Statistical Programming course

Master on Visual Tools to Empower Citizens

Universitat de Girona

2020-2021

Karina Gibert and Marc Comas

Long Term practical work

Elaboration of a complete data science process to extract knowledge from a real data set

Note: This assignment is part of the cross-module assignment that you will have to develop in teams this term
In the next days you will receive the complete document with the part of the assignment correspongint to the other two involved courses

Working asset and delivery

You will work in groups of 3 or 4 persons and apply the contents of the course to the reference dataset. Each group works independently.
Note: The working groups will be determined by the Academic Board of the Master. So, you wil receive information on that in an ongoing additional document

A report will be delivered by Friday January 8th 2021 at 12:00 o'c.
A public presentation of each work will be done on Friday January 8th, 2021 at 17:00 o'c.
Common discussion of the different approaches followed by the working teams on the same dataset will be held as the final part of your training program in this course.
In the following, details on the different aspects of this practical work are detailed.

Reference DATASET

For this practical work you will work with COVID-19 data maintained by *Our World in Data*. Data consists of a different time series for COVID-19 tests, new cases, deaths, and other related information. A codebook for available information can be found as a CSV file at GitHub repository available at https://github.com/owid/covid-19-data/tree/master/public/data

A second dataset is available with information of different (not all) countries (country-info.xlsx). The column location can be used as a primary key to join with COVID-19 dataset. More information for the information related to each country can be find in country-info-description.docx.

The work plan is the following

a) Working data matrix preparation
b) Initial Descriptive analysis
c) Preprocessing
d) Definite descriptive analysis
e) Modelling
f) Conclusions

Working data matrix preparation

1. Decide a level of granularity of your analysis (weekly data, monthly data, quarterly, etc) and build the corresponding aggregate dataset).
2. Find additional QUALITATIVE information on countries that MIGHT be relevant to the COVID-19 crisis (we provide a possible additional dataset called country-info with its corresponding metainformation to be used, but if you like you can search on World Bank, WHO or other open datasets, provided that you find a minium of 5 additional qualitative variables for countries).
3. Enrich your aggregated COVID19 dataset with the additional information on countries.

Preprocessing

1. Start with a complete descriptive analysis (report to a Word document where you can select relevant results and add comments as a chapter of your final report).
2. Data Cleaning according to goals. Detail and report preprocessing steps.
3. REPORT DECISIONS.
4. Repeat descriptive analysis for eventual clean and/or new variables.

Modelling and post-processing

The work consist in practicing a real application of several data mining methods to your dataset.
You are required to use at least 4 different data mining methods, one of each different main families of methods described in DMMCM map presented in class (Profiling methods, Associative, Discriminant and Predictive).
To do that,

1. Imagine 4 scenarios in which a certain question (of different characteristics) requires your data (or part of it) to be answered, thus fitting with the application of the 4 main families of methods.
2. Examples (which COVID-19 characteristics are different in the several countries ? total_deads depend on?.... which countries are behaving similar? ….).
3. For each scenario, choose a suitable DM method.
4. For each DM used.
5. Prove that technical hypothesis hold on target data set.

a. Brief description of method.
b. Results of applying the method to your data.
c. Prove validation of results.

6. Describe the process of converting DM results into useful and understandable results.
7. Show results.

Data Science workflow followed

Represent all steps of your entire process in a complete workflow (visualize graphically).

Reporting (written report by delivery date with a synthesis of the following aspects)

(To be determined which parts go in the written report and which in the slides)

1. Problem description with eventual contextualization of domain field.
2. Data file source and contents specification, including origin of additional qualitative variables for countries
3. Flowchart with the complete DM process followed.
4. Description of software tools used in every step and technical details.
   a. Preliminary descriptive analysis of variables.
   b. Detailed preprocessing steps with results.
   c. Descriptive analysis of clean data.
5. For each data mining method
   a. Question to be answered.
   b. Prove that technical assumptions holds on data set.
   c. Brief description of methods.
   d. Results.
   e. Prove validity of results.
   f. Describe post-processing if any
   g. Interpretation of final results.
6. Conclusions of the study.
7. Future work not solved, possible improvements.
8. Annex with scripts when used.

Deliverable (to be updated accordingly to previous section)

A Folder or zip file containing:

a) Report specified in Reporting section (in PDF).
b) Presentation in PPT (or PDF).
c) The aggregated dataset with additional qualitative variables before preprocessing
d) Preprocessed datasets.
e) Bibliographic references used and corresponding PDF.
f) Font code of the macros used in the different pieces of software used.