Big Data Analytics
Prof. Giacomo di Tollo et al.
Summer Semester 2024
University of Luxembourg
Exercise Sheet #3 – Due on May 15th, 2024

Incremental Exercise Sheet

Raw Data Processing **2 Points**

Problem 1.
Let's assume that the file –ClusterBicluster.zip– on Moodle contains -csv files coming from the execution of a parallel task on a cluster. For each file, rows 0 to the one labelled *Ratio Zero to* represent the raw output of our runs.
Each row identifies a city; each column identifies a variable. Our goal is to resume the outcomes of the experiments similarly to the image *ClusterPlot*, in which, for each subplot (representing a file), x-axis represent variables, and y-axis represent variables' values. Each plot line connects variables values representing the variables values associated to a city.
The goal of the exercise is to produce plots from the image *ClusterPlot*, associated to the new set of experiments.

Decision Trees **8 Points**

Problem 2.
The two Excel files named –Before-Pre-Processing– and –After-Pre-Processing– contain information coming from the same set of attributes, whose *raw data* are firstly manually inspected and explored (file –Before-Pre-Processing–) and then prepared and transformed into training data format (file –After-Pre-Processing–).
Please implement a Decision Tree to predict variables $S1Q061$ and $S1Q06P1$ (one tree for each predicted variable) by using all all other variables in the predictor set. Then, please focus on a smaller predictor set, to be decided by the lecturer. Please report the *precision*, *recall*, and other performance metrics discussed during the lectures.
Last, please determine a more autonomous method to determine what variables have to be contained in the predictor set.