**A report on**

**INFO-H423 - Data Mining. Assignment – part 1.**

By

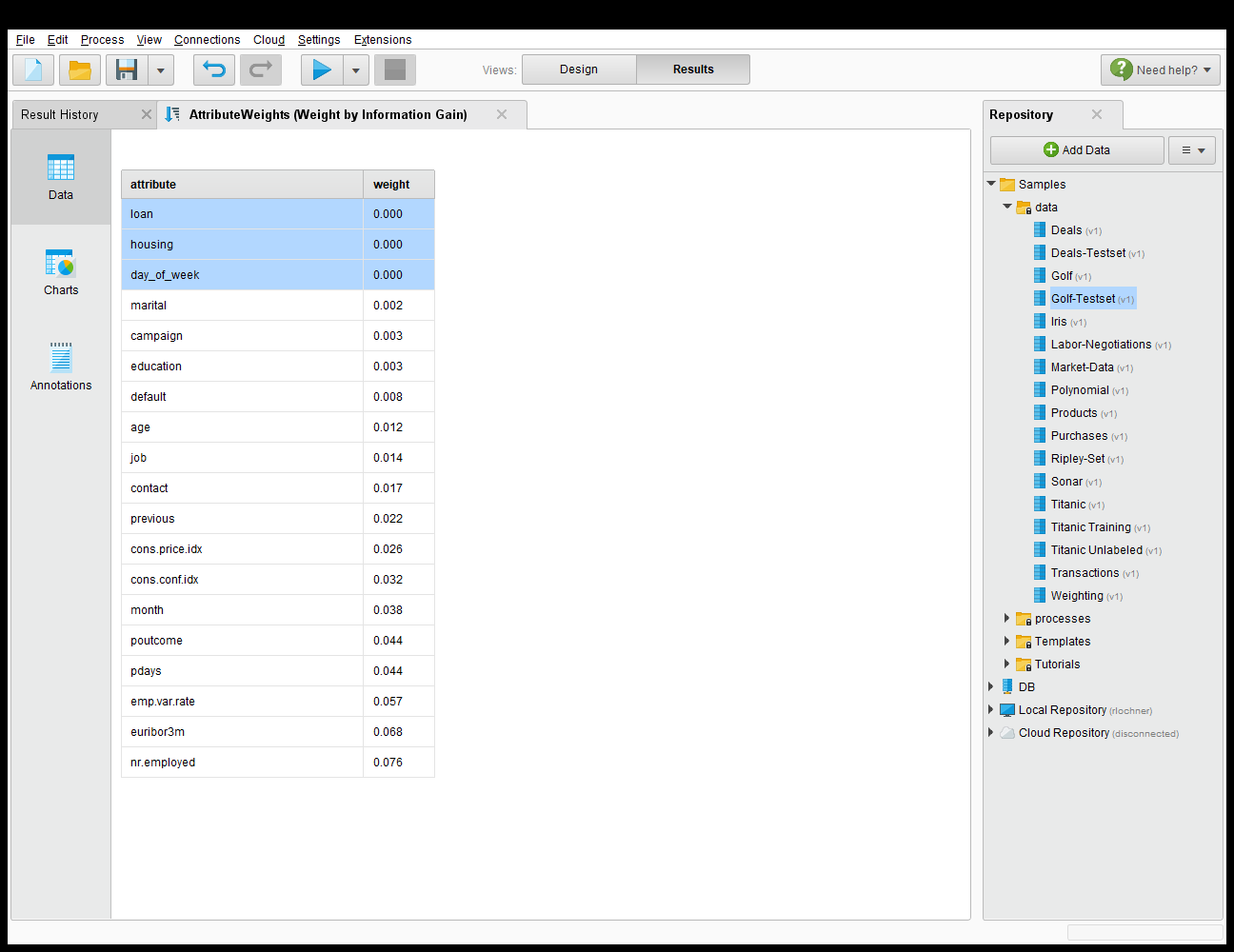
Raymond Lochner (???)

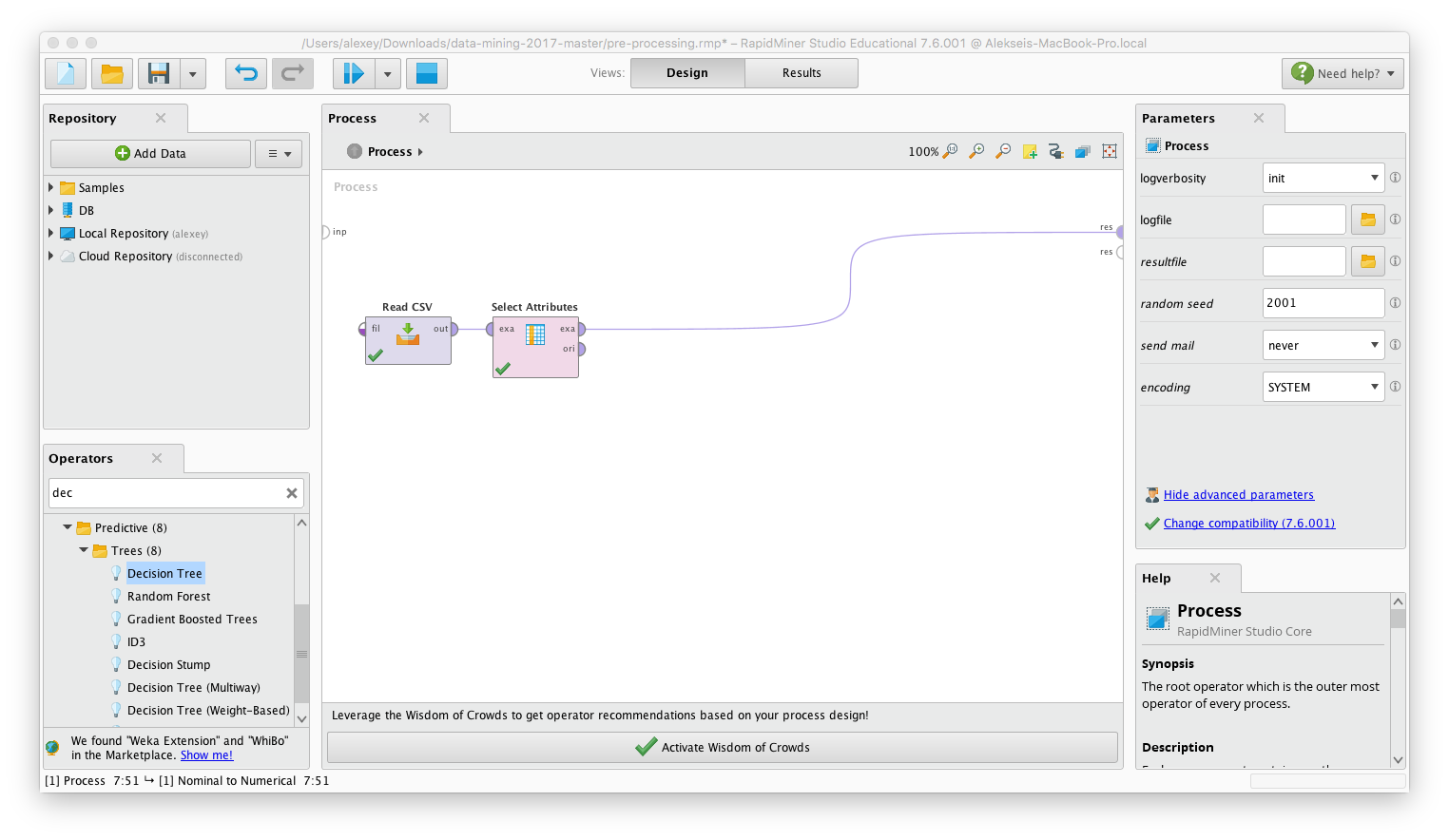
Aleksei Karetnkiov (000455065)

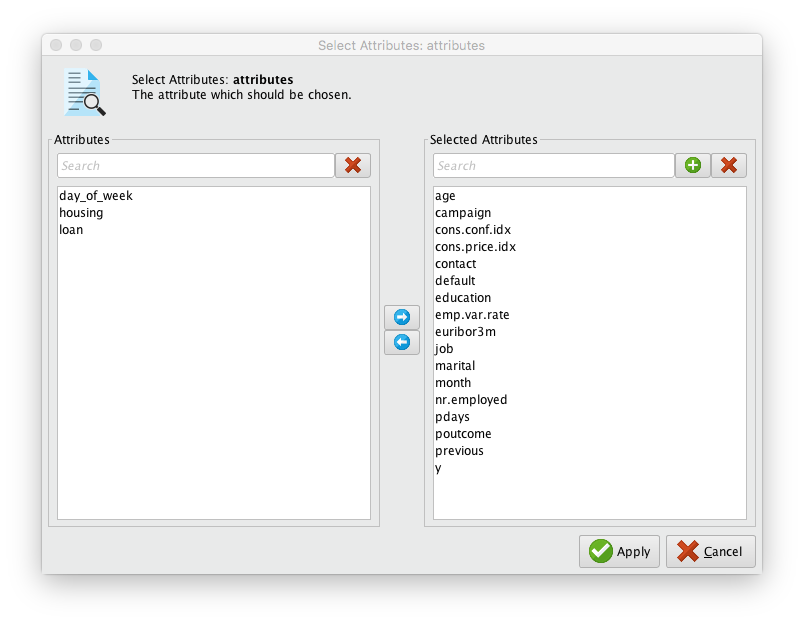
Aldar Saranov (000435170)

# Preprocessing Stage

The input file has been successfully loaded and the appropriate roles have been assigned. In order to detect and remove attributes which do not contribute to the prediction, we measure the information gain for each attribute. We find that certain attributes have negligible information gain (load, housing and day\_of\_week).







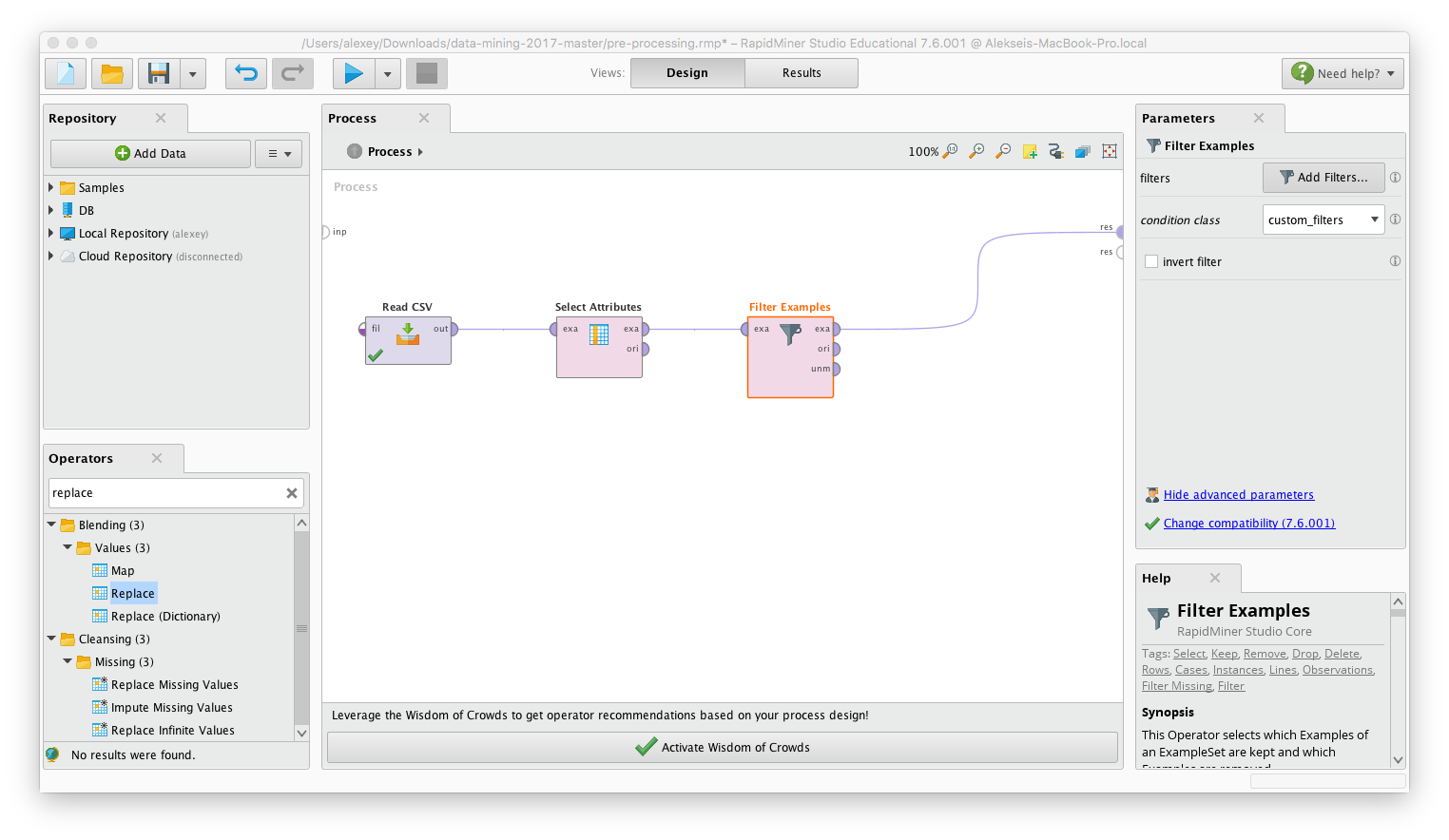
Then, we need to consider the “unknown” values. On condition that our dataset has 41.188 entries we suppose that it is enough to have a 5% significance level, which is about 2.000 entries. It is necessary to optimize the task without any significant influence on the result of the whole research. So, we can remove the entries, which has smaller number of “unknown” values and recovery the others. To recover the damaged values we will use the further method of the interpolation, which based on the correlation of values with values of other attributes:

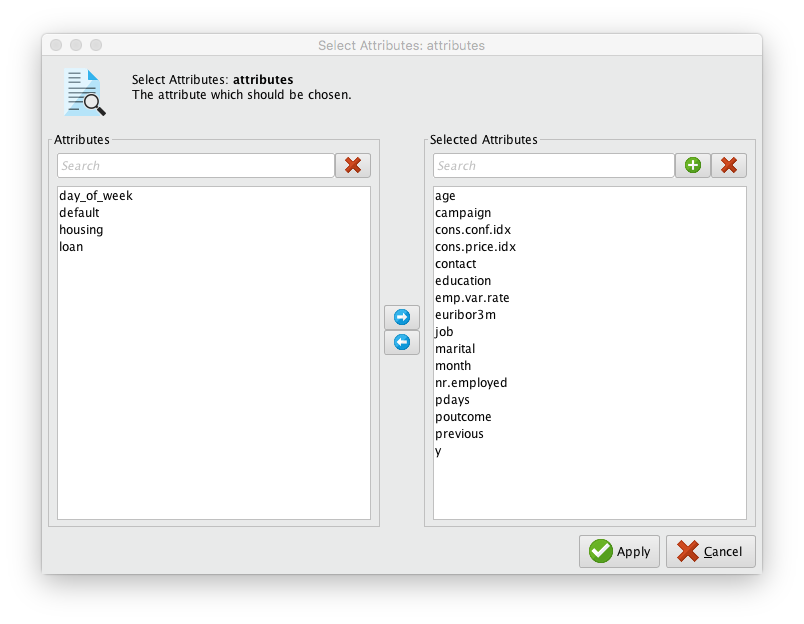
1. We need to find the maximal absolute value of the correlation with other attributes;
2. Consider the sign of the correlation coefficient;
3. Find a pattern;
4. Replace these “unknown” values by the found pattern.

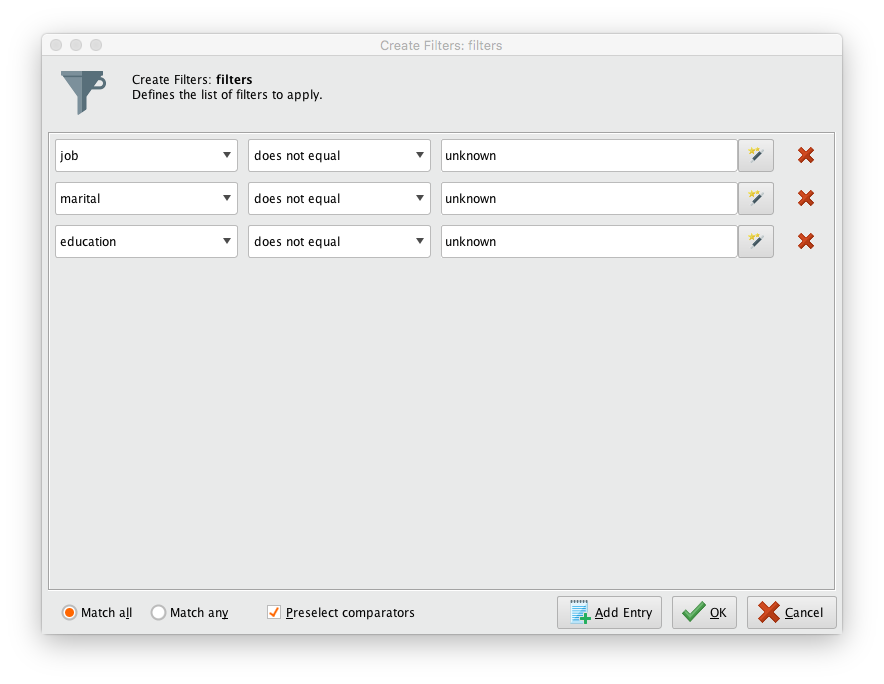
We need to recognize all the attributes with the possible “unknown” values, considering that attributes “load”, “housing” and “day\_of\_week” excluded of the research on the previous step.

|  |  |  |
| --- | --- | --- |
| Attribute | “Unknown” values | Ratio of the “unknown” values |
| Job | 330 | 0.008 |
| Marital | 80 | 0.002 |
| Education | 1731 | 0.042 |
| Default | 8597 | 0.209 |

Regarding the results, we can see the we need to work only with “default” attribute and remove other entries to optimize the task. Regarding this attribute, we can see that we have 32.588 “no” (79.25% of the all entries) and 3 “yes” (0.007%) values. Unfortunately, this set is not enough for building any model. So, we can assume that other “unknown” values could be “no”. Since, more than 99% of the “default” values are the same, we could also exclude this attribute, because it will be useless.







Finally, we have to replace all the nominal values by corresponding numerical to optimize the dataset for the further analysis. We will save the resulting preprocessed dataset to the local repository to simplify access to it.

