UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Estadística
i Investigació Operativa

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Intelligent Data Science and
Artificial Intelligence Research Center

Data Mining
Bachelor Degree on Informatics Engineering
curs 2020-2021

Professorat: Karina Gibert Oliveras (coordinadora), Xavi Angerri, Dave Rodríguez

D1. Definition and projects assignment. Delivery by: September, 23th 2020 (beginning of class)

Every group must present a one page report with the following information:

1. Name of all group components by alphabetical order (sort by Family name)
2. Data source including the url or urls involved
3. One paragraph explaining the process to get your data
   (basic download, more sophisticaded processes when used). It is possible to enlarge your data base with additional variables comming from other sources if you like, but do not invest too much time on that. If it is the case, provide all urls involved in your dataset.
4. One paragraph explaining what data are about
5. Basic structure of data matrix: One paragraph with:
   a. nr of records (better if it is bigger than 500, if you are working with countries in the world or other situations, this might be reconsidered)
   b. nr of variables
   c. nr of numerical variables (minimum of 7 numerical variables)
   d. nr of binary variables (minimum of 2 binary variables)
   e. nr of qualitative variables (minimum of 5 categorical variables)
   f. number and % of missing data per each variable
   g. % of missing data in the whole data matrix.

D2. Project kick-off Delivery by: September, 23th 2020, midnight

6. Once the groups consolidated and approved by the lecturer, the leader of the working team must send an email to to karina.gibert@upc.edu, xavier.angerri@upc.edu, david.rodriguez.segado@upc.edu including CC with mails of ALL members of the working team in the CC. The subject of the email must be:
   "*DM IEdegree WT <number>. <keyword of your practical work>*"
   (ex. DM IEdegree WT 3.videogames)
   The number is the working team number previously assigned to your group by the teacher. The keyword identifies the topic of your practical work.

   This will be used as the basic communication reference between the lecturer and the working team. From that point on, be sure that all questions mailed to the lecturer uses this complete list.

D3. Project development Delivery by: September, 30th 2020

7. Initial working plan (two pages): Including Gantt, division of tasks and brief risk contingency plan (see *Working team resources* slides in the website section entitled *Working team resources*)
8. Metadata file describing the selection of variables considered for the analysis (*see slide nr 8 in Data and Metadata slides from Theme 2. Data Preparation*)

9. Basic initial univariate descriptive statistics of raw variables (see provided R scripts for automatic descriptive analysis in *Lab Session2. Preprocessing (II)*. Markdow produces word files, the R script is interpreted. The scripts ARE ORIENTATIVE. Modify whenever required)
10. Enumerate which steps of the preprocessing process are used with your particular data (consider the steps proposed in slide 4 of *Preprocessing Slides in Theme 2. Data Preparation.* Remember that you have the whole complete information on preprocessing in the reference text from the link *Survey of preprocessing. Reference paper (MIE 2001)* from *complementary materials section of Theme 2. Data Preparation*)
11. List and justify all decisions taken for each preprocessing step
12. Additional descriptive statistics of variables that have been modified or created by preprocessing

Structure of the report to be delivered by October, 28th 2020

Part of the materials have already been made in previous deliveries. Just collect them and make a single final document

1. Cover with title of work, name of course, data and list of working team members by alphabetical order of family name
2. Index
3. Motivation of the work and general description of the problem to be analyzed (max one page)
4. Data Source presentation (repeating what was delivered in first part) (one paragraph)
5. Formal description of Data structure and metadata
    - **a.** What do rows of data matrix contain? (one paragraph)
    - **b.** Metadata Table
    - **c.** Final scope of the study with inclusion and exclusion criteria for both rows and columns (max half a page)
6. Complete Data Mining process performed (one page, including a workflow).
7. Detailed description of Preprocessing and data preparation. Please be sure to justify all decisions made.
8. Basic statistical descriptive analysis
    - **a.** Univariate for all the variables included in the study (half a page per variable)
    - **b.** Bivariate when relevant (half a page per pair of variables)
    - **c.** When required, please include descriptives before and after preprocessing
    - **d.** Conclude the section with one paragraph describing how is your data
9. PCA analysis for numerical variables:
    - a. Scree plot. Specify how many principal components are selected
    - b. Factorial map visualisation: For each factorial map provide. Be sure you use a single landscape pager for each single map in order to guarantee visibility of materials to the readers
        - i. Individuals projections
        - ii. Common projection of numerical variables and modalities of qualitative variables (take care to use correct color codes as explained along the course)
        - iii. Interpretation of relationships among variables observed. When possible, interpret the latent variable associated with the principal axis
        - iv. Conclusions

    Note: All factorial maps must be placed in a single landscape page that makes it visible
10. Hierarchical Clustering on original data:
    - a. Precise description of the data used (which variables have not been included in the analysis, if any, whenever you are using a CURE strategy, provide details about eventual sampling performed on data, etc)
    - b. Clustering method used, metrics and aggregation criteria used (Ward's method is recommended embedded or not in a CURE strategy whenever scaling to big data is required, for messy data Gower dissimilarity coefficient to the square is recommended)
    - c. Resulting Dendrogram (of the total dataset or the sample). USE A SINGLE PAGE for it.
    - d. Discuss about how to get the final number of clusters
    - e. Table with a description of the clusters size
11. Profiling of clusters: Use class variable as a response variable to analyze conditional distributions of variables to clusters and eventual statistical tests to assess which variables are significant in each cluster. Detect commonalities of each cluster and differences between clusters. What is intrinsic of each cluster? What distinguishes clusters among them?
    - f. Profiling graphs, CPGs, multiple boxplots, bivariate barplots, descriptive by groups, etc...

   g. For selected relevant variables, you can also add specific profiling tests to complete clusters interpretation
   h. Synthesize the result of the classes' interpretation process into a set of templates characterizing the clusters, one template per cluster
12. Global discussion and general conclusions of the whole work. Analyze coincidences and divergences between ACP, AMC, Clustering

13. Working plan, including (please be sure you include the working plan at the end of the document, not at the beginning)
   i. Initial and final Gantt
   j. Final tasks assignment grid
   k. Critial discussion about deviances of final scheduling with respect to the originally designed one and discussion about risks avoided by the initial contention plan and unexpected risks appeared during project.
14. R Scripts (only if they have not been embedded along the explanations of the work in previous chapters)

Structure of the PPT for 15min oral presentation: The structure of the ppt is the following

1. Slide with title, name of all group components, delivery date
2. Slide with outline of talk
3. Slide with topics addressed, goals of the work and urls from data sources, overview of BD structure and variables analyzed
4. Slide with the Data mining process schema
5. Slide with the descriptive analysis of one numerical variable and one qualitative variable
6. Slide synthesizing univariate descriptive analysis
7. Slide with additional descriptive analysis issues when relevant
8. Slide describing preprocessing steps (if required add additional slides for any specific aspect to be commented)
9. Slide with PCA specificacions, screeplot
10. Slide wtith first factorial plane for PCA (eventually additional slides for othe planes retained). Lack of visibility of map penalizes.
11. Slide with conlusions of PCA
12. Slide describing the clustering process followed and resulting dendrogramm
13. Slide describing which tools of class interpretation you have been used
14. Slide with CPG or eventual profiling graphs or numerical information about clusters to be highlighted (whenever possible, synthesize important graphics in a single slide... evenctually you can add some extra slide)
15. Slide with final class profiling (synthesis with description of class characteristics)
16. Slide with comparison of conclusions between PCA and clustering
17. Slide with conclusions
18. Slide with original and final scheduling

Discussion session under oral presentation on ppt : October 28th 2020

D4. Material to be presented by final delivery October 28th 2020

Printed version of the report and ppt ON PAPER (not required in courses with online lab sessions)
Bring TWO USB-pen or similar conveniently labelled with name of the goup (not required in courses with online lab sessions)
One of the USB and the Printed report and pdf will be delivered just before presentation.
In the meanwhile the lecturer copies the USB in his machine, ***the students will use the second USB*** in parallel to copy their presentation into the room PC. (not required in courses with online lab sessions)

The USB given to the lecturer will be returned to the students immediately after getting your delivery materials. (not required in courses with online lab sessions)

The USB has to contain a Folder with the following contents (in courses with online lab sessions the folder will be directly uploaded to the virtual and precise indications will be provided by lecturers in due time):

19. Report in pdf and font files (doc o tex).
20. Presentation in PPT and pdf.
21. Folder containing the original dataset used
22. Subfolder with intermediate data files used during the development of the work (at least with cleaned data)
23. Folder containing bibliographic references available in digital support plus the reference files when possible

24. Font code of the R scripts used and macros from other softwares or programming languages used
25. README.txt specifying the structure and contents of the CD, including comments on the contents of the different files

Deficiencies in presentation protocols will be penalized.

If someone cannot attend to the presentation days, please contact the lecturer by mail in advance to agree on a solution (send an email to karina.gibert@upc.edu with copy to his laboratory instructor). Non notified absences to presentation days will penalize.

PS: If the work merits it, it will be considered as a research report in the EIO department and eventually for further publication in a journal.