Roll No: 60 **Name:** Subrato Tapaswi **Practical:**4

Experiment - 4

Aim: Introduction TO WEKA – A Data Mining Tool

THEORY:

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

The GUI Chooser consists of four buttons—one for each of the four major Weka applications—and fourmenus.

The buttons can be used to start the following applications:

Explorer: An environment for exploring data with WEKA

Experimenter: An environment for performing experiments and conduct- ing statistical tests betweenlearning schemes.

Knowledge Flow: This environment supports essentially the same functions as the Explorer but with adrag-and-drop interface. One advantage is that it supports incremental learning.

Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

Roll No: 60 **Name:** Subrato Tapaswi **Practical:**4

Explorer:

At the very top of the window, just below the title bar, is a row of tabs. The tabs are as follows:

Preprocess. Choose and modify the data being acted on.

Classify. Train and test learning schemes that classify or perform regression.

Cluster. Learn clusters for the data.

Associate. Learn association rules for the data.

Select attributes. Select the most relevant attributes in the data.

Visualize. View an interactive 2D plot of the data.

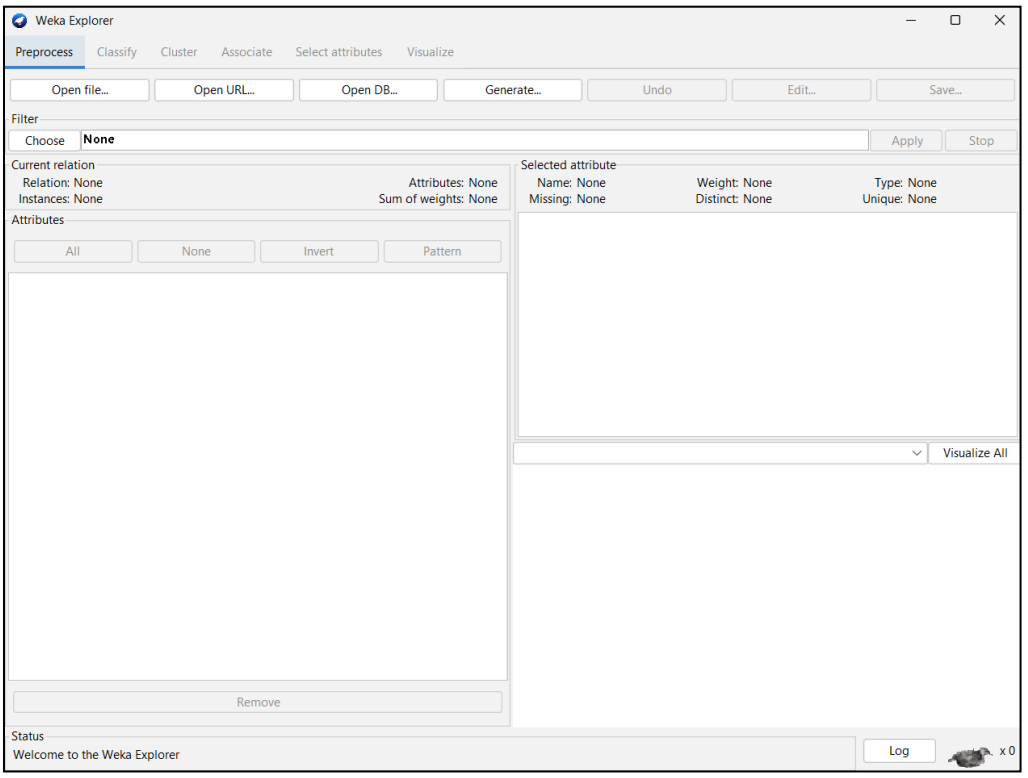
Loading Data:

The first four buttons at the top of the preprocess section enable you to load data into WEKA:

Open file .... Brings up a dialog box allowing you to browse for the data file on the local file system.

Open URL .... Asks for a Uniform Resource Locator address for where the data is stored.

Open DB.... Reads data from a database. (Note that to make this work you might have to edit the file in
weka/experiment/DatabaseUtils.props.)



The Current Relation:

Once some data has been loaded, the Preprocess panel shows a variety of information. The Current relation box (the "current relation" is the currently loaded data, which can be interpreted as a single relational table in database terminology) has three entries:

Relation. The name of the relation, as given in the file it was loaded from. Filters (described below) modify the name of a relation.

Instances. The number of instances (data points/records) in the data.

Attributes. The number of attributes (features) in the data.
When you click on different rows in the list of attributes, the fields change in the box to the right titled Selected attribute. This box displays the characteristics of the currently highlighted attribute in the list:

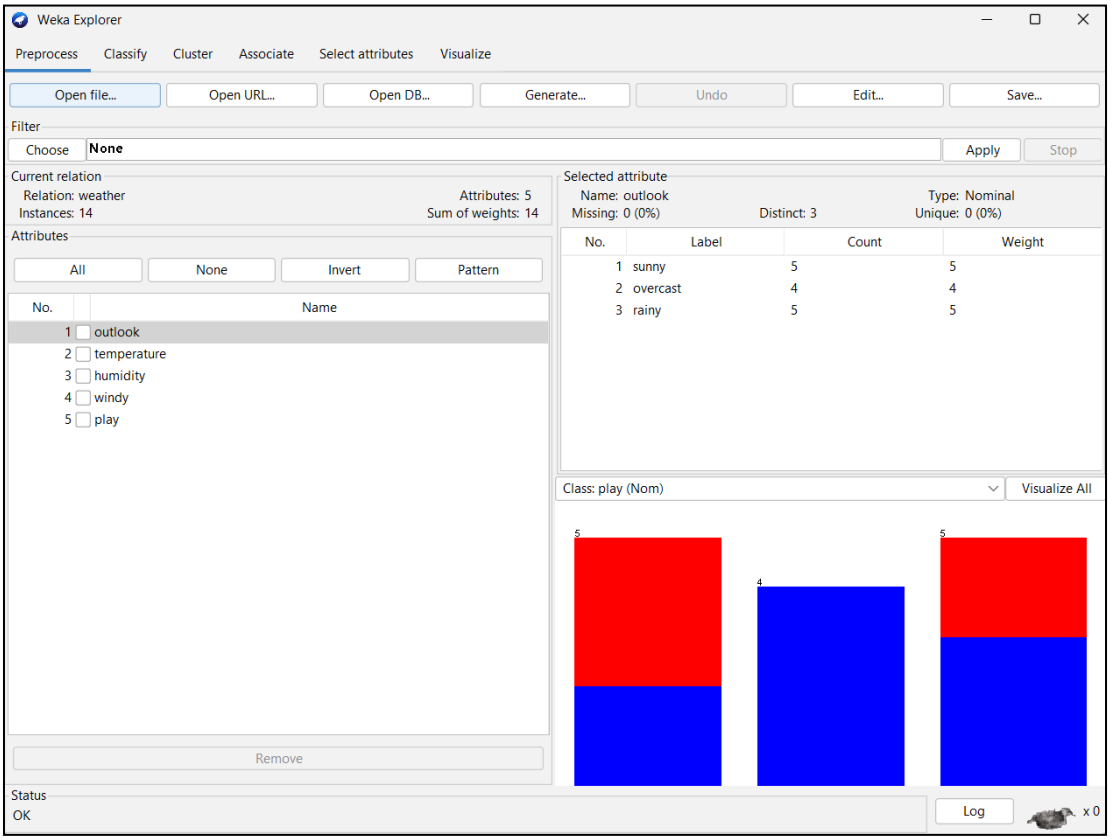
Name. The name of the attribute, the same as that given in the attribute list.

Type. The type of attribute, most commonly Nominal or Numeric.

Missing. The number (and percentage) of instances in the data for which this attribute is missing (unspecified).

Distinct. The number of different values that the data contains for this attribute.

Unique. The number (and percentage) of instances in the data having a value for this attribute that no other instances have.



Below these statistics is a list showing more information about the values stored in this attribute, which differ depending on its type. If the attribute is nominal, the list consists of each possible value for the attribute along with the number of instances that have that value. If the attribute is numeric, the list gives four statistics describing the distribution of values in the data—the minimum, maximum, mean and standard deviation. And below these statistics there is a colored histogram, color-coded according to the attribute chosen as the Class using the box above the histogram. (This box will bring up a drop-down list of available selections when clicked.) Note that only nominal Class attributes will result in a color-coding. Finally, after pressing the Visualize All button, histograms for all the attributes in the data are shown in a separate window. Returning to the attribute list, to begin with all the tick boxes are unticked. They can be toggled on/off by clicking on them individually. The four buttons above can also be used to change the selection:

All. All boxes are ticked.

None. All boxes are cleared (unticked).

Invert. Boxes that are ticked become unticked and vice versa.

Pattern. Enables the user to select attributes based on a Perl 5 Regular Expression. E.g., .* id selects all attributes whose name ends with id.

Result/Output*:*

Classification: - @relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no

Mean absolute error:
In statistics, the mean absolute error is a quantity used to measure how close forecasts or predictions are to the eventual outcomes.

Mean squared error:
In statistics, the mean squared error or MSE of an estimator is one of many ways to quantify the amount by which an estimator differs from the true value of the quantity being estimated.

Root relative squared error:
The root relative squared error is relative to what it would have been if a simple predictor had been used.

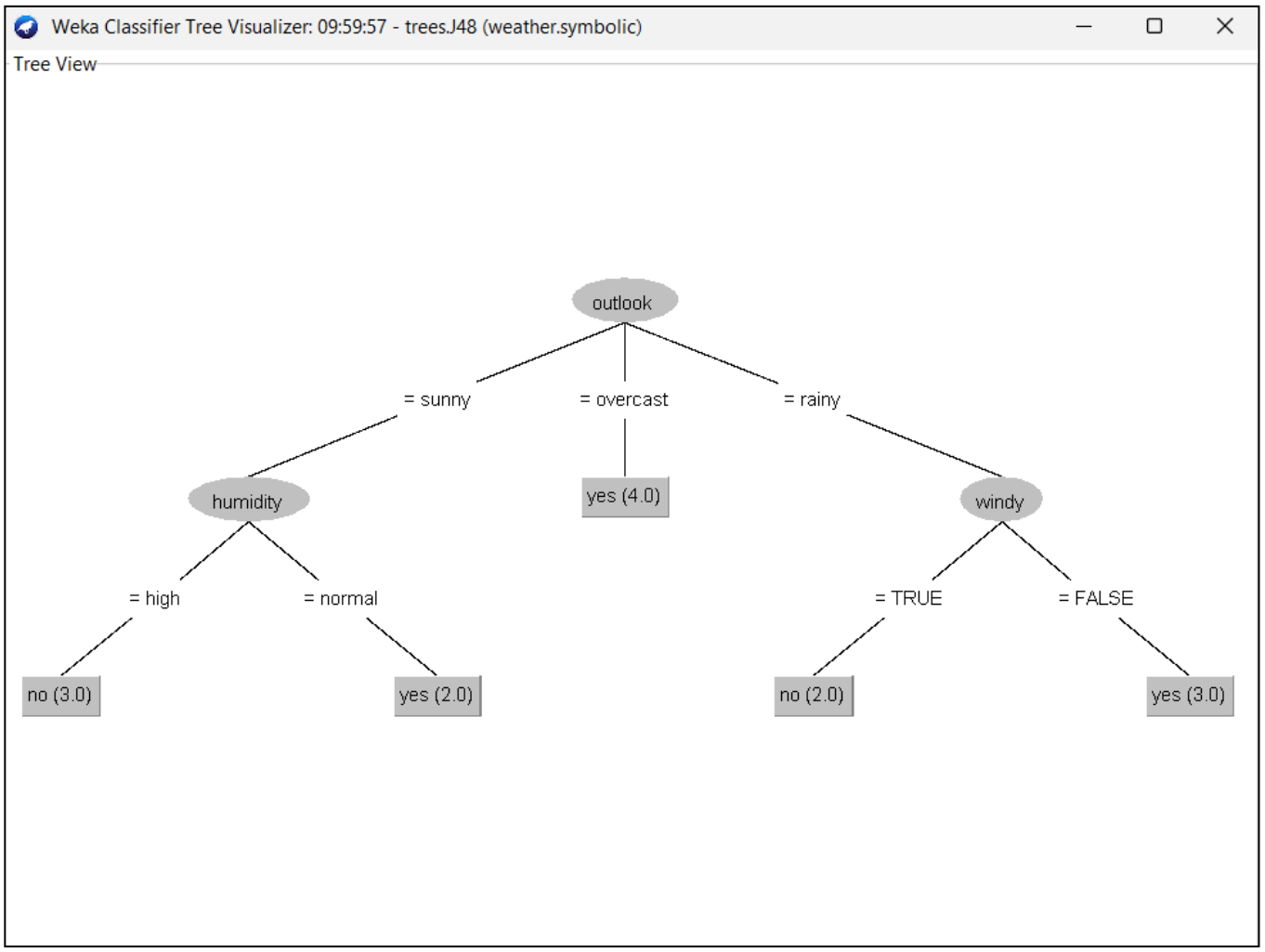
Confusion Matrix:
In Predictive Analytics, a Table of Confusion, also known as a confusion matrix, is a table with two rows and two columns that reports the number of True Negatives, False Positives, False Negatives, and True Positives. The values along the diagonal path are correctly classified instances while those along the non-diagonal path are incorrectly classified instances.

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather.symbolic
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

outlook = sunny
|   humidity = high: no (3.0)
|   humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :     5

Size of the tree :      8


Time taken to build model: 0 seconds
```

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          14               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances               14

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     yes
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     no
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 0 5 | b = no
```

Roll No: 60 **Name:** Subrato Tapaswi **Practical:**4



Conclusion:

Thus, we studied about various Classifiers and implemented them using WEKA.