Assignment

September 14, 2023

0.0.1

Máster en Big Data. Tecnología y Analítica Avanzada (MBD).

Introducción al Análisis Estadístico con Lenguajes de Programación para Machine Learning (IAELPML). 2023-2024.

Assignment

Due Date: 2023-09-21

1 Enter your name here:

2 Preliminaries

- You are expected to submit your answers using the Jupyter notebook format. You can use this notebook as a template, but make sure to add your personal information (name and email) at the top of the notebook
- If you want to include additional files (data, scripts, figures, etc.) create a zip file containing all the files and email that file to us.

3 Questions

- We will be using the data in the file `frami.csv`, that you can downoload using this link: frami.csv
  Be careful: it is similar but not exactly equal to the Framingham data set we have seen in class.
- Load the data set into a pandas DataFrame called `frami`.
- Gather some basic information about the data set: how many observations are there, what are the variables and their types,…
- Check for missing data. If there are any, first obtain the row numbers and variables in which they appear. Then remove those rows from the table. Make sure to rename the resulting table (after the removal) again as `frami` and use this table to answer the following questions.
- If there are factor variables in the table convert them to the appropriate Python type.
- Perform a detailed exploratory analysis of a selection of variables of your choice. This analysis should cover all the types of variables in this table. That is, you should at least study one

variable for each type of variable present in the table (but you don't need to study all the variables). The analysis should contain:

  - For quantitative variables (continuous or discrete):
    * Basic numeric summary (mean, median, quartiles, sd).
    * Graphics (the right ones for that type of data, possibly more than one graph per variable).
  - For factors :
    * Frecuency tables (absolute and relative).
    * Graphics (bar plot).

- Create a new column called `ageGroup` dividing (binning) age in these three levels:

  `(0,40], (40,55], (55,100]`

- Find out how many observations are there for each level of `ageGroup`. Now, using only observations corresponding to women, , what is the mean of cholesterol level `totChol` for each of those age groups?
- Let us analyze the *possible* relation between gender (variable `male`) and `currentSmoker`). Do a graphical analysis of that relation and compute the 2x2 contingency table (with marginal probabilities)) for these two factors.
- Use Python to compute the following probabilities.
  - What is the probability that a current smoker (selected at random from the table) be a woman?
  - What is the probability that a man be a current smoker?
  - What is the probability that a randomly chosen person be a smoker? And the probability of being simultaneously man and smoker?
- If we choose a sample of 10 persons out of this table (random sampling with replacement), what is the probability that exactly 4 of them are smokers? What is the probability that at most 4 of the 10 persons are smokers?
- Run a simulation using Python to confirm your results of the previous question. That is, take $N = 100000$ samples of size 10 from the table and use them to obtain an estimate of the requested probabilities. Use the random number generator features to make your simulation code reproducible; that is, when we run your code we should get exactly the same results.