

---

# 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](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

A dashed vertical line with an upward-pointing arrowhead connects the text 'user input' to the input field '0.05' in the text box above.

## 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/](http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/)

- fpgrowth

- [http://rasbt.github.io/mlxtend/user\\_guide/frequent\\_patterns/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)

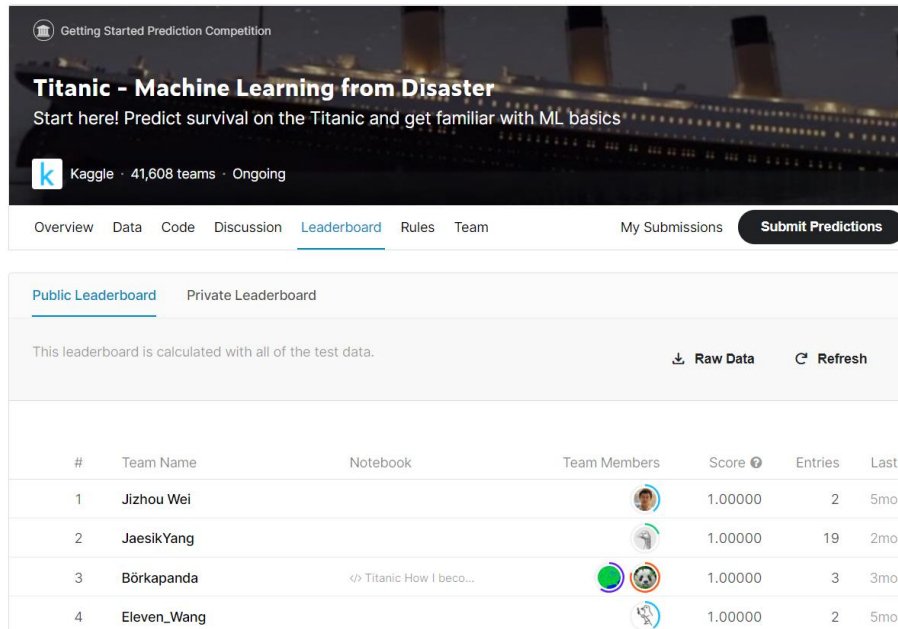
---

- 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](http://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



# 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



The screenshot shows the Kaggle interface for the "Titanic - Machine Learning from Disaster" competition. At the top, there's a header with the competition title and a "Submit Predictions" button highlighted with a red arrow. Below the header, there's a navigation bar with tabs for Overview, Data, Code, Discussion, Leaderboard (selected), Rules, and Team. The main content area displays the "Public Leaderboard" with a table of top teams. The table has columns for Rank, Team Name, Notebook, Team Members, Score, Entries, and Last Update. The top four teams are listed, all with a score of 1.00000.

| # | Team Name   | Notebook                              | Team Members | Score   | Entries | Last |
|---|-------------|---------------------------------------|--------------|---------|---------|------|
| 1 | Jizhou Wei  |                                       |              | 1.00000 | 2       | 5mo  |
| 2 | JaesikYang  |                                       |              | 1.00000 | 19      | 2mo  |
| 3 | Börkapanda  | <a href="#">Titanic How I beco...</a> |              | 1.00000 | 3       | 3mo  |
| 4 | Eleven_Wang |                                       |              | 1.00000 | 2       | 5mo  |



# 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

| Overview  | Data                      | Code | Discussion | Leaderboard | Rules | Team | My Submissions | Submit Predictions |
|---|---------------------------|------|------------|-------------|-------|------|----------------|--------------------|
| 32855   | Hjalte P                  |      |            |             |       |      | 0.76555        | 1 37m              |
| 32856   | Kaoru Matsumoto           |      |            |             |       |      | 0.76555        | 1 34m              |
| 32857   | LJW&WH                    |      |            |             |       |      | 0.76555        | 1 11m              |
| 32858   | Alex Eponon               |      |            |             |       |      | 0.76555        | 1 3m               |
| 32859   | Ki Yong Lee               |      |            |             |       |      | 0.76555        | 1 ~10s             |
| <b>Your First Entry</b>   |                           |      |            |             |       |      |                |                    |
| Welcome to the leaderboard!   |                           |      |            |             |       |      |                |                    |
| Your score represents your submission's accuracy. For example, a score of 0.7 in this competition indicates you predicted Titanic survival correctly for 70% of people.   |                           |      |            |             |       |      |                |                    |
| What next? You've got a few options:  |                           |      |            |             |       |      |                |                    |
| <ul style="list-style-type: none"><li>🧠 Learn skills that can improve your score in <a href="#">our Intro to Machine Learning course</a> by Dan Becker.</li><li>🔍 Check out <a href="#">the discussion forum</a> to find lots of tutorials and insights from other competitors.</li><li>🏆 Find a new challenge by entering one of our <a href="#">open, active competitions</a> or searching our <a href="#">public datasets</a>.</li></ul> |                           |      |            |             |       |      |                |                    |
| 32860   | HUANGMEIHUA1              |      |            |             |       |      | 0.76315        | 1 5mo              |
| 32861   | Zhuowen Ye                |      |            |             |       |      | 0.76315        | 3 5mo              |
| 32862   | 109_AlandCar_CMGS_1083... |      |            |             |       |      | 0.76315        | 7 5mo              |
| 32863   | Ctios1997                 |      |            |             |       |      | 0.76315        | 1 5mo              |

- 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  $\geq 1$
  - 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](mailto:uo3359@sookmyung.ac.kr)