CSI5155: Assignment 1 report

Adrien Heymans

# Lessons learned

One of the main lessons observed is that one single classifier, using the same features for each drug will not give satisfactory results. Depending on the targeted output, it is necessary to drop some features. This demonstrate that before even starting the classification task, the pre-processing step is crucial to ensure optimal results. It is a time-consuming but necessary step. If the pre-processing is correctly completed, the classification part of the problem will be easier to complete and give more accurate results. The bottom-line is that the model and features are not uniform for a dataset, it depends on the classification task and context.

# Results

The results obtained presented both similarities and differences from the ones in the paper.

In terms of pre-processing, we have discovered that some drugs have little to no correlations with other drugs. In details “Chocolate”, “Caffeine”, “Alcohol” and “Semeron” (a fictional drug used to discard respondents to the survey), have almost no correlations with other drugs. Our common sense would tell us that this is probably because everyone consumes them frequently, and it tells nothing about an individual’s predisposition to drug use.

Regarding the feature selection, our algorithms have made different choice when it came to select features for individual drugs. However, the main similarity is that the “Age” feature was always selected by our model, it was also the most selected feature in the paper.

In terms of classification, we focused on 4 models (DT, RF, KNN, SVM), the research paper did not use SVM but used a variety of other models, including Naïve-Bayes or Logistic Regression, it would have been in interesting to see how our dataset would have reacted with our models and our feature selection.