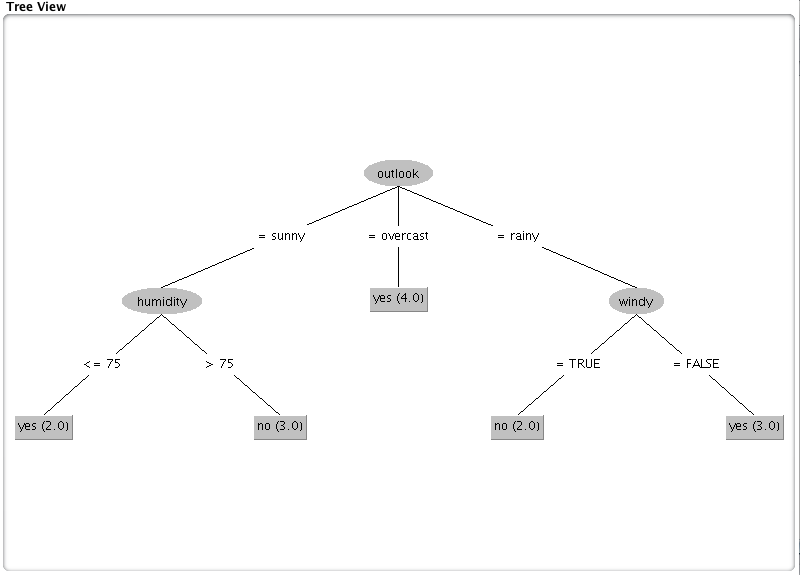
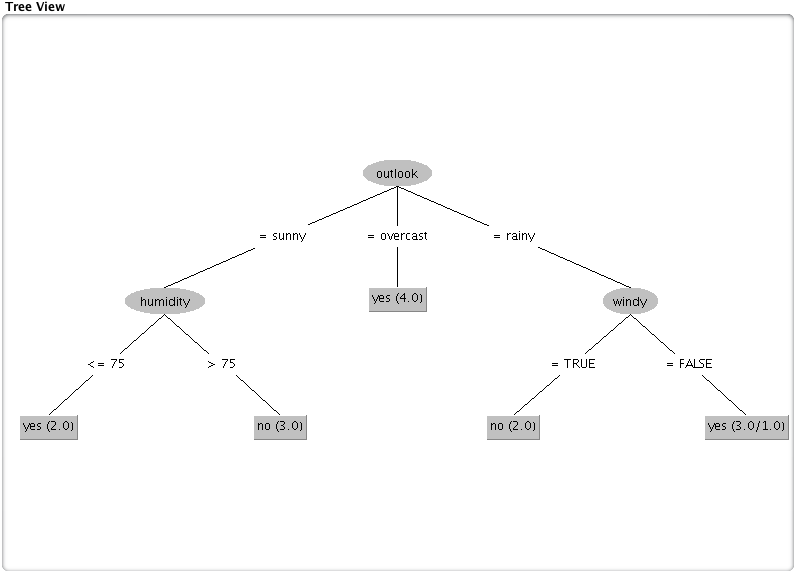
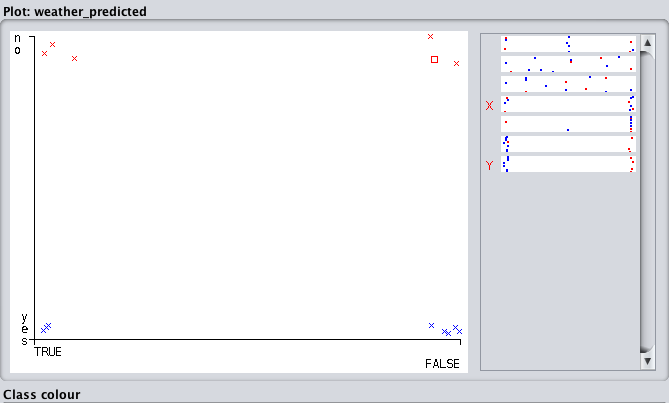
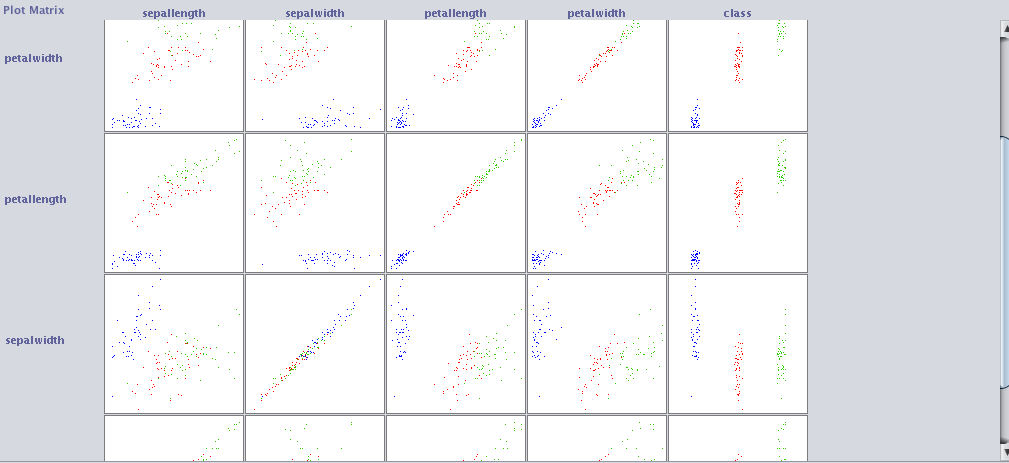
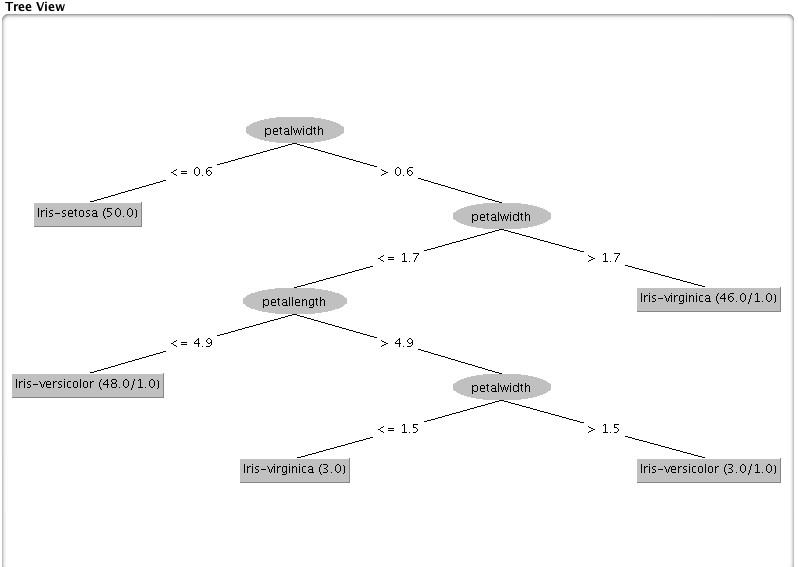
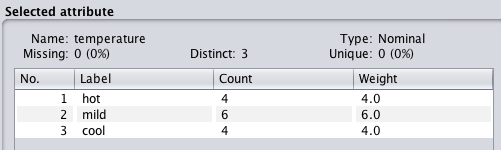
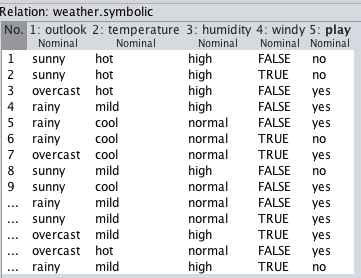
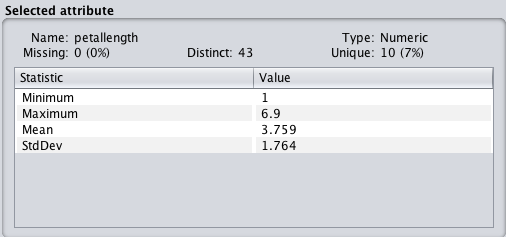
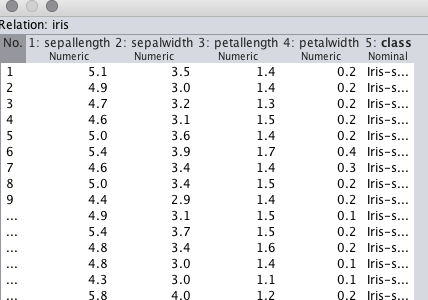
Lab 1: Introduction to the *Weka Explorer*

In this lab, you will learn the basics of looking at data using the *Weka Explorer*. The information in this lab is expressed in more detail in the text Data Mining by Witten et al. (starting on page 369 (edition 2), page 559 (edition 3) and page 553 (edition 4)).

1. Start the *Weka Explorer*.
   1. Weka 3.8 is located in the O: drive under O:\Program Files\Weka3.8.0
   2. Double-click the **Weka shortcut** or the **RunWeka.bat** file.
   3. You should see the *Weka GUI Chooser*.
   4. Press the **Explorer** button.
   5. You should see the *Weka Explorer*.
2. Load the sample weather data.
   1. In the **Preprocess** tab, click the **Open file** button.
   2. In the **data** directory, open **weather.numeric.arff**
3. Investigate the attributes of the weather sample data.
   1. Notice that in the **Current relation** pane, you can see the number of instances and the number of attributes in the data.
   2. Selecting an individual attribute such as **temperature** will show some basic statistics (e.g. mean temperature).
   3. The lower-right pane shows the distribution of a target attribute (e.g. **play**) for the values of the selected attribute.
4. Run a J48 classifier.
   1. Go to the **Classify** tab, click **Choose** and select the **J48** classifier (which can be found under the **trees** group).
   2. Under **Test Options**, select **Cross-validation**, and hit **Start**.
   3. Examine the output generated in the right-hand pane. In particular, you should get a tree that looks something like this:

|  |
| --- |
| J48 pruned tree  ------------------  outlook = sunny  | humidity <= 75: yes (2.0)  | humidity > 75: no (3.0)  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 |

* 1. Figure out what the tree means in the text format.
  2. Verify your understanding by looking at the pictorial representation. Do this by right-clicking on the entry in the **Result list** and clicking **Visualize tree**.  
     
  3. Notice that a number is shown next to each leaf of the tree. This is the number of instances which reached that leaf. For example, the leaf “windy = FALSE: yes (3.0)” means that 3 instances reached that leaf and they're classified as “play=yes”, furthermore, they are all correctly classified.
  4. If 1 of the 3 instances had been incorrectly classified, the leaf: “windy = FALSE” would indicate: “yes (3.0/1.0)”. Verify this by going back to Preprocess, clicking Edit, and changing the 10th instance from play = yes to play = no. Click OK on the editor. Go back to the Classify tab and run J48 again and look at the tree.  
     
  5. Now under **Test options**, try **Use training set** and click **Start** again.
  6. Examine the output to see the differences. You can go back to your previous result by selecting it in the **Result list**.
  7. Right click on your last result and select **Visualize classifier errors**. This gives you a way of plotting one attribute against another and seeing the distribution of the target class. You can distinguish instances which are classifying correctly (plotted as **x**) from those that are errors (plotted as boxes). In this case, you should see one red box. For example, set X: windy and Y: play. You will see that for Y = yes, there are 8 blue x's, three are at windy = TRUE and 5 are at windy = FALSE. At the level of Y = play = no, you see five red x's and one red box.  
     

1. Look at a more complex data set.
   1. Re-run steps 2 through 4 on **iris.arff**, noticing that with this data set, not all instances are classified correctly.
   2. Click on the **Visualize** tab. This tab shows a plot for each pair of attributes. Adjust the **Jitter** and press **Update** to see the data points more clearly.
2. Hand in:
   1. From your observations in “5. b)” predict which pairs of attributes separate the three classes well. Also, predict which pairs are less effective. You can restrict your consideration to pairs (x, y) where neither x nor y is the class attribute and where x and y are not equal.  
      - petallength + petalwidth separates the three classes well  
      - sepallength + sepalwidth are less effective  
      
   2. Compare your predictions (above) with the **J48** tree for this data.  
      
   3. Load weather.nominal.arff. What values can the attribute temperature have?  
      hot, mild and cold 
   4. What is the class value of instance number 8 in the weather data?  
      play = no  
      
   5. In iris.arff, what is the range of possible values of the attribute petallength?  
      1-6.9  
      
   6. In the iris data, how many numeric and how many nominal attributes are there?  
      4 numeric and 1 nominal attributes  
      
   7. Load weather.nominal.arff. In the Preprocess tab, click on Choose. You will see a hierarchical filter selection. Choose the filter weka.unsupervised.instance.RemoveWithValues. After clicking on that filter, on the main Preprocess panel click on “RemoveWithValues” to get a panel where you can set the attribute and values to remove. Remove all the instances for which the humidity attribute has the value high. (Hint: humidity is the third attribute (3) and the index of “high” can be seen before going into the filter selection by clicking on humidity in the Attribute section.) Click on “Apply” after you have made the filter selection. You can check that you have done it right by clicking on humidity again and looking in the Selected attribute panel. Now run J48 with Test option = “Use training set” and report the percent of “Correctly Classified Instances”.  
      Correctly Classified Instances 6 85.7143 %  
      