# Lab 03 - Data Pre-processing in Weka

1. Fire up the Weka (Waikato Environment for Knowledge Analysis) software, launch the explorer window and select the “Pre-process” tab.

2. Open the iris dataset. What information do you have about the data set (e.g., number of instances, attributes and classes)? What type of attributes does this data-set contain (nominal or numeric)? What are the classes in this dataset? Which attribute has the greatest standard deviation? What does this tell you about that attribute? (You might also find it useful to open “iris.arff” in a text editor).

3. Under “Filter” choose the “Standardize” filter and apply it to all attributes. What does it do? How does it affect the attributes' statistics? Click “Undo” to un- standardize the data and now apply the “Normalize” filter and apply it to all the attributes. What does it do? How does it affect the attributes’ statistics? How does it differ from “Standardize”? Click “Undo” again to return the data to its original state.

4. At the bottom right of the window there should be a graph which helps to visualize the dataset, making sure “Class: class (Nom)” is selected in the drop-down box click “Visualize All”. What can you interpret from these graphs? Which attribute(s) discriminate best between the classes in the dataset? How do the “Standardize” and “Normalize” filters affect these graphs?

5. Under “Filter” choose the “Attribute Selection” filter. What does it do? Are the attributes it selects the same as the ones you chose as discriminatory above? Try different evaluators? Are the outcomes different?

6. Select the “Visualize” tab. This shows you 2D scatter plots of each attribute against each other attribute (similar to the F1 vs F2 plots from lab 1). Make sure the drop-down box at the bottom says “Color: class (Nom)”. Pay close attention to the plots between attributes you think discriminate best between classes, and the plots between attributes selected by the “Attribute Selection” filter. Can you verify from these plots whether your thoughts and the “Attribute Selection” filters are correct? Which attributes are correlated?

7. Under the Classify tab, apply the JRip classifier to the data set. Examine the confusion matrix and note the results of how accurate the resulted model is.

8. Click on the Pre-process tab and restore all the original attributes by clicking ‘Undo’. Apply the JRip classifier to the full data set. How well does this model compare with the model that was generated when only the selected attributes were was used. What are your reasons?

9. Open the diabetes data set and repeat steps 5 to 8 above. Is the selection consistent with your impression formed in lab 1? Is there anything else you can learn about the dataset?