Introduction to Data Science

21KDL

Lab03 - Regression

Deadline: **23h59 - 03/05/2023**
Submitting via email: hduc.lee@gmail.com

Tools and Language: Python / Jupyter Notebook / Google Colab.

PROBLEM

Let's use the dataset at this link to build models for predicting beer consumption with respect to data features that are described as follows:

- Tempetura Media: median of temperature in *Celsius* degree
- Tempetura Minama: min of temperature in *Celsius* degree
- Tempetura Maxima: max of temperature in *Celsius* degree
- Precipitacao: precipitation (lượng mưa) in *mm*
- Final de Semana: whether the weekend or not *(0 or 1)*
- Consumo de cerveja: the amount of beer consumed in a day (*liters*)

Requirements:

- Build 3 models with the provided dataset: Linear Regression, Lasso Regression, and Ridge Regression.
- Choose one of the following preprocessing techniques and apply it to the dataset: Standardization, or normalization. Notice that you can determine by yourself which features are important to build your models. *(OPTIONAL, BONUS)*
- Spit the dataset for training and testing phrases. The ratio of splitting is up to you.
- Illustrate the regression line against the dataset or relevant charts to visualize your work if possible and needed.
- Evaluate the models with metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R2-score. Then figure out the best model for the problem and *explain* your decision in terms of the *values* of those metrics. Please write your explanation with the mode of *markdown* in your notebook.

References:

- Scikit-learn - Regression

NOTICE:

- The lab can be done in **both Vietnamese and English**.
- Students are NOT required to re-implement the algorithms, i.e, be able to use the library.
- Please send me your work before the due date. You can send the Jupyter Notebook file (*.ipynb) or the Google Colab link. **DON'T FORGET** to share access to your Google Colab link. (Click to share button in the upper-right).

- Named **CORRECTLY** your notebook AND the subject of the submission email by the following pattern (same for Google Colab notebook's title): DS2023_Lab<LabID>_<StudentID>_<StudentName>.ipynb.
    - Example:
      Email subject: DS2023_Lab01_21280075_NguyenVanA
      File name: DS2023_Lab01_21280075_NguyenVanA.ipynb

- Inside the coding file, there should be a brief introduction (as in the example below).

```
"""
Introduction to Data Science
Programming Exercise: 01
Name: Nguyen Van A
Student ID: 21280075
"""
```

There is NO acceptance for cheating or copying.