Project instructions
Data Mining I

The objective of the Data Mining I project is to test yourself on a real knowledge discovery process using the methods and algorithms learned during the course. This is, by design, an **independent** activity, where you are expected to identify relevant questions that can be answered using Data Mining methods, autonomously reason about the encountered problems and identify appropriate solutions, and it is a task based on *real data*: nothing has been edulcorated to simplify your life or to force some educational concepts to emerge.

These are the basic instructions:

GROUPS
- The project is performed by the groups formed on the student portal.

EXAMINATION
- The project is on the G/U grading scale (pass/not_pass).
- To pass the project, you must:
    1. Choose one Data Mining question regarding the data.
        a. That is, you must recognize that what you are asking for requires the application of at least one of the Data Mining algorithms studied in this course, and cannot be performed e.g. just using SQL.
    2. Identify the Data Mining algorithm(s) that is (are) appropriate to answer your question.
    3. Preprocess the data to make it ready for the selected Data Mining algorithms.
    4. Execute the chosen Data Mining algorithms and present the obtained results. Given that you work on real data, it may happen that you cannot identify any patterns in the data.
    5. Interpret your results and be able to convince your examiner & fellow students that your results can be trusted (whether you have found patterns or you are claiming that there are no patterns in the data).
- You must submit your presentation slides latest on the Friday before your presentation.
- You will have 15 minutes (including questions) to present your work and your results during the time slot assigned to your group and indicated on the Student Portal.

DATA & TOOLS
You are free to choose any tools to perform your analysis. A recommended combination is MySQL or sqlite (if you need a lot of initial preprocessing power and you want to exploit your knowledge of SQL) and RapidMiner, or only RapidMiner if you do not need database methods, but you are free to choose other tools if you prefer, e.g., R, Weka, Orange, etc. You do not need any approval for this.

For your analysis, you can use a dataset of your choice and interest (that must be approved by your tutor) or one of the following three datasets (that do not need approval):

- A traditional dataset containing data from the 1990 U.S. Census, that can be downloaded from the following link: https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)
- A dataset obtained by monitoring an Online Social Network, with user posts, likes and a following/follower network. Please, notice that you *do not have to* use all the tables, but you can (and should) choose those suiting your analysis questions: for example, you can focus on the "entries" or "comments" tables. These can be obtained from the following link: https://drive.google.com/folderview?id=0B_D5tuT1vDQtckFGWkk1aTh5VlE&usp=sharing
- Data from the Global Health Observatory Data Repository. Please, notice that the data are spread through several tables, which will need to be integrated following the formulation of your data mining questions. The data are available from the following link: http://apps.who.int/gho/data/?theme=home

Independently of the chosen dataset, consider that **you will probably need to spend most of the project time understanding, retrieving and pre-processing the data**.

SUPPORT

This project tests your ability to independently design and execute a knowledge discovery process on real data – that is, not "academically" prepared to be simple to understand or where existing patterns have been "made ready for discovery".

Therefore, you should work independently on this project.

However, you can have multiple meetings with your tutor to check that you are on the right track and get feedback. You also need your tutor to check and approve your dataset if you use one that is not in the list above. **You are encouraged to use your own datasets**, but do not have to.