ASSIGNMENT 2 - POLITECNICO DI TORINO

DATA VISUALIZATION AND PREPARATION FOR REGRESSION

Dataset

Data to be used are in **A2_student-mat.csv** file. Original data may be found at this link. The target variable is named “G3” and represents the student final grade in the Math course.

Instructions

DATA LOADING

- load the dataset
- visualize the first rows of the dataset
- inspect the data types of the columns

DATA CLEANING

- check for the presence of null values and remove the rows containing them (if present)

DESCRIPTIVE STATISTICS and VISUALIZATIONS

- display the number of students
- display the number of features
- describe "G3" feature with its main statistics (mean, standard deviation, quartiles, etc.)
- plot "G3" empirical distribution
- plot "Age" empirical distribution
- plot "Age" distribution grouped by "Gender"
- plot a boxplot of "Age" vs "G3"
- plot a boxplot of "Gender" vs "G3"
- plot a boxplot of "Age" vs "G3" grouped by "Gender"
- count how many students live in Rural and Urban areas ("Address" feature)
- plot the estimated CONTINUOUS distributions of "G3" for students living in Urban and Rural areas on the same plot
- compute and display all correlations between features (optionally, find a suitable visualization)

MANIPULATIONS AND FEATURE ENGINEERING

- drop "G1" and "G2" features (perfectly correlated with "G3")
- Create feature "Social disadvantage" = "True" IF "Address" == "R" AND "famsize" == "GT3" AND "Pstatus" == "A" AND "internet" == "no" ELSE "False"
- Convert all categorical variables in one hot encoding
- Find correlations again and keep only the 8 features with highest correlation with "G3"

- In a 4x2 plot grid (4 rows, 2 columns) plot each retained feature against the target variable “G3” (using suitable plots)

REGRESSION

It is assumed that this section is carried out using the dataset resulting from the previous parts of the exercise.

- Split the dataset in two randomly sampled subsets: training set (80% of data) and test set (20% of data)
- Fit a linear regression model of the retained variables vs. “G3” using the training set
- Test the obtained model predicting the target variable “G3” for samples of the test set
- Evaluate results

Recommended libraries: Pandas, Seaborn/Matplotlib

Note: Don’t reinvent the wheel! Explore libraries documentation and exploit already defined functions.