DDA4210 Mini-Project

- **The competition will start at 6:00 PM February 17, 2023 and end at 1:30PM February 27, 2023.** The deadline of the supplementary materials (report and code) is 11:59PM March 5, 2023.
- **Submission:** Turn in your prediction results on Kaggle (`https://www.kaggle.com/t/0b0a516a6d8520fdc5629f5e1dc2df65`). You can create your favourite team name, but it has to end with your *studentID*. For example, *A-Hardworking-Student-1155014367*. Turn in your **report and code** electronically via Blackboard. Be sure to submit your report and code as a **single compressed file.** Please name your file as *P1_studentID_name*.
- **Important: If your prediction submission does not contain your student ID, it will NOT be considered, even if it's shown up on the leaderboard!**
- No **late submissions** are allowed.
- **Project Guru : Mr. Zhengyang Tang**

  Guru Email: 222010059@link.cuhk.edu.cn

  Start early and come to TA office hours with your questions on the mini-project! The tutorials will be helpful too!
- **Collaboration policy:** You need to complete this mini-project independently and collaboration between students is **not** allowed.

Project Website

`https://www.kaggle.com/t/0b0a516a6d8520fdc5629f5e1dc2df65`

Problem Description

The COVID-19 pandemic has had a significant impact on the world, and in this mini-project, you will use machine learning tools to perform a binary classification task that aims to identify changes resulting from the pandemic.

Specifically, the stay-at-home restrictions and other policies have led to notable changes in the behaviors of students and staff on campus, such as fluctuations in arrival and departure times at workplace parking stations. To address this, we have prepared a dataset consisting of anonymized and aggregated EV charging sessions, with each data sample represented by a four-dimensional feature vector $x \in \mathbb{R}^4$ that includes the following information:

x_1: Arrival time (UTC)

x_2: Charging duration (hours)

x_3: A mysterious feature

x_4: Another mysterious feature

Your task is to build a binary classification machine learning model that can accurately determine whether a given data sample was created during the pre-COVID-19 period (label=1) or the post-COVID-19 period (label=0). We encourage you to explore and experiment with different approaches to achieve the best possible performance. Good luck!

Evaluation Metric

The evaluation metric is the *accuracy*, defined as

$$\text{Accuracy} := \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is the number of True Positives, TN is the number of True Negatives, FP is the number of False Positives, and FN is the number of False Negatives. There are two leaderboards – a public and a private. Each of them is generated by part of the testing data. The public leaderboard can be seen by everyone in the competition while the private one will remain unseen. The overall accuracy will be computed based on all testing data.

Kaggle Submission Rules

1. Starting from 10:00 AM on February 18, 2023, you will have a total of **5 submissions** available to use every day. It's important to use your submission quotas wisely, so make sure to consider your priorities before submitting.

2. To submit your prediction results, please prepare a CSV file and submit it through the project website. For more details on the required format and other submission guidelines, please refer to the project website.

3. Please note that results must be submitted individually and not as a group. Each participant is responsible for submitting their own results, so please make sure to submit your work separately from other students.

Grading Policy

The total grade for this mini-competition will be **15 points**, with **10 points** allocated for your accuracy ranking on the leaderboard and **5 points** for your report. Your report should be **one to two pages** (longer reports will not receive any credits) in length and provide a clear and concise description of your binary classification algorithm, including its key components and the rationale behind your design decisions. We encourage you to include the **pseudocode** of your algorithm in the report, as your algorithm and the implementation will be a significant factor in the overall grading of your report. Below is the scheme for your accuracy ranking on the final leaderboard (considering **both** public and private ones).

1. As a reward for outstanding performance, the students with **top 3** accuracy scores in the class will receive **1 bonus point**;

2. In addition to the top 3 students, the top 10% students with accuracy scores will get 10 points;

3. The students with accuracy scores between $10\% - 20\%$ will get 9.5 points;

4. The students with accuracy scores between $20\% - 40\%$ will get 9 points;

5. The students with accuracy scores between $40\% - 60\%$ will get 8.5 points;
6. The students with accuracy scores between $60\% - 80\%$ will get 7 points;
7. The students with accuracy scores between $80\% - 100\%$ will get 5 points;
8. In addition to above rules, if you can get an accuracy that is **above** 92%, you will get at least **9 points**. If you can get an accuracy that is above 90%, you will get at least **8 points** (This can easily be obtained by simple algorithms we have covered in our lectures).
9. If there are **ties**, i.e., two or more students have the same overall accuracy score, the students who submitted the results earlier will get higher ranks.
10. If your accuracy is less than 60%, you will get only 1 point.