Class Project 2: A Case Study in Data Mining

In this project, each team will use a publicly available dataset, define a data mining problem, and then use at least **three different** mining algorithms to mine the dataset, compare and contrast the performance of your chosen models, and interpret your findings. Teams will write a report summarizing their findings.

To look for a dataset, you can go to Kaggle.com or UC Irvine machine learning data repository. There are plenty of data for you to choose from. Make sure you read files that describe the data before you engage in mining.

You can use Rapid Miner, Python, Excel, or any tool of your choice – as we discussed in class, the key skill of a data analyst is NOT the tool that he/she uses, but how he/she applies the right model for the problem at hand, and how he/she interprets the results to make a difference.

I would STRONGLY recommend that you choose a **supervised learning problem**, as it is a lot easier to measure the performance of your outcomes. In case you choose a supervised problem, the resulted **confusion matrix** and **performance parameters** need to be shown in your report and discussed. Specifically, what are the **precision and recall rates**? Which one serves as a more important measurement metric to this problem and why? Which algorithm performs best? What parameters were tuned, and how does that make a difference?

In your discussion, focus on your findings and their significance, both mathematically and practically.

As a team project, I would strongly recommend that you choose ONE dataset to work with. However, each team member should **INDIVIDUALLY** work on **one** algorithm and report his/her results.

Here is a suggested structure of your final report:

PART I. Introduction (to be completed as a team): Introduce the problem at hand to the reader, including the background of the problem and any domain knowledge necessary.

PART II. Data (to be completed as a team): Summary statistics of the data shall be presented here, including general measurements such as number of records, number of features, the label (if you have any), mean, standard deviation, etc.

PART III. Method (to be completed by each member **INDIVIDUALLY**): In part three, each team member will choose ONE data mining algorithm to work on the data set and report the result. Please report the type of algorithm chosen, a screenshot of your workflow and screenshots of the parameter setup if you are using RapidMiner, or your code if you are using python, R, or any other language. If it is a supervised learning problem, please include the confusion matrix and discuss your findings and performance. If you are working on an unsupervised problem such as clustering or association rules, please present the clustering or rules that you found and discuss the implications. Each team member will be evaluated based on his / her performance.

PART IV. Conclusion (to be completed as a team): This is where you put everyone's model together and compare their results. Please discuss:

- Which evaluation metrics should be used (Recall or precision) and why
- Which algorithm works best, at what parameter setup
- What does the finding tell you? Can you take action to improve the performance and solve a problem based on the findings? Do you have any recommended action?