Assignment 2: Data Preprocessing

DSA 8401: Applied Machine Learning

Master in Data Science and Analytics

Strathmore University

@iLabAfrica Centre

1 Objective

The objective of this assignment is the students perform a preprocessing process on a dataset from the MIMIC-III database, which contains information on in-hospital mortality from the monitoring of patients in the Intensive Care Unit (ICU) for 48 hours..

2 Data

In the virtual classroom you will find the data in a file named: ihm_48_hours.csv

3 Instructions

Following the guidelines provided in the class, and using mainly PANDAS, students are required to prepare a notebook that describes step by step the preprocessing process necessary to obtain a dataset without outliers or incomplete variables. The report must include evidence in the form of plots and tables that show the following information:

- Enumeration of Variables: List the categorical variables, indicating which are binary and which are considered nominal. List the numerical variables present in the data set.
- Initial Statistical Description: Carry out a statistical description of the variables before carrying out the cleaning process.

- Identification of Outliers in Numerical Variables: Identify numerical variables that contain outliers. Support this detection with plots that illustrate the procedure used.
- Elimination of Records with Incomplete Values: Delete records that contain incomplete values in categorical variables. Document this process through screenshots of the applied filter and the resulting variables after the operation.
- Treatment of Outliers and Missing Values in Numerical Variables: Delete values considered outliers in numerical variables. Replace missing values with central tendency values.
- Post-Processing Statistical Description: Recalculate the statistical description of the variables. Verify that the data set does not contain outliers or missing values.
- Analysis of Advantages and Disadvantages: Perform an analysis that evaluates the advantages and disadvantages of the data preprocessing process carried out. Provide an evaluation of the results obtained.

4 Hand in

The students will upload in the eLearning platform the Jupyter notebook. The notebook will contain the python code, the plots and especially personal responses to the previous questions. If there is no personal comment the mark will be F. Marks C,B and A will depend on the variety and quality of the personal comments.