'Data Mining and Knowledge Discovery'
Data Mining Project with R

Master 1 MLDM
Saint-Étienne, France

Fabrice Muhlenbach

Laboratoire Hubert Curien, UMR CNRS 5516
Université Jean Monnet de Saint-Étienne
18 rue du Professeur Benoît Lauras
42000 SAINT-ÉTIENNE, FRANCE
https://perso.univ-st-etienne.fr/muhlfabr/

1 Objectives

The data mining project is not about finding a "good" dataset. The objective is (1) to search a real life and/or everyday life problem and (2) to find a sufficiently large dataset for being able to apply data mining techniques with R to find some answers to this problem.

An important part of this project is to find "amazing knowledges" (interesting, unexpected, or valuable structures) that are embedded in a large dataset.

2 Report

The data mining project report must be a 6-page document (PDF version) including an analysis describing the following six points: (1) Problem Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation, and (6) Deployment.

3 Datasets

With the arrival of the "Open Data," it is now possible to find data on multiple subjects. For example, you can find some datasets on:

- **Kaggle** website: a platform for predictive modeling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

- Official open data portals for some countries, cities, or institutions.

- Less interesting: on *UC Irvine Machine Learning Repository*, the datasets are often too small to apply interesting data mining techniques, the problems have already been studied by many data scientists, therefore it is difficult to find some surprising elements.

4 Storage

Your ® code as well as the dataset must be hosted on a GitHub repository.

Note that you can easily work with GitHub within RStudio (see for example this tutorial).

Moreover you can create your own ® package. An ® package stored on GitHub can be automatically installed with devtools or githubinstall packages (see for example this tutorial).

5 Content

The report should not be very big (6 pages) but must follow the logical data miner/data scientist guidelines and be able to answer the following questions:

- Who has created the dataset? How did you find this dataset? Where can we find it? If you created the dataset yourself, how did you do it?

- How many data? (number of observations, number of variables, file size)

- What are the data mining and machine learning methods used? Why this choice? Have you achieved good results?

- How long did the computational process take? (We will see how to time a process with ®)

- Why did you chose this problem?

- What can you conclude from your study?

6 References and Plagiarism

It is important to clearly mention the different bibliographic sources used in your work, be they books, articles, tutorials found online or explanatory videos on YouTube or elsewhere. It is thus necessary to have these different information in the reading of the report with a section "References" at the end of your report recapitalizing all the bibliographic resources used.

For the writing of this bibliography, it is recommended to follow the indications that you will be able to find for example on the ACM Citation Style and Reference Formats.

Note that we are particularly vigilant on the issue of plagiarism. We use anti-plagiarism software that looks for all forms of similarity between your own report and web pages, books, and academic report libraries. With some specific tools, it is also possible to detect texts or codes automatically generated by Transformers and large-language based models like the chatbot *ChatGPT*.

7 Date

You must send your report before Tuesday, April the 30th 2023, 11:59 PM (CET, Paris or Saint-Etienne local time) by electronic mail to your teacher: fabrice.muhlenbach@univ-st-etienne.fr (note that the enclosed document must have less than 10 MB).