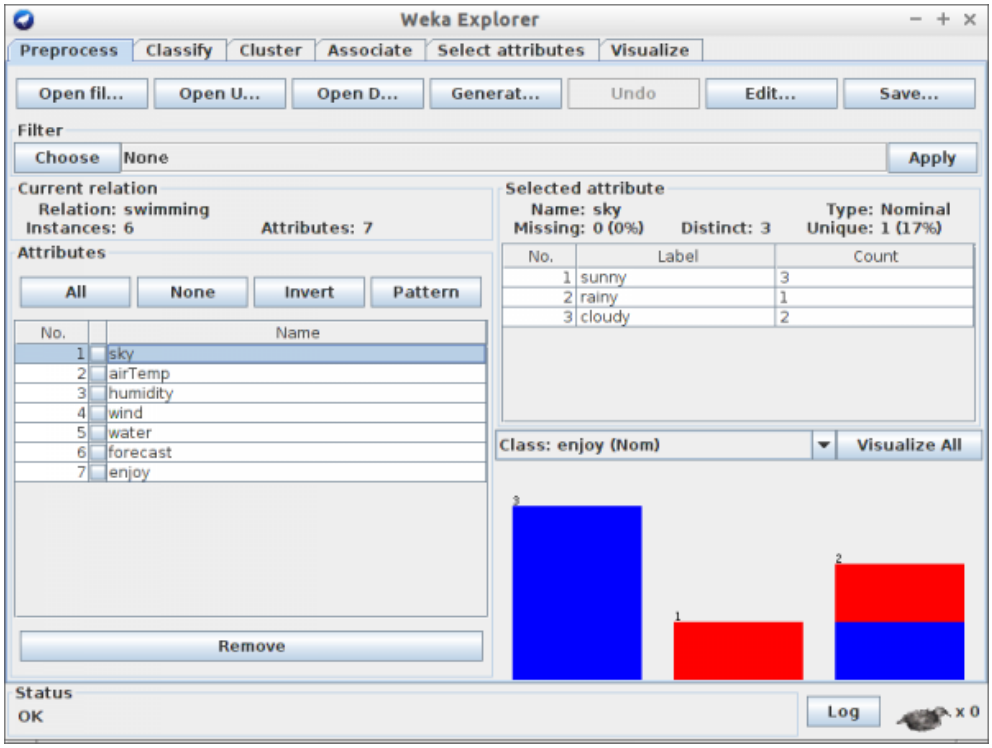
A decision tree is a graphical method of supporting the decision-making process, used in decision theory. The decision tree algorithm is also used in machine learning to generate knowledge based on given examples.

The aim of this laboratory is to use the WEKA package to generate a decision tree (decision table).

DataSets: Link : https://github.com/caiomsouza/ml-open-datasets/tree/master/weka-dataset-arff

Data loading and analysis

1. Open any dataset and learn thestructure of this learning file with its symbolic data vectors.
2. Start Weka, click the Explorer button and load data file.
3. Analyse the first '*Preprocess'* tab and answer the questions below:

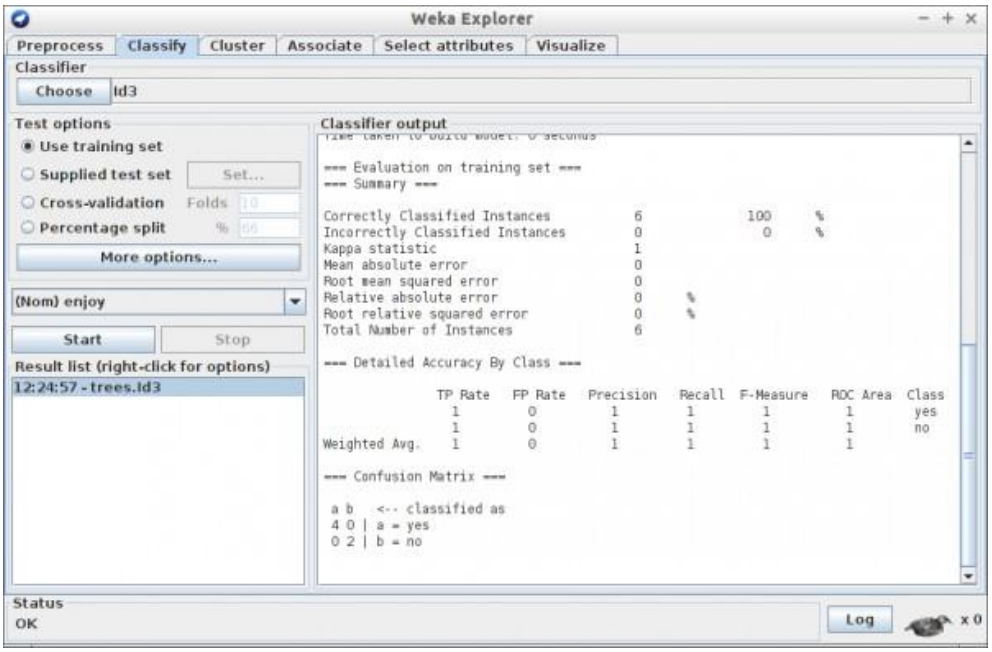


a) What is the size of the training set?
b) How many attributes exist in the training set?
c) How many instances are positive (Enjoy = yes) and how many negative?
d) Which attribute best separates the data?
e) How many elements from the data set have the humidity attribute set as high?

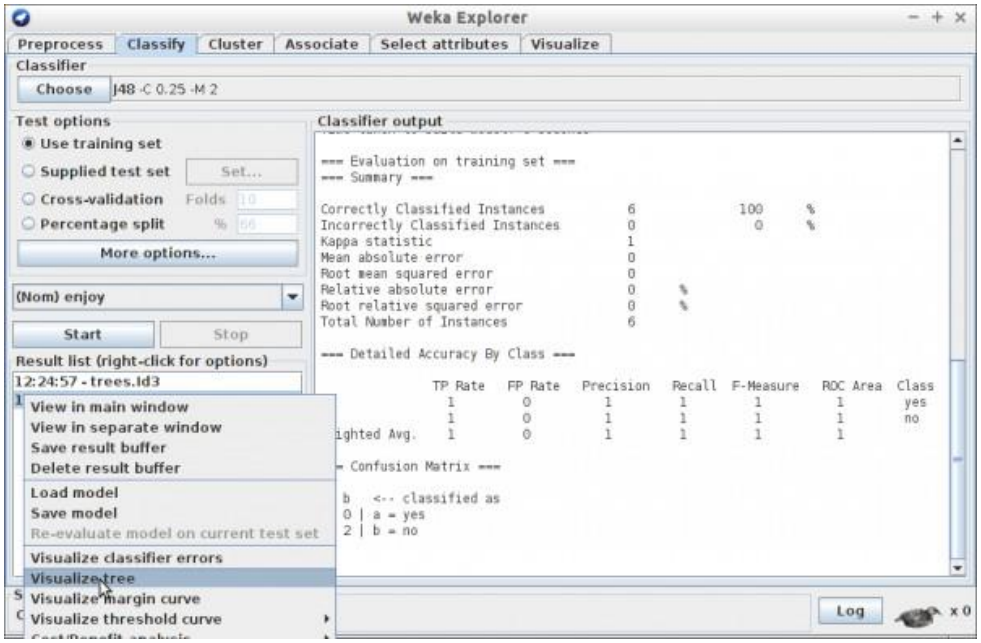
II. Load and Analyse data

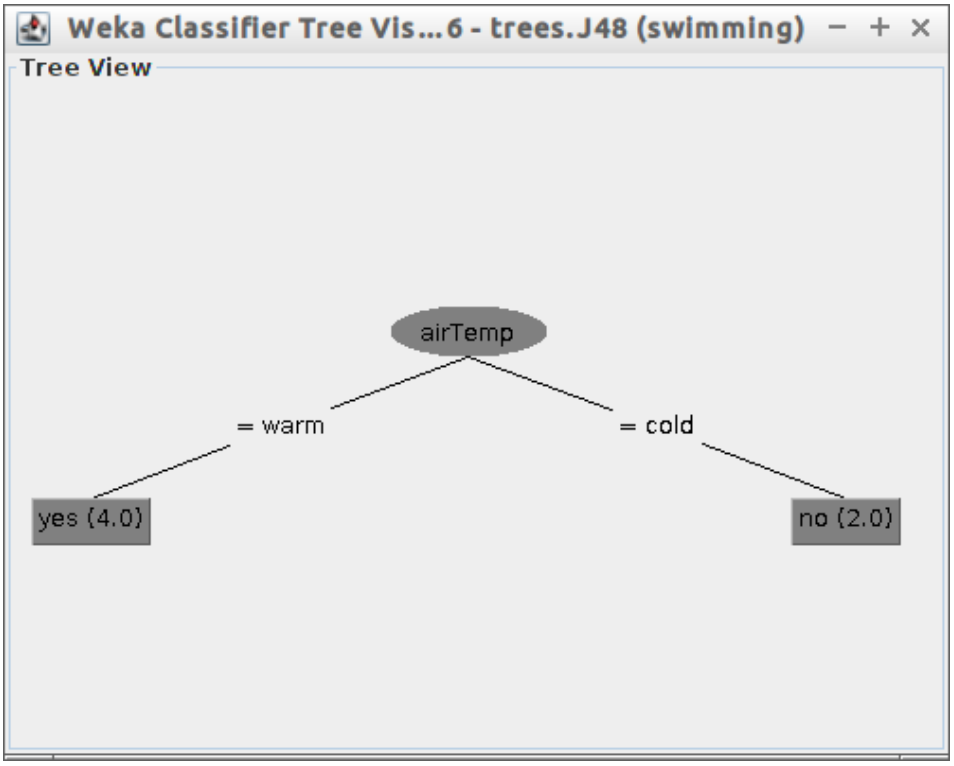
1. Open the Classify tab.
2. Select the J48 classifier using the Choose button.
3. Make sure that '*Use training'* set is checked in the '*Test options'* window. Attention! In the future, we will **not** use this form of testing - we are forced here because of the small training set.

4. Click on the Start button. Look at the result. What do the results mean?



5. Select the J48 classifier using the Choose button and click Start, then visualize the tree as shown below:



6. Does the tree look like this?
7. 

III. Classification accuracy

1. Load the file creditg.arff to Weka. It contains learning data for the system, which on the basis of the attributes contained in the file, should determine whether a given set of attribute values indicates a credibility of bank customers – i.e. whether the bank should grant him a loan or if it is too risky to do so.
2. Open Classify tab and choose J48 algorithm.
3. In the '*Test options'* area, select Percentage split and type in 66%. IT means that 66% of the data will be used for learning and 34% of this data set will be used for validation.
4. Run the algorithm. How many percent of cases were correctly classified? Is this a good result?
5. Change the classifier to ZeroR from the rules branch. What is the obtained result? Better or worse than J48?
6. Try **3 other classifiers**. What are their results?
7. Go to the '*Preprocess'* tab and see how the distribution of the attribute defines whether the set is good or bad. What would be the effectiveness of an algorithm that regardless of the value of attributes would "shoot" that the user is reliable or not?
8. Why is it worth taking a look at the data before attempting a classification task?