**DAMO510, Module 05: Working with Categorical Variables**

**Objective**

* Apply train-test split before preprocessing
* Working with categorical variables and utilizing them in predictive tasks appropriately
* Getting insights about advantages and limitations of using categorical variables in predictive tasks.
* Gain hands-on experience with real-world data modeling using Python and optionally MS Excel (Data Analysis Toolpak).

**Dataset Overview**

The *student-por.csv* dataset is part of the Student Performance Data Set from the UCI Machine Learning Repository. It contains detailed information on 649 students from Portuguese secondary schools, focusing on their performance in the Portuguese language course. The dataset includes 30 attributes covering academic grades (G1, G2, G3), demographic details (age, gender, address), family background (parental education and jobs), and social factors (study time, extracurricular activities, alcohol consumption, and romantic relationships). The data was collected through school reports and questionnaires, and it's commonly used for classification and regression tasks to predict final grades. Notably, the final grade (*G3*) is highly correlated with earlier grades (*G1* and *G2*), which raises important considerations around data leakage when building predictive models.

The data dictionary is as follows:

| **Variable Name** | **Type** | **Description** |
| --- | --- | --- |
| *school* | Nominal | Student's school (*GP* = Gabriel Pereira, *MS* = Mousinho da Silveira) |
| *sex* | Nominal | Student's gender (*F* = female, *M* = male) |
| *age* | Integer | Student's age (15–22) |
| *address* | Nominal | Home address type (*U* = urban, *R* = rural) |
| *famsize* | Nominal | Family size (*LE3* = ≤3, *GT3* = >3) |
| *Pstatus* | Nominal | Parent's cohabitation status (*T* = living together, *A* = apart) |
| *Medu* | Ordinal | Mother's education (*0* = none to *4* = higher education) |
| *Fedu* | Ordinal | Father's education (*0* = none to *4* = higher education) |
| *Mjob* | Nominal | Mother's job (*teacher*, *health*, *services*, *at\_home*, *other*) |
| *Fjob* | Nominal | Father's job (same categories as *Mjob*) |
| *reason* | Nominal | Reason to choose this school (*course*, *reputation*, *home*, *other*) |
| *guardian* | Nominal | Student's guardian (*mother*, *father*, *other*) |
| *traveltime* | Ordinal | Home-to-school travel time (*1* = <15 min to *4* = >1 hour) |
| *studytime* | Ordinal | Weekly study time (*1* = <2 hrs to *4* = >10 hrs) |
| *failures* | Integer | Number of past class failures (0–3) |
| *schoolsup* | Binary | Extra educational support (*yes* or *no*) |
| *famsup* | Binary | Family educational support (*yes* or *no*) |
| *paid* | Binary | Extra paid classes (*yes* or *no*) |
| *activities* | Binary | Extracurricular activities (*yes* or *no*) |
| *nursery* | Binary | Attended nursery school (*yes* or *no*) |
| *higher* | Binary | Wants to take higher education (*yes* or *no*) |
| *internet* | Binary | Internet access at home (*yes* or *no*) |
| *romantic* | Binary | In a romantic relationship (*yes* or *no*) |
| *famrel* | Ordinal | Family relationship quality (*1* = very bad to *5* = excellent) |
| *freetime* | Ordinal | Free time after school (*1* = very low to *5* = very high) |
| *goout* | Ordinal | Going out with friends (*1* = very low to *5* = very high) |
| *Dalc* | Ordinal | Workday alcohol consumption (*1* = very low to *5* = very high) |
| *Walc* | Ordinal | Weekend alcohol consumption (*1* = very low to *5* = very high) |
| *health* | Ordinal | Current health status (*1* = very bad to *5* = very good) |
| *absences* | Integer | Number of school absences |
| *G1, G2, G3* | Integer | Grades in first, second, and final periods (0–20) |

**Instructions**

Before start discussing about the steps, it is highly recommended to practice naming conventions for the different steps of our analysis. It is an approach well adopted by a lot of data analysts and data scientists especially in the case of working on “small data” analytics. In this activity, if the data is going to be labeled as *df*, it is supposed to be called as *df\_a*. Thereafter, in the case of any dataset editing activities like data cleaning, data transformation and so on, instead of overwriting, the letter version will be updated such as *df\_b*, *df\_c*, and so on. At the end of each activity, getting report such as **shape** or **head**, could be helpful for any possible further debugging or development tasks. The instruction as below:

1. Load the dataset. **Be careful about the delimitation!**
2. Implement data cleaning (Module 2)
3. Define the target and features. To do this, *G3* can be determined as the target and the other two ones (*G1* and *G2*) to be dropped (**Why?!**).
4. Apply train-test split. Note that, up to this moment, **you haven’t done any encoding yet**.
5. Identify the types of features (Nominal, Ordinal and Numerical). For the nominal and ordinal variables, use **OneHotEncoding** and **OrdinalEncoding** respectively. Consequently, apply the results to both training and testing datasets. Leave the numerical as is; however, apply it on training and testing datasets. At the end, combine all of these three categories to obtain one training and one testing dataset. Note the shape of the new data frames.
6. Implement correlation analysis.
7. Implement linear regression and report *R*2, adjusted *R*2**\***, and RMSE.
8. Implement ANOVA test and report significance of the features.
9. Bonus task: Apply everything instructed on MS Excel and Data Analysis Toolpak for educational reasons.

**Reflection Questions**

1. Why did we drop *G1* and *G2* from the very beginning?
2. Why did we implement data splitting before data encoding?
3. What are the differences between applicability of **OneHotEncoding** and **OrdinalEncoding**? Coud we use other methods such as **LabelEncoding** or **get\_dummies**?
4. What is the advantage of using adjusted *R*2 instead of the original one? Is there anything strange about the value?
5. Are the results satisfactory as a reliable model? If not so, what are the possible reasons for it? How can we improve the performance?

Adjustted *R*2 = 1 – (1 – *R*2) \* (*n* – 1) / (*n* – *k* – 1)

*n* is the number of rows of the testing dataset

*k* is the number of predictors