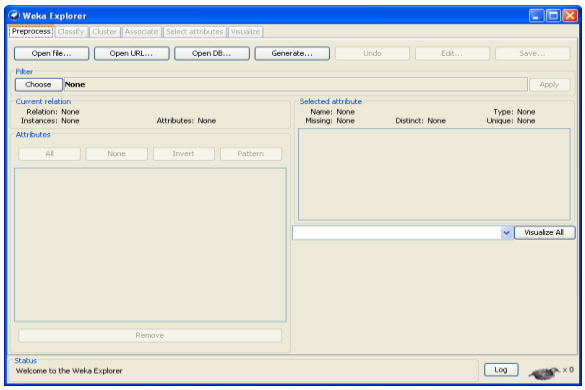
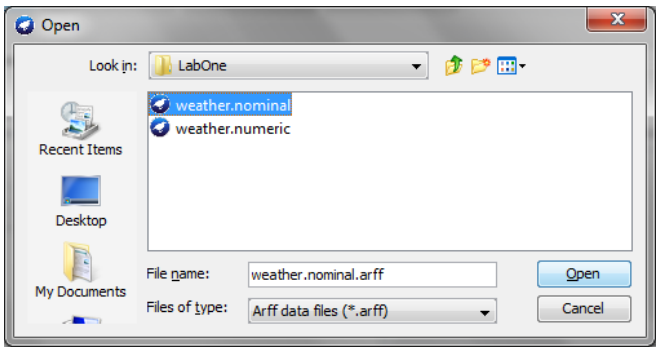
Lab Exercise One
Data Preprocessing with WEKA Explorer

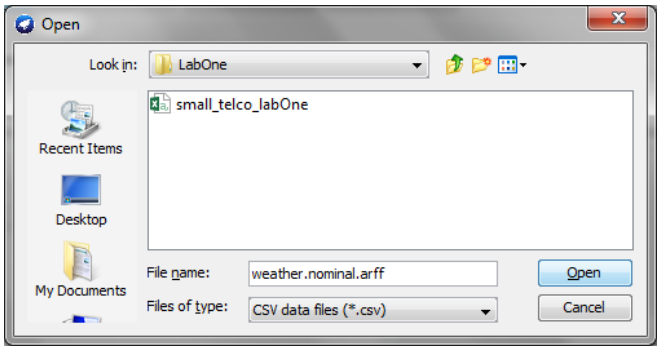
Preview of the raw data

1. Fire up WEKA to get the GUI Chooser panel. Select Explorer from the four choices on the right side.

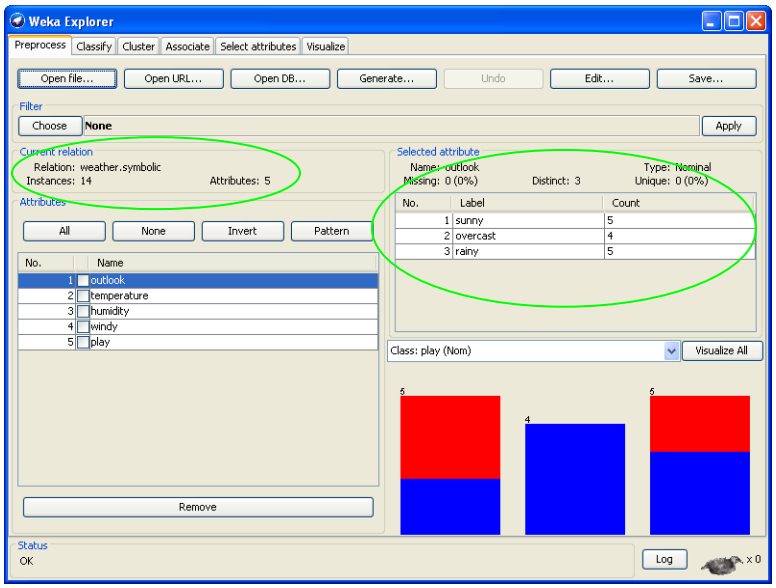


2. The above is the main Explorer screen. There are six tabs along the top are the basic operations that the Explorer supports. We are on ***Preprocess*** now. Click the ***Open file*** button to bring up a standard dialog through which you can select a file. Choose the ***weather.nominal.arff*** file. If you have a file in ***CSV*** format, change from ***ARFF*** data files to ***CSV*** data files. When you specify a .csv file it is automatically converted into ARFF format.

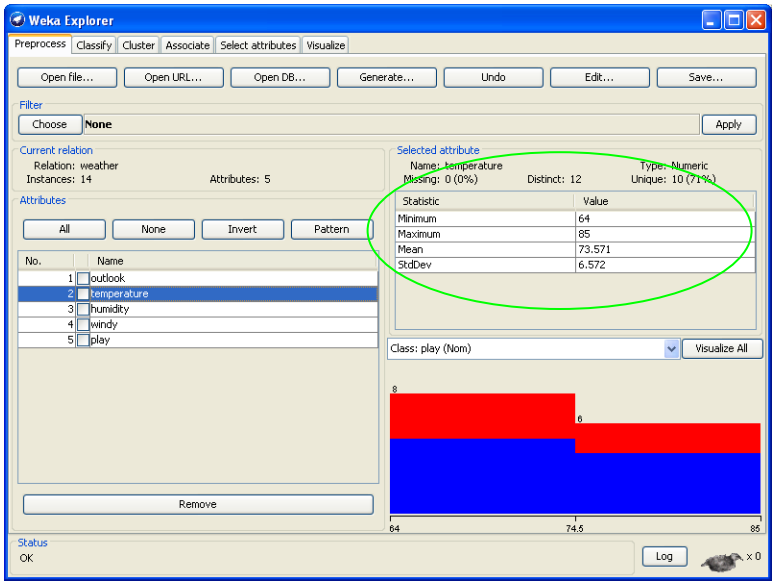




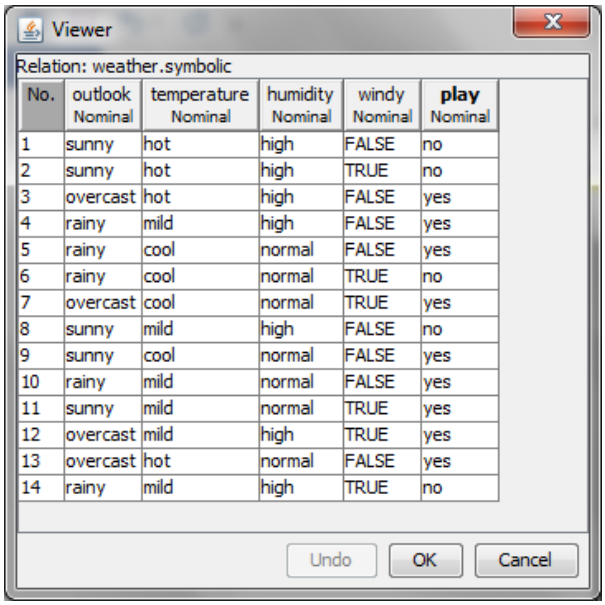
3. The following figure shows the GUI when the ***weather.nominal*** dataset is loaded. The first attribute, ***outlook***, is selected by default. The features of this attribute are shown. A histogram at the lower right shows how often each of the two values of the ***play class*** occurs for each value of the ***outlook*** attribute. You could have the analysis for other attributes by selecting them on the left.



4. If you open the other Weather file, ***weather.numeric.arff***, the visualization of the attribute is different. The second attribute, ***temperature***, is selected and you see its minimum and maximum values, mean and standard deviation. The histogram shows the distribution of the class as a function of this attribute.



5. To see the original dataset, click the Edit button, a viewer window opens with dataset loaded.



No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

6. Click the Visualize tab to visualize the 2D plots of the dataset.