MULTIVARIATE STATISTICS (Bachelor's degree in Mathematics)
Final evaluable practice

In this final project, the aim is for students to prepare a professional report in relation to a problem of interest based on a data set of their particular choice. With this objective, an exploratory analysis will be carried out and decisions will be made based on what is learned from the data. This analysis will be carried out in two phases:

1. **Univariate exploratory analysis**

   In this phase it is recommended to carry out a preliminary exploratory analysis of the data contained in the considered data set. To do this, apply the different numerical and graphical techniques used in class. At first, it will focus on the **analysis of each of the variables independently** without yet looking for possible interactions between them (**univariate analysis**). It is recommended to perform **numerical and graphical analysis** of each variable, to detect:

   a) **Recodings or data groupings** if you consider it appropriate by viewing the structure of the data file.

   b) **Missing values** by loading and viewing data. The following steps must be carried out:

      i.) For each variable, identify the **% missing values**.

      ii.) Of the variables that have more than 5% missing values **analyze the random pattern** or not. To do this, study homogeneity according to groups (NA and non-NA) with other variables. If they are continuous, with a Student test, if they are qualitative or discrete with a Chi-square independence test, etc.
      (*Investigate functions for the contrast of means such as t.test(), etc. from the R language.*)

      In the case of **homogeneity** the **pattern is random** and, in this case, it is chosen to **replace** the NA with the mean or median, depending on whether it is quantitative or qualitative variable.
      (Use the source code of the class practices.)

      In the case where **there is no homogeneity**, the **pattern is not random**. This would have to be discussed with the researcher who poses the problem under analysis because **they should neither be eliminated nor replaced**, but since in this case it is not feasible, **it is decided to act as in the case of a random pattern, warning of this fact in the final report**.

   c) **Classical numeric** descriptive analysis (measures of central tendency, dispersion, quantiles, symmetry, kurtosis, etc.)
   (*Use the source code of class practices.*)

   d) **Extreme values** (outliers) based on the numerical results of the previous section as well as graphical results (boxplots).
   (*Use the source code of the class practices.*)
   In the event that there are extreme values, the decision will be made **to eliminate them**, if the data file has enough records, **or replace them with the mean or median** depending on whether the variable is quantitative or qualitative.
   (Use the source code of the class practices.)

   e) Many statistical techniques cannot avoid the **assumption of normality**. In this sense, analyze this assumption for the different continuous variables of the database. To do this, you must try to justify or discard it graphically with normality graphs (qqplots, etc.)
   (*Investigate qqplot() of the R language.*)

f) Any other issue that is considered of interest for a good understanding of the data.

2. **Multivariate exploratory analysis**

   Secondly, the assumptions underlying the application of the different multivariate dimension reduction techniques, such as PCA or FA, will be checked before being applied. For this, it is requested:

   a) Check the assumptions of correlation between variables with the Bartlett test.
   (Use the source code of the class practices.)

   b) It is assumed that in the previous univariate analysis the outliers have been identified and treated; if not, they must be analyzed before applying dimension reduction techniques.
   (Use the source code of the class practices.)

   c) If the NA missing values have not been treated because they did not exceed 5% as indicated in the previous univariate analysis, at this point decisions must be made about them in order to use dimension reduction techniques.
   (Use the source code of the class practices.)

   d) At this point it is requested to carry out a study of the possibility of reducing the dimension through observable variables. It is convenient to choose the optimal number of principal components using the different graphical techniques introduced in class.
   (Use the source code of the class practices.)

   e) Likewise, it is requested to perform a dimension reduction using latent variables, previously choosing the optimal number of factors to consider.
   (Use the source code of the class practices.)

   f) Before constructing classification methods, analyze the multivariate normality of the data with the test proposed in discriminant analysis practical class (unit 5).
   (Use the source code of the class practices.)

   g) Next, a classifier will be built using a linear and quadratic discriminant analysis.
   (Use the source code of the class practices.)

   h) Finally, we will carry out a very basic validation of the classifiers obtained by graphically representing their respective confusion matrix, COR curve and different security and internal validity measures (sensitivity, specificity, positive predictive value and negative predictive value).
   (Use the source code of the class practices.)

   i) Additionally, a cluster analysis can be performed to confirm that the grouping of the response variable used in the classification models is appropriate. If there was no response variable for section g) above, this cluster analysis would have to be done first to define it and then use it in the classification.

INDICATIONS

1. **Use RMarkdown** to carry out the previous exploratory analysis to have an overview of the different outputs obtained with R. This code may have text that describes the opinions of the students based on them.

2. Prepare the **final report with a scientific text processor**, preferably LaTeX. Reports written in Word, Writer, etc. will also be accepted. You can also embed LaTeX in the RMarkdown document and present the pdf or html it generates.

3. The final report could include the following sections:

    - **Summary or abstract** of no more than 200 words putting the chosen problem in context, indicating what techniques have been applied and with what objective to end with a line or two that describes a final conclusion.
    - **Introduction** of no more than 400 words that slightly extends the previous summary. This section should end with a paragraph of two or three lines defining the objective of the work to be carried out.
    - **Materials and Methods**. This section could include a subsection, 'Materials', that briefly describes the database, reporting what the different variables store and providing a table with the basic descriptive statistics (mean and standard deviation for quantitative variables; % and totals for categorical variables). The second subsection, 'Statistical methods', of no more than 400 words will indicate the different statistical techniques used. It is emphasized that in this subsection **the techniques are indicated, they are not explained** nor are master classes given.
    - **Results**. This section should show a summary of the most notable results obtained in the development of this practice. It must be an objective presentation of results without interpretation in the context of the problem.
    - **Discussion** of no more than 600 words that interprets the results obtained. This section should begin by remembering what the objective or objectives were announced in the final paragraph of the introduction, and then discuss which ones have been achieved and how, based on the results.
    - **Conclusion** of no more than 250 words that summarizes what has been achieved, talks about the strengths of the work carried out, reports on its limitations and makes proposals for improvement or indicates other paths that could be followed or opened in the analyzed context.

4. **DELIVERY METHOD**

    The students will upload four files to the task created for this practice on the PRADO platform: the **RMarkdown source code**, the **html output obtained with RMarkdown**, the **LaTeX source file** with the final report (if another word processor is chosen, the editable file created by the student will be requested) and a **pdf file with the final report** compiled in the case of LaTeX, or saved as a .pdf file from any other text editor used.

5. **DELIVERY DEADLINE**

    Until the day before the exam of the ordinary assessment session for the subject (**January 17, 2024**).