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

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

MULTIVARIATE ANALYSIS
Official Master in Data Science
Universitat Politècnica de Catalunya

Lecturers: Karina Gibert Oliveras (Coordination), Dante Conti.

Check delivery dates in the slides of introduction to lab sessions and the ORIENTATIVE Scheduling.

D1. Definition and projects assignment. (see delivery date in course schedule)
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 sophisticated processes when used). It is possible to enlarge your database with additional variables coming from other sources if you like, but do not invest too much time on that. Deliver dataset on time. 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) Suggestion, at least 2000-5000 records.
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 (see delivery date in course schedule)
6. Once the groups consolidated and approved by the lecturer, the leader of the working team must send an email to karina.gibert@upc.edu, dante.conti@upc.edu, including **mails of ALL members of the working team in the CC.** The subject of the email must be:

"*MVA DSdegree WT <number>. <keyword of your practical work, provided by the lecturer>*"
(ex. MVA DSdegree WT3.videogames)
The number is the working team number previously assigned to your group by the instructor. The keyword identifies the topic of your practical work.
This will be used as the basic communication reference between the instructor and the working team. From that point on, be sure that all questions mailed to the lecturer **uses this complete list.**
7. Initial working plan (two pages): Including Gantt, division of tasks (assignment grid) 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*)

D3. Project development (and partial delivery)

9. Cover with title of work, name of course, data and list of working team's members by alphabetical order of family
10. Index
11. Motivation of the work and general description of the problem to be analyzed (max one page)
12. Data Source presentation (repeating what was delivered in first part, D1) (one paragraph)
13. 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)
14. List and justify all decisions taken for each preprocessing step
15. Basic initial univariate descriptive statistics of preprocessed variables. Compare raw and preprocessed variable distribution when relevant
16. Half page describing the dataset according to the main conclusions of the univariate and bivariate statistics
17. PCA analysis for numerical variables:
a. Scree plot. Specify how many principal components are selected
b. Factorial map visualization (must be placed in a single landscape page that makes it visible): For each factorial map, be sure you use a single landscape page for each single map in order to guarantee visibility of materials to the readers. Be sure all legends required are included: i. Individuals projections. ii. Common projection of numerical variables and modalities of qualitative variable (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
18. MCA: analysis of multiple qualitative variables
19. Multiple Factorial Analysis by combining all numerical and qualitative variables together.

In the D3 delivery provide a folder or zip file with pdf with the report, ppt, raw datasets, pre-processed datasets, and R scripts. Be sure the instructors will be able to open the files. Upload this folder at RACO-FIB (please, upload only one folder per team).

D4. Project development (second part) and final delivery

20. Association rules mining analysis:
a. Use the ECLAT and Apriori methods to identify the frequent itemsets and the association rules extracted from the database
b. Explain the results including the support, confidence and lift of the induced rules
c. Explain the top 20 rules sorted by decreasing confidence
21. Include as first part of the report, the entire materials generated in D1, D2 and D3, except Gantt and task distribution grid that will be placed at the end of the report
22. Hierarchical Clustering on original data:
a. Precise description of the data used (which variables have not been included in the analysis, if any)
b. Clustering method used: metrics and aggregation criteria used (Ward's method is recommended, for messy data Gower dissimilarity coefficient to the square is recommended)
c. Resulting Dendrogram. 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

22b. Model-based Clustering.
23. Profiling of clusters: Use the 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. Find commonalities of each cluster and differences between clusters. What is intrinsic of every cluster? What distinguishes clusters among them? a. Profiling graphs, CPGs, multiple boxplots, bivariate bar plots, descriptive by groups, etc...
b. For selected relevant variables, you can also add specific profiling tests to complete clusters interpretation
c. Synthesize the result of the classes' interpretation process into a set of templates characterizing the clusters, one template per cluster
24. Decisions tree (CART-Classification And Regression Trees)
a. Determine which input parameters have you used to build the model. Determine the response variable and all explanatory variables
b. Include the tree plot results
c. Explain the results obtained from the tree plot
d. Validate the model using cross validation and other criteria, like AIC
e. Explain the predictive power of the resulting model
f. Write conclusions with this method and compare the method among them
25. Discriminant analysis (LDA)
26. Discussion and conclusions
27. Planned Gantt and task distribution grid. Final real executed Gantt, tasks assignment grid and real risks addressed along the project

Structure of the report to be delivered by D4
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 (why did you selected those data? ….) and general description of the problem to be analyzed what is data about?) (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 (according to structure presented in class)
c. Final scope of the study with inclusion and exclusion criteria for both rows and columns (max half a page)
6. Detailed description of Preprocessing and data preparation. Please be sure to justify all decisions made.
7. 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 descriptive before and after preprocessing
d. Conclude the section with half a page describing how is your data
8. Synthesize the descriptive analysis of sample data (80% size of original data)
9. For each predictive model a. Variables included in the model
b. Goodness of fit indicators of the initial model
c. After simplification, equation of the resulting best model
d. Validation of the model
10. Slide describing the clustering process followed and resulting dendrogram
11. Slide describing which tools of class interpretation you used
12. Slide with CPG or eventual profiling graphs or numerical information about clusters to be highlighted (whenever possible, synthesize important graphics in a single slide... eventually you can add some extra slide)
13. Slide with final class profiling (synthesis with description of class characteristics)
14. Slide with PCA specifications, scree plot

a. Slide with first factorial plane for PCA (eventually additional slides for other planes retained). Lack of visibility of map penalizes.
b. Slide with conclusions of PCA
c. Same scheme for the ACM results
15. Discussion, conclusions, comparison of results among several methods. What have you learned from your data?
16. Working plan, including (please be sure you include the working plan at the end of the document, not at the beginning) a. Initial and final Gantt
b. Final tasks assignment grid
c. Critical 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.
17. 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
CARE: Legibility of slides is taken into consideration in the final marks
Discussion session under oral presentation on ppt :the day after D4
D4. Material to be presented by final delivery
Printed version of the report and ppt ON PAPER (not required in courses with online lab sessions)
Bring **ONE USB-pen** or similar conveniently labelled with name of the group (not required in courses with online lab sessions) and the Printed report and pdf will be delivered to lecturers just before presentation.
While the lecturer opens the materials delivered in the raco in his machine, ***the students will use the USB*** in parallel to copy their presentation into the room PC. (do not assume that the PC will have free access to your drive folders or similar) (not required in courses with online lab sessions)
Prepare Folder with the following contents (to be uploaded in the virtual campus by D4 date):
1. Report in pdf and font files (doc o tex).
2. Presentation in PPT and pdf (up to 41 books).
3. Folder containing the original dataset used
4. Subfolder with intermediate data files used or created during the development of the work (at least with cleaned data)
5. Folder containing bibliographic references available in digital support plus the reference files when possible
6. Font code of the R scripts used and macros from other software or programming languages used
7. README.txt specifying the structure and contents of the folder, 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 / dante.conti@upc.edu). **Non notified absences to presentation days will be penalized.**