HW4 - Data Modeling

CSC14119 - Introduction to Data Science

Lecturer: Dr. Le Ngoc Thanh, lnthanh@fit.hcmus.edu.vn
Teaching Assistant: Mr. Le Nhut Nam, lnnam@fit.hcmus.edu.vn

1 Objectives

This assignment is to continue the HW3 - Exploration Data Analysis (Part 2). In this assignment, we will continue to use the Wine dataset. By leveraging your understanding of this dataset and your preparation in previous HW, you have to build classification models to predict the wine quality. Also, fine-tune the hyperparameters and compare the evaluation metrics of various classification algorithms. The complexity arises because the dataset has fewer samples, & needs to be more balanced. Can you overcome these obstacles & build a good predictive model to classify them?

2 Dataset Introduction

This dataset is related to red variants of the Portuguese "Vinho Verde" wine. The dataset describes the amount of various chemicals present in wine and their effect on its quality. The datasets can be viewed as **classification** or **regression tasks**. The classes are ordered and not balanced (e.g., there are many more normal wines than excellent or poor ones). Your task is to predict the quality of wine using the given data. This data frame contains the following columns:

- 1 - fixed acidity
- 2 - volatile acidity
- 3 - citric acid
- 4 - residual sugar
- 5 - chlorides
- 6 - free sulfur dioxide
- 7 - total sulfur dioxide

- 8 - density
- 9 - pH
- 10 - sulphates
- 11 - alcohol

And the Output variable (based on sensory data):

- 12 - quality (score between 0 and 10)

3 Notes and Constraints

List of constraints when doing this lab:

- Work without a Jupyter Notebook will not be graded.
- Reference sources (if any) need to be fully recorded in the report in the References section. Note that it is necessary to distinguish between referencing and plagiarism.
- Individuals or groups that commit cheating and dishonesty will receive 0 points in the course.
- Please name the Jupuyter Notebook as HW04.ipynb. After that, compress all your work into **one file** and name it MSSV.zip. If the size is > 20MB, upload it to an external storage service such as Google Drive or OneDrive, then submit the link. Lastly, please keep the link public for at least 2 years.

4 Limitations

- The exercises are designed to be completed within a basic Python programming environment.
- You are welcome to employ foundational libraries such as **NumPy**, **Pandas**, **Seaborn**, and **Matplotlib** for your tasks. You can use **Scikit-lear**n to build a prediction model. Should you wish to explore additional libraries, please consult with your instructors before hand to gain their approval.

5 Evaluation Criteria

Your assignment will be evaluated based on the following criteria:

Criteria	Mark
Data preparation (Splitting train, valid, test)	5%
Building predicting models (It should be more than two models for comparision)	50%
Model evaluations	20%
Conclusion	10%
The notebook presents a logical and clear layout and format.	15%
There is analysis, visualization with novel charts, and drawing of useful information.	5%
Overall comprehension of the submitted source code.	5%
Total	**110%**