Homework #1, 2, 3

Implementing a simple data mining program

To Do

- Implement a simple data mining program that works as follows:

- Initial screen

```
[ Student ID: your student ID ]
[ Name: your name ]

1. Titanic Survivor Predictor
2. Market Basket Analyzer
3. Quit
```

1. Titanic Survivor Predictor (1/2)

① Read the training data

- `train.csv` (https://www.kaggle.com/c/titanic/data)

② Preprocess the data if necessary

- Feature selection, missing values, discretization, binarization, …

③ Build ***one*** of the following classifiers using ***scikit-learn***

- Neural network
  - https://scikit-learn.org/stable/modules/neural_networks_supervised.html#classification
- AdaBoost
  - https://scikit-learn.org/stable/modules/ensemble.html#usage
- Random forest
  - https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees

1. Titanic Survivor Predictor (2/2)

④ Read the test data

- `test.csv` (https://www.kaggle.com/c/titanic/data)

⑤ Generate a prediction file for the test data

- (ex) `submission.csv`
- Refer to the example in https://www.kaggle.com/c/titanic/data

⑥ Return to the initial screen

2. Market Basket Analyzer (1/2)

① Read the dataset

- `Market_Basket_Optimisation.csv` (https://www.kaggle.com/roshansharma/market-basket-optimization)

② Receive the minimum support from the user

```
Enter the minimum support: 0.05
```

user input

2. Market Basket Analyzer (2/2)

③ Find frequent sets using Apriori ***or*** fpgrowth provided by ***mlxtend***

- Apriori
  - http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/
- fpgrowth
  - http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/fpgrowth/

④ Print the found frequent sets using the following code

```
apriori(df, min_support=..., use_colnames=True)
```

or

```
fpgrowth(df, min_support=..., use_colnames=True)
```

⑤ Return to the initial screen

Notes (1/2)

- You can assume that the following data files exist in the ***same*** folder as the program file (`.py`)
  - `train.csv`, `test.csv`, `Market_Basket_Optimisation.csv`

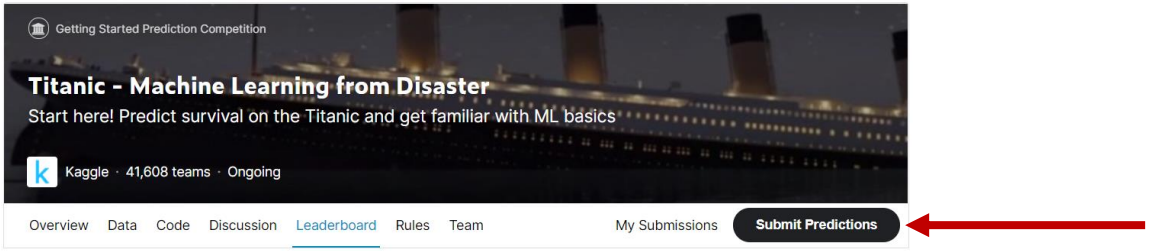
- You must register to the Kaggle (www.kaggle.com)
  - An online community of data scientists and machine learning practitioners
  - Allows users to
    - Find and publish data sets
    - Explore and build models in a web-based data-science environment
    - Work with other data scientists and machine learning engineers
    - Enter competitions to solve data science challenges

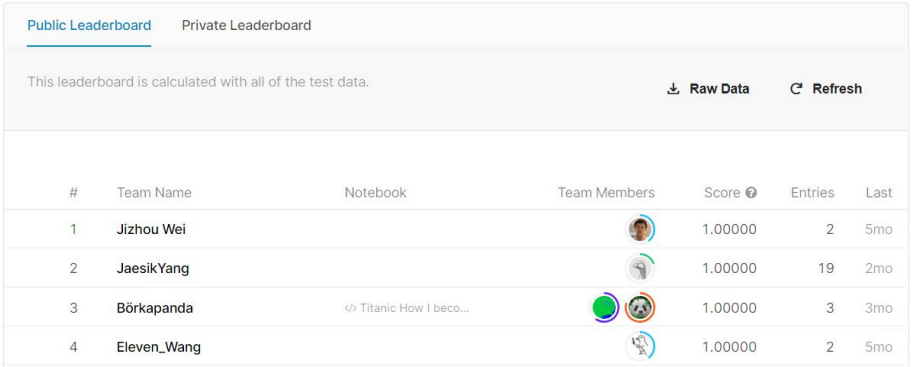
kaggle™

Notes (2/2)

- The accuracy of your classifier must be ***higher than 0.76555***
  - This is when we assume all and only female passengers survive

- You can measure the accuracy of your classifier on Kaggle
  - By submitting your csv file directly (i.e., “Submit Predictions”)
  - Note that you may submit a maximum of ***10*** entries per day





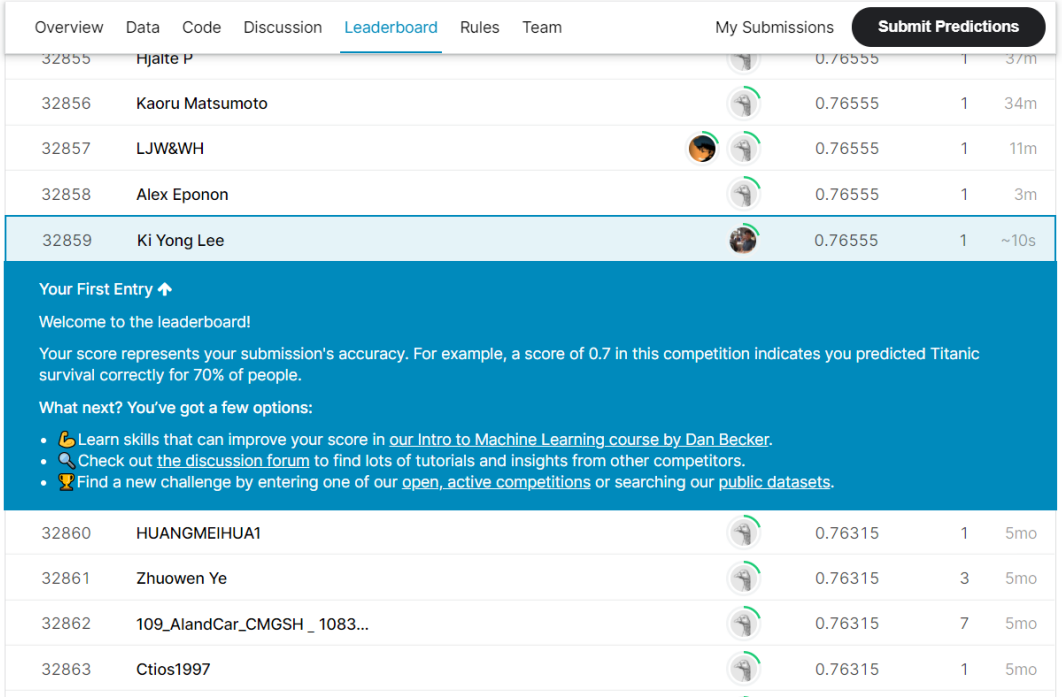
For Your Information

- For Titanic Survivor Predictor, you can refer to
  - https://www.kaggle.com/c/titanic/code
  - Google search: ‘kaggle’ ‘titanic’ ‘scikit-learn’

- For Market Basket Analyzer, you can refer to
  - https://www.kaggle.com/roshansharma/market-basket-analysis
  - https://www.kaggle.com/roshansharma/market-basket-optimization/code

Submission (1/2)

- Compress the following files to create a zip file
  - Source file (`.py`)
  - Screen capture of the Kaggle leaderboard that shows your accuracy



  - Zip file name: *`studentID`*`.zip` (e.g., `1234567.zip`)

Submission (2/2)

- Upload the zip file to the SnowBoard
  - SnowBoard → 데이터마이닝및분석 → 13주차 → Homework

- Due: **2021.6.18 (Fri) 23:55**
  - 1-day delay: 80% credit
  - 2-day delay or more: 0% credit

Evaluation Criteria (30 pts)

- [Homework #1] Titanic survivor predictor (10 pts)
  - The accuracy of your classifier must be higher than ***0.76555***
  - ***Additional points can be given to the students with the highest accuracy***

- [Homework #2] Market basket analyzer (10 pts)
  - You must print frequent sets correctly

- [Homework #3] Program completeness & correctness
  - Whether your program runs in accordance with the requirements
  - (ex) input file reading, output file writing, program behavior

Homework Support

- **[Off-line]** If you need help from me, I recommend you to request an off-line meeting with me
  - The number of students ≥ 1
  - Schedule a meeting time with me

- **[On-line]** Slack
  - Of course, you can use the Slack for Q&A, discussion, etc.

- **[Teaching Assistant]** You can get help from the T.A.
  - Han-Seul Kim (Master's student)
  - uo3359@sookmyung.ac.kr