Labs Data Mining

Informatics Engineering degree

Barcelona School of Informatics

D. Conti [1,2], X. Angerri [1,2], S. Ramírez [1,2], L.Mandadapu [1,2], K. Gibert [1,2,3]

[1]Department of Statistics and Operations Research

[2] Knowledge Engineering and Machine Learning group at

Intelligent Data Science and Artificial Intelligence Research Center

[3] Institute for Cience and Technology of Sustainability

karina.gibert@upc.edu, sergi.ramirez@upc.edu

http://www.eio.upc.edu/homepages/karina, ----

Universitat Politècnica de Catalunya, Barcelona

Phases of First Practical Work

- Form the working teams
- Choose a working topic and find data
- Make a working plan
  - Schedulling
  - Tasks assignments
  - Prior risk analysis
- Development
  - Data Cleaning and descriptive analysis
  - PCA
  - Clustering + Profiling
- Conclusions
- Comparative analysis
- Critical Analysi of working plan, management and incidences
- CAREFULL presentations preparation

Working teams

- Form working group in the class (Gender, foreign people,…)
- Present group proposal: First week (orange deadline)
- Lecturer might refine proposal

Assume importance of working well in a professional team leaving appart personal issues

Working Teams and practical work

- See additional working team resources in the website
- See instructions for deliveries in the website

Working teams: Soft skills to be considered

- Form working groups
- Planning and incident management
- Reasoning
- Synthesis and clarity in the written report
- Communication skills (oral presentation)
- Knowledge integration (comparative analysis)
- Performance throughout the course in follow-up sessions

Care: Cross evaluation activities

Select working topic

- Determine one (or more) topics of interest for the whole group (avoid topics repeated in previous courses. See web)

Mantain work motivation

- Search data providing information about those topics
  - Suggestion: download open data available in Internet (unless you have direct access to other data)
  - Open source links list available in the website of the course
  - Ensure to download data readable in R (or similar)

Database features

- Ensure that the formats are compatible with R
- Rectangular database structure
  - Individuals in rows
  - Variables in columns
- BD with 2000-5000 rows
  - If you have more, make a selection of a profile of interest or random
- At least 7 numeric variables
- At least 7 categorical variables
- Propose the database to teachers for their approval: date D1

Make a workplan (date D2)

- Identify big tasks (include documentation, etc)
- Sequenciation of tasks (carefully think about precedences between tasks, use the scheduling of the course as a reference)
- Assign human resources to each task (who does what and when) (Ensure viability, i.e. , everyone has a balanced load and no precedences that could collapse the group). In table format.
- Deserve some time before delivery for unforeseen incidents. Plan to finish sooner.
- Deserve time to prepare presentations as well
- Workplan format:
  - Gant's Diagram + assignment table + risk table *(look workingTeam resources)*

Materials

- Materials course

  https://www-eio.upc.edu/~karina/datamining/

  Username: dataminingGEI

- Bibliography (see website)
- OpenData Links (see website)
- List of previous practical works (see website)

- Software. R, RapidMiner, Weka, Knime …..