CSED 226
Introduction to Data Analysis
Competition: Clustering

Overview

- The goal is to improve your model’s performance **as much as possible.**

- You can improve your model’s performance through **data preprocessing, model selection, and hyperparameter tuning**.

- You can submit your model’s results to Kaggle and view your scores in real-time on the leaderboards.
  - You can only submit a **maximum of 20 per day.**

Rules

- In this competition, you can build a model using any clustering method.

- You must participate as an individual team in this competition.

Evaluation

- Submissions are evaluated on the Adjusted Rand Index between **the ground truth cluster labels of the data** and **your predicted cluster labels (assignment)**.
    - Please refer to the link.

- **You are not given the number of ground truth clusters or any training labels.**
    - This is a completely unsupervised problem.
    - We recommend cluster indices that start with 0. (e.g., 0, 1, 2, ...)

Dataset

- **Health Condition levels Dataset**
  - You are given data with features such as demographic features and lifestyle features to cluster people’s health condition levels.

Dataset Description

Feature Name	Feature Type
Lifestyle_feature1	Categorical
Lifestyle_feature2	Numerical
Lifestyle_feature3	Categorical
Liefstyle_feature4	Numerical
Height	Numerical
Smoker	Categorical
Medical_history	Categorical
Weight	Numerical
Age	Numerical
Sex	Categorical
DFC	Numerical
SC	Categorical
EF	Numerical
FI	Numerical
AC	Categorical
EDUT	Numerical
HSM	Categorical

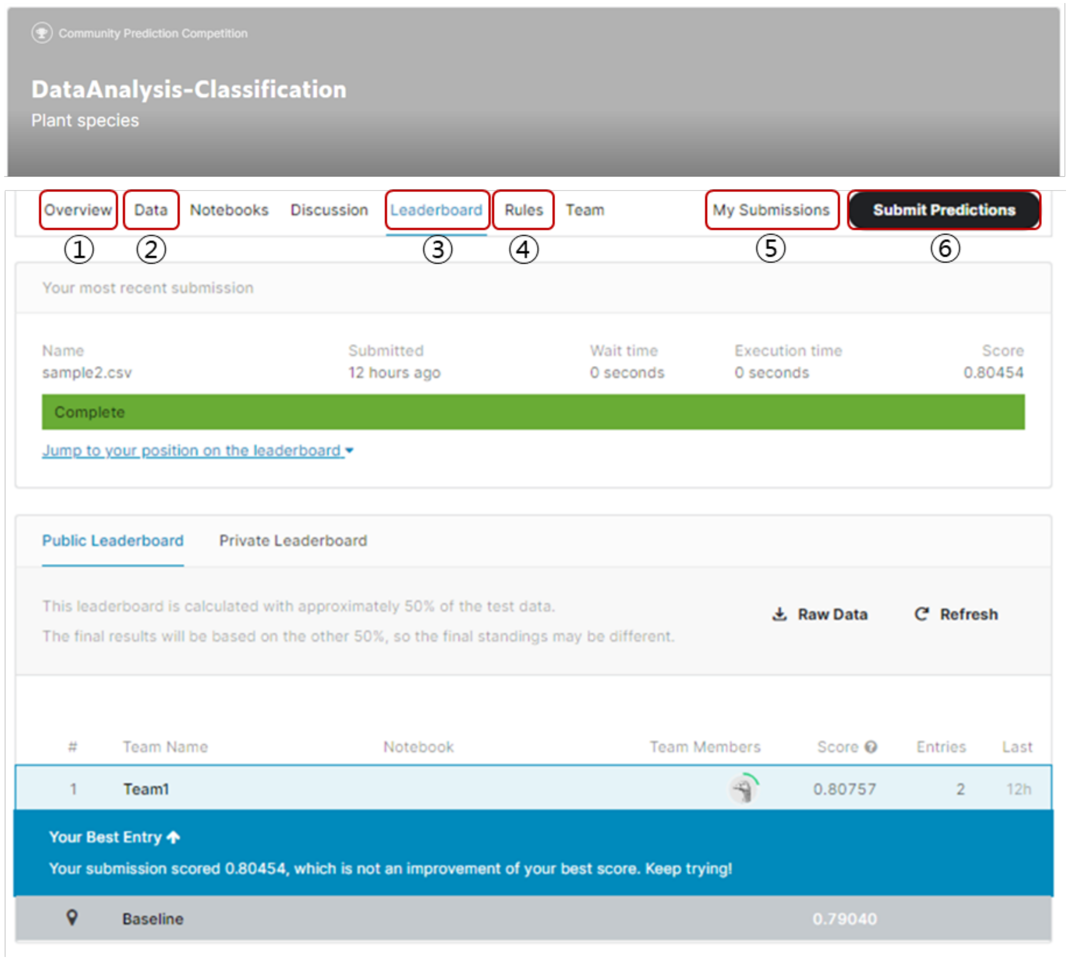
Dataset Description

- **Abbreviations**
  - Dietary Fiber Consumption (DFC)
  - Snack Consumption (SC)
  - Exercise Frequency (EF)
  - Fluid Intake (FI)
  - Alcohol Consumption (AC)
  - Electronic Device Usage Time (EDUT)
  - Health Signal Monitoring (HSM)

Competition

- You will enter the competition individually through Kaggle.
- Kaggle Site: https://www.kaggle.com
- Competition URL: https://www.kaggle.com/t/1459d64d96ce491e9b7aa42c40b6fa66

Competition



Community Prediction Competition

DataAnalysis-Classification

Plant species

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
sample2.csv	12 hours ago	0 seconds	0 seconds	0.80454

Complete

Jump to your position on the leaderboard ▾

Public Leaderboard   Private Leaderboard

This leaderboard is calculated with approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Team1			0.80757	2	12h

Your Best Entry ↑
Your submission scored 0.80454, which is not an improvement of your best score. Keep trying!

	Baseline			0.79040		

Kaggle-site

① Overview: Description of the task

② Data: Files to be used in the competition

③ Leaderboard: Current score updated in real time

④ Rules: What you must follow in the competition

⑤ My submissions: Files you submitted

⑥ Submit Predictions: Submission