CS 412 - Machine Learning
Project Assignment   —   Spring 2020-2021

Due Date: Wednesday, June 9, 2021, 23:55

Late Submission: Late submissions will **not** be accepted.

The goal of this project is to develop high performance classifiers for identifying American Sign Language. There are 29 classes in this language: 26 of them are for the letters A-Z and 3 classes for SPACE, DELETE and NOTHING.

1   Dataset

The data set is a collection of images of characters from the American Sign Language, separated in 29 folders which represent the various classes.

Download the dataset from the link. The training data set contains 29,000 images which are 200x200 pixels.

You will be provided with two test sets. The *first test set*, which is already available for download, contains 7250 images in the same format as the training data.

On June 7, 2021, you will be provided with the *second test set*. You will run your models over this test set as well and submit your prediction results to SUCourse. This test set will be in the same format as the *first test set*.

For both test sets only the input images will be shared, the gold standard labels will not be shared.

We have setup a Kaggle competition for this task at the following link link.. You should submit your *first test set* results to this kaggle page to appear in the public leader board.

2   Terms and Rules

- You should form the same team that you have been informed the email.

- Only one person for each team will form a team on Kaggle site and invite the team members to join.

- Name your team as follows: `"SU_GroupX"`, e.g. if you are team 11, your team name should be `"SU_Group11"`.

- You may use any approach (any classifier and any data normalization; dimensionality reduction, feature selection and extraction techniques, etc.). You can also use Transfer Learning approaches provided that the pre-trained models you use **must not** be trained for this specific task. Make sure your preprocessing/feature extraction can be

done easily on the test data (e.g. compactly located at the end of your code); since when given the *second test data*, you will need to apply the same feature processing steps on this data as well.

- You can submit your model predictions for *first test set* to Kaggle in order to observe the performance of your model. Each group can do at most 3 submissions per day. This will form the Public Leader board where you can see your standing until June 7, 2021. Accuracy will be used as the main metric to rank systems at leader board.

- You should not use any additional features. This is to prevent you focusing on feature engineering through domain knowledge (not within the scope of this project).

- You are not allowed to use any extra/outside dataset (this dataset related or other data). Only given training data should be used for training your system. Furthermore, your submitted code must give the exact test results as the one in Kaggle when trained with the given training data.

- You will have to submit your submissions until (June 9, 23:55) and late submissions will not be accepted.

- For this project you are expected to use Python 3 notebook and submit only the .ipynb and .py, no other format will be accepted.

- Your report should be a 1 or 2-page report explaining your approach and findings and should be submitted along with your code and *second test set* predictions, by the project closing time.

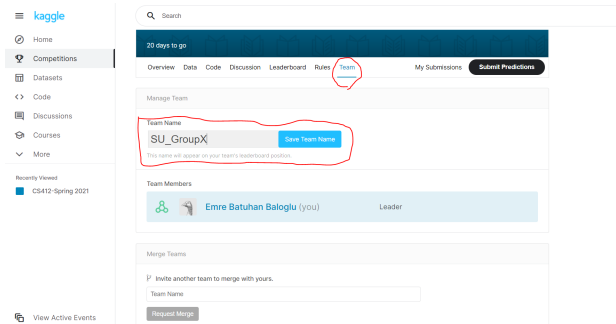- Any violation of the above rules will result in large penalty points.

3 Submission Instructions

- Join the competition by clicking the button and accept the competition rules (all group members should do this):
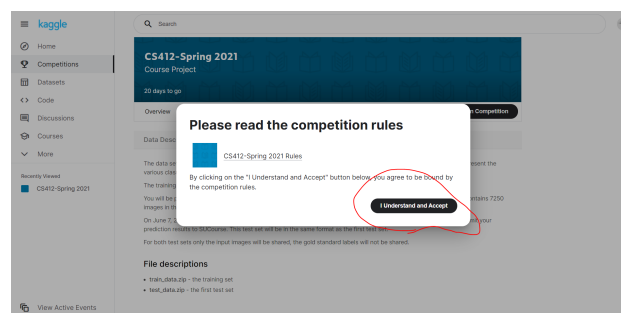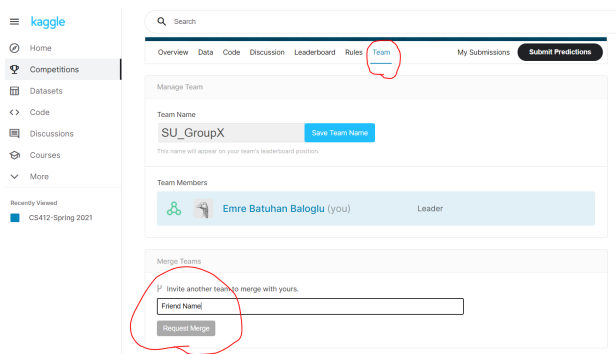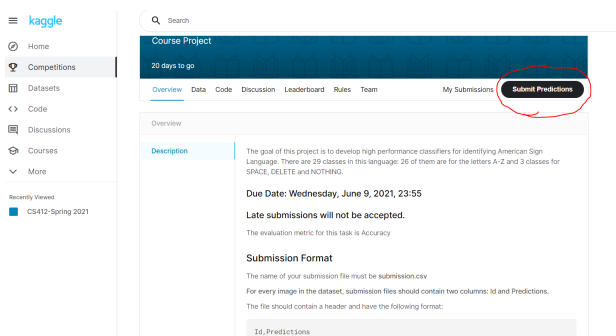

- One of the group members should go to the "team" tab and rename your team as instructed.

- That member should invite other group members using the "Merge Teams" section in the teams tab.

- After training your model, you should submit the predictions of the model for the test set that is provided at that time (*first test set* before June 7, 2021 and *second test set* after June 7, 2021.

- Your submission file **must** be a .csv file, having the same number of predictions as the size of the test set. Name your file as "submission.csv". Also you must name the columns as "Id" and "Prediction". (case sensitive)