University of Reading

Module convenor: Dr. Carmen Lam
Exercises created by Prof. Giuseppe Di Fatta

MSc Data Science and Advanced Computing
CSMDM21 - Data Analytics and Mining

Practical 1: KNIME basics

These are many publicly available datasets (e.g., http://archive.ics.uci.edu/ml/) that can be used to practice. These two datasets are required for this practical and are also available as on Blackboard.

- http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
- http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data

Exercise 1: I/O and Data Manipulation

Workflow 1: Data reading and cleaning

1. Create a KNIME workflow and use the node *File Reader* to read the data file directly from the URL http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
   - The **.data** file **does not contain column headers**, which can be found in the corresponding files with extension **.names**: http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names
   - It is a good practice to add column headers to make data exploration more user friendly. This can be achieved in different ways.
     1. The simplest method is to change the column headers directly into the node *File Reader*.
     2. Another simple way is to use the node *Column Rename*, which is specifically meant for this purpose. However, in both cases the process is tedious as the data contains many (15) attributes.
     3. An advanced method is to build and use a dictionary to change the column headers of a table to the desired values. Create a text file with the list of desired column headers (one header per line) or download **adult_headers.txt** from Blackboard. Read the column header file in the workflow and use it to create a dictionary old_name vs new_name for the column headers. For this task you can use the nodes: *Extract Column Header, Transpose, RowID, Joiner,* and *Insert Column Header*.

2. Download **adult_file_1_raw_data.csv** from Blackboard (this one contains column headers) and use the node *File Reader (Complex Format)* to load this data file in the KNIME workflow. Then, remove or replace missing values:
   - If missing values is in a String column, then remove rows.
   - If missing values is in a Number column, then replace with the mean value.

   After that, you may want to reset the row IDs of the table with the node *RowID*. Finally, write the cleaned data into a file **adult_file_2_clean_data.csv**. Do not forget to include the column headers in the new file by enabling the option "Write column headers" in the configuration dialog of the node *CSV Writer*.

Workflow 2: Row and column filtering

1. Create a KNIME workflow to read data file **adult_file_2_clean_data.csv**. Process data by using *Row* and *Column Filter* nodes, then view the filtered data using *Interactive Table.*

   For row filtering, select:
   a. all data sample where the person's age is under 30 and view in an interactive table;
   b. all female data samples and view in an interactive table;
   c. only first 100 data samples and view in an interactive table;
   d. remove all records with education labelled as the "some-college" and view in an interactive table.

   For column filter select:
   - specific data features/attributes that are relevant only for personal information.

Workflow 3: Horizontal and vertical data splitting

1. Create a KNIME workflow to read data file **adult_file_2_clean_data.csv**. Process data to create four subsets of data using *Row* and *Column Splitter* nodes and write the subsets into separate CSV files:
   - The first subset contains all Male data records **only.** Write subset into a file **adult_file_3_male.csv**.
   - The second subset contains all Female data records **only.** Write subset into a file **adult_file_4_female.csv**.
   - The third subset contains all data with only occupational related information. Write subset into a file **adult_file_5_occupational.csv**.
   - The fourth subset contains only personal information. Write subset into a file **adult_file_6_personal.csv**.

Workflow 4: Concatenation and Joining

1. Create a KNIME workflow to read data files **adult_file_3_male.csv** and **adult_file_4_female.csv** and concatenate the subsets (after that, you may want to reset the row IDs of the table with the node *RowID*), then view the concatenated data in an interactive table.
2. Also read data files **adult_file_5_occupational.csv** and **adult_file_6_personal.csv** and join the subsets, then view the joined data in an interactive table.

Exercise 2: Data Visualization

Workflow 1: Scatter Plots

1. Create a KNIME workflow to read data file **iris.data** using the node *File Reader* directly from the URL http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data or from a local copy of the file (download from Blackboard).
2. Rename the five columns respectively as "Sepal length", "Sepal width", "Petal length", "Petal width", and "Species".
3. Use the scatterplot to visualize the data and experiment with assigning different colours toeach species, e.g., red for Iris setosa, purple for Iris versicolor, green for Iris virginica. Similarly, experiment with assigning a different shape to each species.
4. Use the node *Column Resorter* to update the columns for plotting them using the *Scatterplot* node.
5. Also, plot each feature of the iris data into a 4×4 Scatter matrix to view all feature simultaneously in one plot.

Workflow 2: Simple Statistics and Other Plots

1. Copy nodes created in step 1 and 2 of Workflow 1 to Workflow 2.
2. Use the node *Statistics* to view statistical moments such as minimum, maximum, mean, etc.
3. Plot data using *Line Plot*, *Bar Chart*, *Box Plot*, *Conditional Box Plot*, *Parallel Coordinate*, *Pie chart* and *Histogram*. Explain each plot under nodes label.

Solutions:

Sample solutions to these exercises are available on Blackboard (Bb) in two forms: images of the workflows and the actual KNIME workflows (as a single zip archive). You should first try to build your own workflows for each exercise. During the practical session, you may use the images to see the proposed solutions and to reconstruct them. During or at the end of the practical session, the archives with the actual KNIME workflows will be made available for you to import and test them. These are KNIME archives (file extension ". knwf ") that can be imported in KNIME. Note that after importing the KNIME archives, you may need to change the file locations of the source data file and the output file destination folder.