

# AIST4010-Spring2022-A0

This is a non-grading assignment for you to get familiar with Kaggle.

4 teams · 7 days to go

[Overview](#)[Data](#)[Code](#)[Discussion](#)[Leaderboard](#)[Rules](#)[Team](#)[My Submissions](#)[Submit Predictions](#)[...](#)

## Overview

### Description

### Evaluation

This is a non-grading assignment for you to get familiar with the Kaggle environment.

For official assignments, half of them will be fixed-answer questions while Half of them will be held on the Kaggle platform as competitions.

In assignment 0, you need to deal with the IRIS classification problem. Try any algorithm you like, just beat the very low baseline!

Data can be downloaded from the 'Data' tab on this same page on Kaggle.

Regarding submission, you need to submit all your codes (including data-processing, training, prediction, etc.) to this Kaggle page.

You can click the "New Notebook" button to create a notebook from the 'Code' tab on this same page to write your codes or upload your local notebook.

Your codes are expected to generate a '.csv' including your predictions on the X\_test data. The F-1 score will be calculated by the output file.

Good luck!

### Description

### Evaluation

The evaluation metric for this competition is [Mean F1-Score](#). The F1 score, commonly used in information retrieval, measures accuracy using the statistics precision (  $\text{p}$  ) and recall (  $\text{r}$  ).

Precision is the ratio of true positives (  $\text{tp}$  ) to all predicted positives (  $\text{tp} + \text{fp}$  ). Recall is the ratio of true positives (  $\text{tp}$  ) to all actual positives (  $\text{tp} + \text{fn}$  ). The F1 score is given by:

$$\left[ \text{F1} = 2 \frac{\text{p} \cdot \text{r}}{\text{p} + \text{r}} \mid \text{where } \text{p} = \frac{\text{tp}}{\text{tp} + \text{fp}}, \text{r} = \frac{\text{tp}}{\text{tp} + \text{fn}} \right]$$

The F1 metric weights recall and precision equally, and a good retrieval algorithm will maximize both precision and recall simultaneously. Thus, moderately good performance on both will be favored over extremely good performance on one and poor performance on the other.

## Submission Format

For every author in the dataset, submission files should contain two columns: `Id` and `Predicted`.

The file should contain a header and have the following format:

```
Id,Predicted
0,1
1,1
2,0
```