DSC3006 Assignment 1: Kaggle InClass Competition

Goal

The goal of assignment 1 is to practice data preprocessing and classification through a Kaggle InClass Competition. You are expected to understand how Kaggle works and how you can improve your classification model's performance.

Task

You are provided with a classification dataset and your task is to build a series of models with the goal of improving the performance. You can use any data preprocessing technique and classification method.

Data description

A detailed description is available at
https://archive.ics.uci.edu/ml/datasets/student+performance
The dataset used in the assignment 1 is a slightly modified version with 30 features and 1 categorical target variable. The goal is to use the 30 features and classify each student into one of the FIVE categories.

How

1) Go to https://www.kaggle.com/t/9f9472665a4747cf8f15d1dd9cea1261 and create an account if you don't have.
2) Go to Data tab and download data files.
   - **X_train.csv**: 264 samples, 30 features (Id should not be counted as a feature)
   - **y_train.csv**: 264 samples, 1 target (from 1 to 5, each number represents a category)
   - **X_test.csv**: 131 samples, 30 features (the dataset you test your model)

- **sample_submission.csv**: This is a sample submission file and when you submit your classification result for **X_test**, your final submission file should have the same format. It is a csv file with two columns **Id** and **Category**. Because it is a sample submission file, it has only 9 samples. The final submission file should have **131** samples (the same as **X_test**) with two columns **Id** and **Category**. **Id** column in your submission file is from **X_test**, and **Category** column should include your precited results (i.e., 1 or 2 or 3 or 4, or 5). After you do the prediction, you should generate an output file that has the same format with **sample_sbmission.csv** and submit it to the Kaggle. The file name can be arbitrary.

3) After submitting the result file. you will be able to see the score. The evaluation method is simple classification accuracy.
4) Try to improve the score by testing different preprocessing and classification methods. You are allowed to submit up to 20 times a day.

Deliverable

One-page short summary on

- Your Kaggle account
- How many submissions have you tried to improve the performance?
- What methods have you tried?
- Did the methods improve the performance? Why or Why not?
- Please explain your best solution with the highest score (e.g., what classification method + how you preprocessed the data)
- What have you learned from the competition?

IMPORTANT

1) You don't need to use your real name for the Kaggle account because the goal of the competition is to compete with yourselves, not with your peers.
2) The assignment will not be graded based on the Kaggle score. The assignment will be evaluated based on your **one-page summary**. Please write it carefully so that I can evaluate your efforts.

3) As I already disclosed the data source, you can find the correct answers easily on the web. Any attempts to artificially make submission files using correct answers will be regarded as PLAGIARISM.
4) If you have questions on reading the dataset and generating the submission files you can ask me or your peers. However, solutions should be your own.