

# 인공지능

---

Artificial Intelligence

# 캐글(Kaggle) 사용법

---

# 1. 캐글이란: 데이터사이언스 경진대회 플랫폼

- 캐글 (Kaggle)

- 가장 유명한 데이터 과학 경진대회 플랫폼

- 2010년 예측모델 및 분석을 위한 플랫폼 서비스로 출발하여 2017년구글에인수
    - 2019년 기준 13,000여개의 데이터를 공개
    - 의료, 경제, 자연과학, 공학 등 거의 모든 분야의 데이터를 다루며 무려 190개 이상의 국가로부터 100만명 이상의 회원이 가입하여 활동 중
    - 주어진 과제에 예측모델을 만들고 학습 결과를 업로드 하면 정확도가 평가됨
    - 이를 기반으로 포인트를 획득하여 레벨을 업그레이드 할 수 있음
    - 레벨에 따라 데이터 과학자로 취업할 수 있는 기회가 주어지기도 함
    - 챌린지에서 입상을 하게 되면 다양한 범주의 상금 획득 가능

- 데이콘 (Daicon) / AI.Factory

- 국내 최대의 데이터 사이언스 경진대회 플랫폼 (한국형 캐글)

# 1. 캐글이란: 데이터사이언스 경진대회 플랫폼

- **기업**에서 본인들의 문제를 공개적으로 **해결**하고 싶었다.
- **기업**에서 훌륭한 데이터사이언스를 **채용**하고 싶었다.
- **정부기관/단체**에서 데이터사이언스를 **양성**하고 싶었다.
- **개인**은 데이터사이언스로 **성장**하고 싶었다.

기업, 정부기관, 단체, 연구소, 개인

Dataset  
With Prize

kaggle

Dataset & Prize  
개발 환경(kernel)  
커뮤니티(follow, discussion)

전 세계 데이터 사이언티스트

## 2. 캐글소개

- 목표
  - 개인의 실력 향상을 위한 툴로 사용하는 것이 가장 좋음
- 캐글 내 활동 가능 분야
  - Competition: 대회 순위에 따른 메달
  - Notebook: 좋은 설명, 좋은 코드에 따른 메달
  - Dataset: 좋은 데이터셋
  - Discussion: 댓글 및 좋은 토론
- 캐글 내 등급 (Kaggle Performance Tier)



초록색(Novice) 다음은 하늘색(Contributor) 다음은 보라색(Expert) 다음은 주황색(Master) 다음은 금색(Grandmaster)

기업 인턴십 조건


## 2. 캐글소개

### Competition

- 대회에서 좋은 결과를 얻는 것을 목표로 함

대회 참여 숫자에 제한 없음

InClass → 교육용



- Home
- Compete**
- Data
- Notebooks
- Discuss
- Courses
- Jobs
- More

Recently Viewed










- 2020.AI.중간고사.문제5
- 2020.AI.MNIST
- External Data Thread
- Mental Health in Tech ...
- 2020.AI.Boston

Search

### All Competitions

Active (Not Entered) Completed InClass

All Categories ▾ Default Sort ▾

	<b>OSIC Pulmonary Fibrosis Progression</b> Predict lung function decline Featured • a month to go • Code Competition • 1382 Teams	\$55,000
	<b>Lyft Motion Prediction for Autonomous Vehicles</b> Build motion prediction models for self-driving vehicles Featured • 3 months to go • Code Competition • 247 Teams	\$30,000
	<b>Cornell Birdcall Identification</b> Build tools for bird population monitoring Research • 13 days to go • Code Competition • 1208 Teams	\$25,000
	<b>Google Landmark Recognition 2020</b> Label famous (and not-so-famous) landmarks in images Research • a month to go • Code Competition • 495 Teams	\$25,000
	<b>Halite by Two Sigma</b> Collect the most halite during your match in space Featured • 13 days to go • Simulation Competition • 1104 Teams	Swag
	<b>Conway's Reverse Game of Life 2020</b> Reverse the arrow of time in the Game of Life Playground • 3 months to go • Code Competition • 21 Teams	Swag
	<b>Predict Future Sales</b> Final project for "How to win a data science competition" Coursera course Playground • 4 months to go • 8517 Teams	
	<b>Titanic: Machine Learning from Disaster</b> Start here! Predict survival on the Titanic and get familiar with ML basics Getting Started • Ongoing • 19264 Teams	Knowledge

## 2. 캐글소개

### Competition

- 대회에서 좋은 결과를 얻는 것을 목표로 함 → 메달획득

<https://www.kaggle.com/c/landmark-retrieval-2020/leaderboard>

≡ kaggle

🏠 Home

🏆 Compete

📁 Data

📄 Notebooks

💬 Discuss

🎓 Courses

💼 Jobs

⌵ More

Recently Viewed

👤 OSIC Pulmonary Fibros...

📊 2020.AI.중간고사.문제5

📊 2020.AI.MNIST

👤 External Data Thread


👤 Mental Health in Tech ...

🔍 Search

Research Code Competition

Google Landmark Retrieval 2020

Given an image, can you find all of the same landmarks in a dataset?

 Google · 541 teams · 16 days ago

Overview

Data

Notebooks

Discussion

Leaderboard

Rules

\$25,000

Prize Money

Late Submission

Public Leaderboard

Private Leaderboard

The private leaderboard is calculated with approximately 66% of the test data. This competition has completed. This leaderboard reflects the final standings.







Refresh

In the money

Gold

Silver

Bronze

#	Δpub	Team Name	Notebook	Team Members	Score ?	Entries	Last
1	—	keetar			0.38677	97	16d
2	—	bysj			0.36278	125	16d
3	▲ 1	TRT			0.34686	72	16d
4	▼ 1	import tensorflow as torch			0.34649	44	16d
5	—	Open Neural Network Exchange			0.32878	81	16d
6	▲ 1	fiSHpAM			0.32600	17	16d

## 2. 캐글소개

### ■ Dataset

- 개인/단체/회사의 데이터셋 공유, 가치 있는 데이터셋 공개 및 가공
- 데이터셋을 통한 커뮤니티 기여

The screenshot displays the Kaggle website interface. On the left is a navigation sidebar with the Kaggle logo and links to Home, Compete, Data (highlighted), Notebooks, Discuss, Courses, Jobs, and More. Below these are 'Recently Viewed' items including Google Landmark Retrieval, OSIC Pulmonary Fibrosis, and 2020 AI competitions. The main content area shows a search bar and a list of datasets. Each dataset entry includes a thumbnail, title, author, upload time, size, rating, file format, and number of tasks. The datasets listed are: Solar Power Generation Data (Ani Kannal, 15d, 2 MB, 10.0 rating, 4 CSV files, 3 tasks), Book-Crossing: User review ratings (Ruchi Bhatia, 22d, 25 MB, 10.0 rating, 3 CSV files, 1 task), AV: Healthcare Analytics II (Neha Prabhavalkar, 5d, 7 MB, 10.0 rating, 4 CSV files, 1 task), arXiv Dataset (Cornell University, 6d, 880 MB, 8.8 rating, 1 JSON file, 3 tasks), 60k Stack Overflow Questions with Quality Rating (Moore, 2d, 21 MB, 10.0 rating, 1 task), US Elections Dataset (Bojan Tunguz, 5d, 7 MB, 9.7 rating, 2 CSV files, 1 task), LEGO Minifigures Classification (Yaroslav Isaienkov, 10h, 5 MB, 9.4 rating, 80 other CSV files, 2 tasks), and Bee or wasp? (George Rey, 10d, 559 MB, 9.4 rating, 11425 other CSV files, 2 tasks).

Dataset Name	Author	Time	Size	Rating	Files	Tasks
Solar Power Generation Data	Ani Kannal	15d	2 MB	10.0	4 Files (CSV)	3 Tasks
Book-Crossing: User review ratings	Ruchi Bhatia	22d	25 MB	10.0	3 Files (CSV)	1 Task
AV : Healthcare Analytics II	Neha Prabhavalkar	5d	7 MB	10.0	4 Files (CSV)	1 Task
arXiv Dataset	Cornell University	6d	880 MB	8.8	1 File (JSON)	3 Tasks
60k Stack Overflow Questions with Quality Rating	Moore	2d	21 MB	10.0		1 Task
US Elections Dataset	Bojan Tunguz	5d	7 MB	9.7	2 Files (CSV)	1 Task
LEGO Minifigures Classification	Yaroslav Isaienkov	10h	5 MB	9.4	80 Files (other, CSV)	2 Tasks
Bee or wasp?	George Rey	10d	559 MB	9.4	11425 Files (other, CSV)	2 Tasks



## 2. 캐글소개

- Notebook
  - 커뮤니티 내 소통의 창구, 설명과 시각화에 노력
  - Jupyter Notebook의 캐글 판

The screenshot displays the Kaggle Notebooks interface. On the left is a sidebar with navigation links: Home, Compete, Data, Notebooks (selected), Discuss, Courses, Jobs, and More. Below these is a 'Recently Viewed' section listing notebooks like 'Google Landmark Retri...', 'OSIC Pulmonary Fibros...', '2020.AI.중간고사.문제5', '2020.AI.MNIST', and 'External Data Thread'. The main content area features a search bar at the top, followed by the 'Notebooks' title and a sub-header 'Explore and run machine learning code with Kaggle Notebooks! Find help in the [Documentation](#).' A '+ New Notebook' button is in the top right. Below this, a 'GPU quota: 41h remaining' progress bar is shown. The notebook list is filtered by 'Public' and sorted by 'Hotness'. Each entry includes a rank, a user profile picture, the notebook title, a brief description, tags, and icons for viewing the notebook, its code, and comments.

Rank	User	Notebook Title	Description	Tags	Views	Code	Comments
51	[Profile]	Heart Failure Prediction & Visualization	1h ago in Heart Failure Prediction	beginner, exploratory data analysis, data visualization, classifi...	43	Py	0
13	[Profile]	You're In!	4h ago in Campus Recruitment	exploratory data analysis, random forest, logistic regression	0	Py	0
55	[Profile]	Top 10%. Efficient ensembling in few lines of code	4h ago in Titanic: Machine Learning from Disaster	ensembling, model comparison, t...	28	Py	0
125	[Profile]	Amazon Alexa Reviews	8h ago in Amazon Alexa Reviews	nlp, data visualization, classification, spaCy	42	Py	0
7	[Profile]	BBC News Categorization using Embedding	4h ago in BBC News Archive	ensembling, dnn, keras	0	Py	0
10	[Profile]	Mall Customer Segmentation Using K-Means	7h ago in Mall Customer Segmentation Data	exploratory data analysis, data visualization, cluste...	0	Py	0

## 2. 캐글소개

### ■ Notebook

- 커뮤니티 내 소통의 창구
- Jupyter Notebook의 캐글 판
  - <https://www.kaggle.com/sanchitakarmakar/heart-failure-prediction-visualization>



#### Heart Failure Prediction & Visualization

Python notebook using data from [Heart Failure Prediction](#) · 1,493 views · 1h ago · beginner, data visualization, exploratory data analysis, +2 more



#### Task Submission

This notebook is a submission for a [Task on Heart Failure Prediction](#).

Version 6 of 6

Notebook

Input (1)

Output

Execution Info

Log

Comments (43)

In [1]:

```
# Importing the libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
# Importing the dataset

dataset = pd.read_csv('../input/heart-failure-clinical-data/heart_failure_clinical_rec
ords_dataset.csv')
```

In [3]:

```
# Lets look at the top 5 rows
dataset.head()
```

Out[3]:

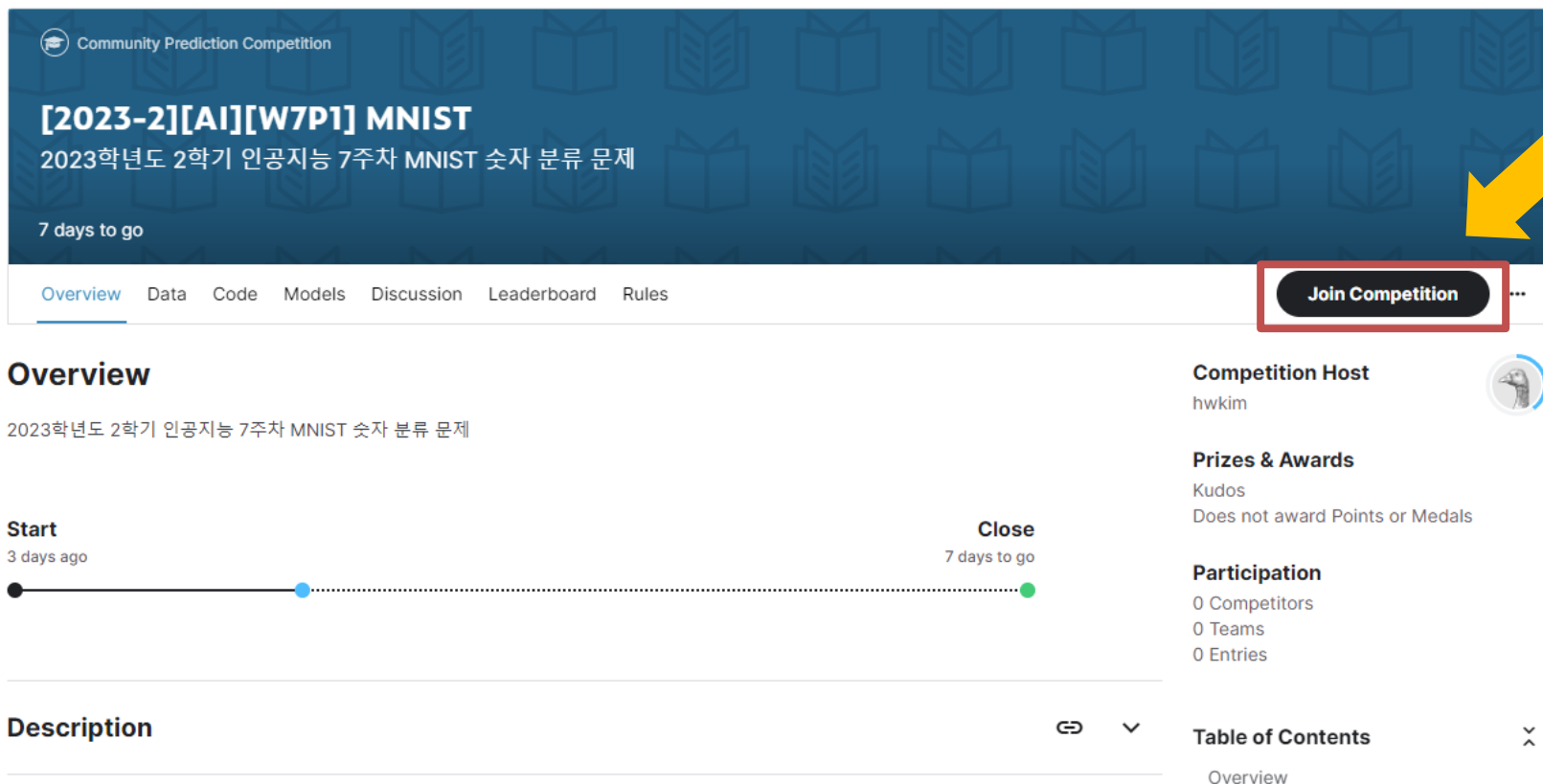
	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets
0	75.0	0	582	0	20	1	265000
1	55.0	0	7861	0	38	0	263350
2	65.0	0	146	0	20	0	162000
3	50.0	1	111	0	20	0	210000
4	65.0	1	160	1	20	0	327000

데이터 분석 및 시각화

코드 설명

# 3. 캐글사용법

- 캐글사용법예시: 인공지능 7주차 실습문제
  - [\[링크\]](#)
  - 참가를 위해 **[Join Competition]** 을 클릭



The screenshot shows the Kaggle interface for a competition titled "[2023-2][AI][W7P1] MNIST". The header includes the Kaggle logo and the text "Community Prediction Competition". Below the title, it says "2023학년도 2학기 인공지능 7주차 MNIST 숫자 분류 문제" and "7 days to go". A navigation bar contains links: Overview, Data, Code, Models, Discussion, Leaderboard, and Rules. A red rectangular box highlights the "Join Competition" button in the top right corner, with a large yellow arrow pointing towards it. The main content area is divided into two columns. The left column has a section titled "Overview" with the subtitle "2023학년도 2학기 인공지능 7주차 MNIST 숫자 분류 문제". Below this is a timeline showing the competition's duration from "Start" (3 days ago) to "Close" (7 days to go). The right column contains information about the "Competition Host" (hwkim), "Prizes & Awards" (Kudos, Does not award Points or Medals), "Participation" (0 Competitors, 0 Teams, 0 Entries), and a "Table of Contents" (Overview).

### 3. 캐글사용법

- 참가를 위해 [I Understand and Accept] 클릭

## Please read the competition rules

By clicking on the "I Understand and Accept" button below, you agree to be bound by the competition rules for [2023-2][AI][W7P1] MNIST.

Don't cheat!

Apply yourself!

Have fun!

I Understand and Accept

# 3. 캐글사용법

- **Overview** 탭에는 해당 실습 문제에 대한 전반적인 설명/목표가 있음
- 주의사항을 제대로 숙지하지 않고 문제를 풀면 시험에서 감점이 발생할 수 있음

OverviewDataCodeModelsDiscussionLeaderboardRulesTeam


SubmissionsSubmit Predictions...

Description

교과목 정보

- 세종대학교 지능기전공학과 (최유경 교수)

MNIST 숫자 분류 문제



본 데이터셋은 인공지능의 대표적인 손글씨 이미지 데이터셋인 MNIST입니다.

MNIST를 구성하는 숫자는 0~9까지 총 10개의 클래스로 구성됩니다. (상단 이미지 참고)

제출 시 주의사항

label을 int형으로 변환하여 제출하시기 바랍니다.

목표

선형분류를 사용하여 MNIST 데이터의 숫자를 분류합니다.

주의사항

랜덤시드는 반드시 고정하여, 매번 실행할 때 마다 성능이 변경되지 않도록 주의 해야 합니다.

코멘트

과제 진행에 어려움이 있다면, 담당 조교와 상의하세요.

Table of Contents

Overview

Description

Evaluation

Citation

### 3. 캐글사용법

- Data 탭에는 문제 해결을 위한 학습/테스트 데이터 그리고 정답 제출 템플릿 파일이 있음
- Description 탭에는 제공되는 제공된 데이터의 설명이 있음
- 데이터 분석 후 정답 제출 템플릿에 정답을 적어 파일을 리더보드에 제출

The screenshot shows the Kaggle Community Prediction Competition page for the 2023-2 AI W7P1 MNIST competition. The page has a dark blue header with the competition title and a countdown timer showing 7 days to go. Below the header is a navigation bar with tabs for Overview, Data, Code, Models, Discussion, Leaderboard, Rules, and Team. The Data tab is selected. The main content area is titled 'Dataset Description' and contains a section for '제공되는 파일 설명' (Provided File Description) with a list of files: train.csv (60,000 rows, 784 columns), test.csv (10,000 rows, 784 columns), and sample\_submit.csv (submission file example). To the right of the description is a sidebar with 'Files' (3 files), 'Size' (127.99 MB), and 'Type' (CSV). Below the description is a section for 'Column name과 정보' (Column name and information) with a list of columns: label (digit 0-9) and 1x1 ~ 28x28 (pixel values). At the bottom, there is a 'sample\_submit.csv' file viewer showing the first two columns: id and label. To the right of the file viewer is a 'Data Explorer' sidebar showing the file size (127.99 MB) and a list of files: sample\_submit.csv, test.csv, and train.csv.

Community Prediction Competition

**[2023-2][AI][W7P1] MNIST**  
2023학년도 2학기 인공지능 7주차 MNIST 숫자 분류 문제  
7 days to go

Overview Data Code Models Discussion Leaderboard Rules Team Submissions **Submit Predictions** ...

#### Dataset Description

제공되는 파일 설명

- train.csv: 60,000개의 데이터가 있고, 각 행은 label과 784개의 픽셀값으로 구성 ( 60,000 x 785 )
- test.csv: 10,000개의 데이터가 있고, 각 행은 784개의 픽셀값으로 구성 ( 10,000 x 784 )
- sample\_submit.csv: submission 파일 예시

Column name과 정보

- label: 손글씨 숫자값 ( 0~9 )
- 1x1 ~ 28x28: 칠점명이 해당하는 위치의 픽셀값

**Files**  
3 files

**Size**  
127.99 MB

**Type**  
CSV

**sample\_submit.csv** (68.9 kB)

Detail Compact Column 2 of 2 columns

id	label
----	-------

**Data Explorer**  
127.99 MB

- sample\_submit.csv
- test.csv
- train.csv

# 3. 캐글사용법

- Leaderboard 탭에는 해당 실습 문제에 대한 **베이스라인** 성능을 확인할 수 있음
- 실습 문제에 대한 해결 여부는 베이스라인을 기준으로 판단

Community Prediction Competition

**[2023-2][AI][W7P1] MNIST**  
2023학년도 2학기 인공지능 7주차 MNIST 숫자 분류 문제  
7 days to go

Overview Data Code Models Discussion Leaderboard Rules Team

Submissions **Submit Predictions** ...

## Leaderboard

Raw Data Refresh

Search leaderboard

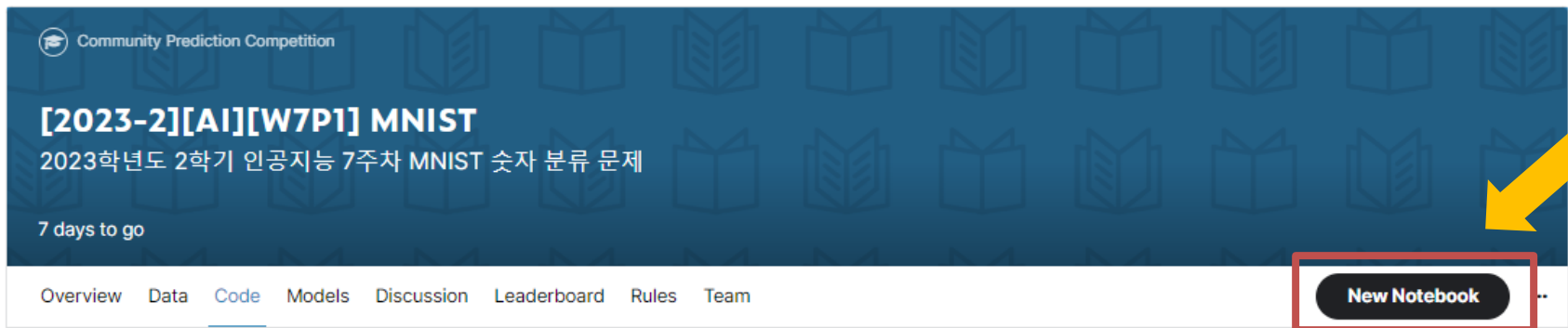
This leaderboard is calculated with all of the test data.

#	Team	Members	Score	Entries	Last	Join
1	SVM Baseline		0.94170			
2	Perceptron Baseline		0.87690			

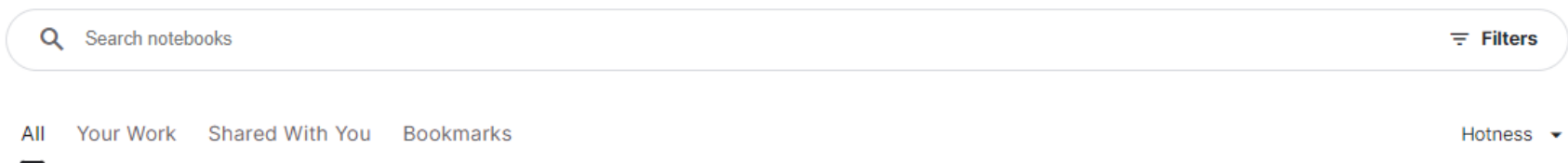


# 3. 캐글사용법

- **Code** 탭에는 해당 실습 문제를 해결하기 위한 실습 노트북을 작성할 수 있음
- 모든 답안 제출은 **실습 노트북 공유**를 통해서만 진행됨



## Notebooks





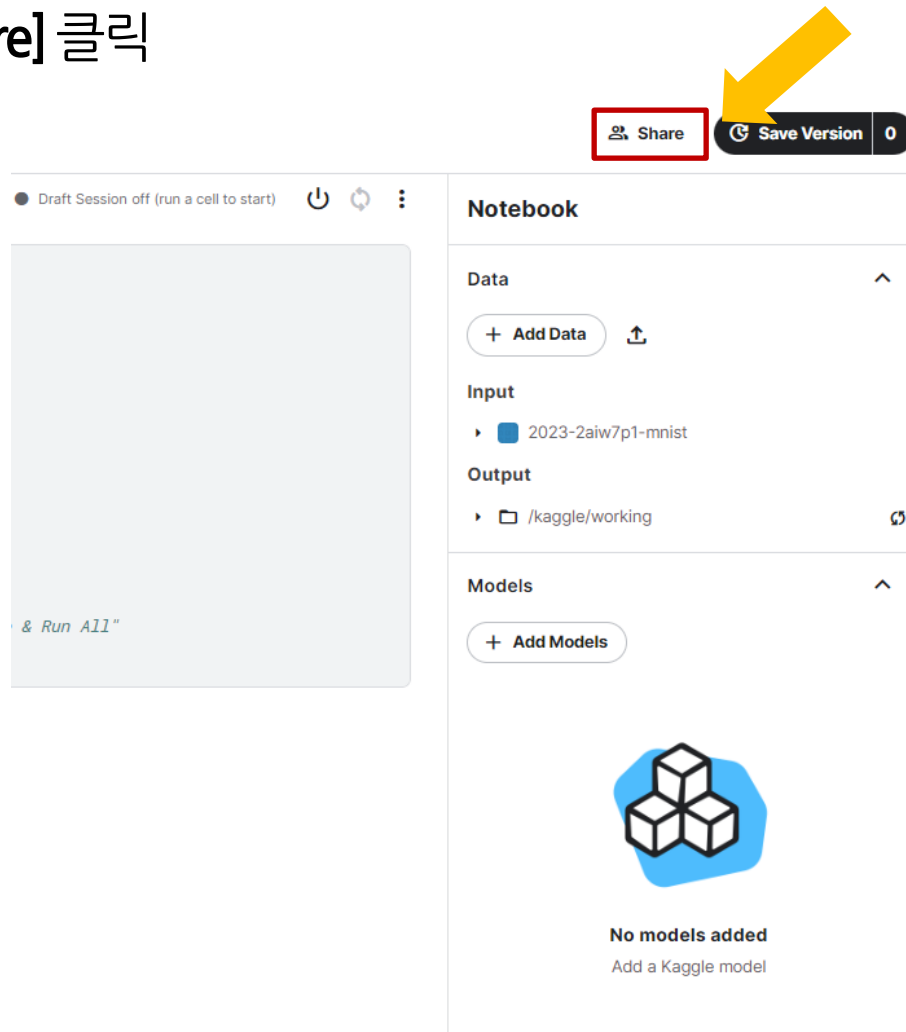
### 3. 캐글사용법

- 노트북 제목은 반드시 양식에 맞출 것 (중간고사, 기말고사는 따로 제공)
- [2023-AI][W7P1]20000000\_홍길동



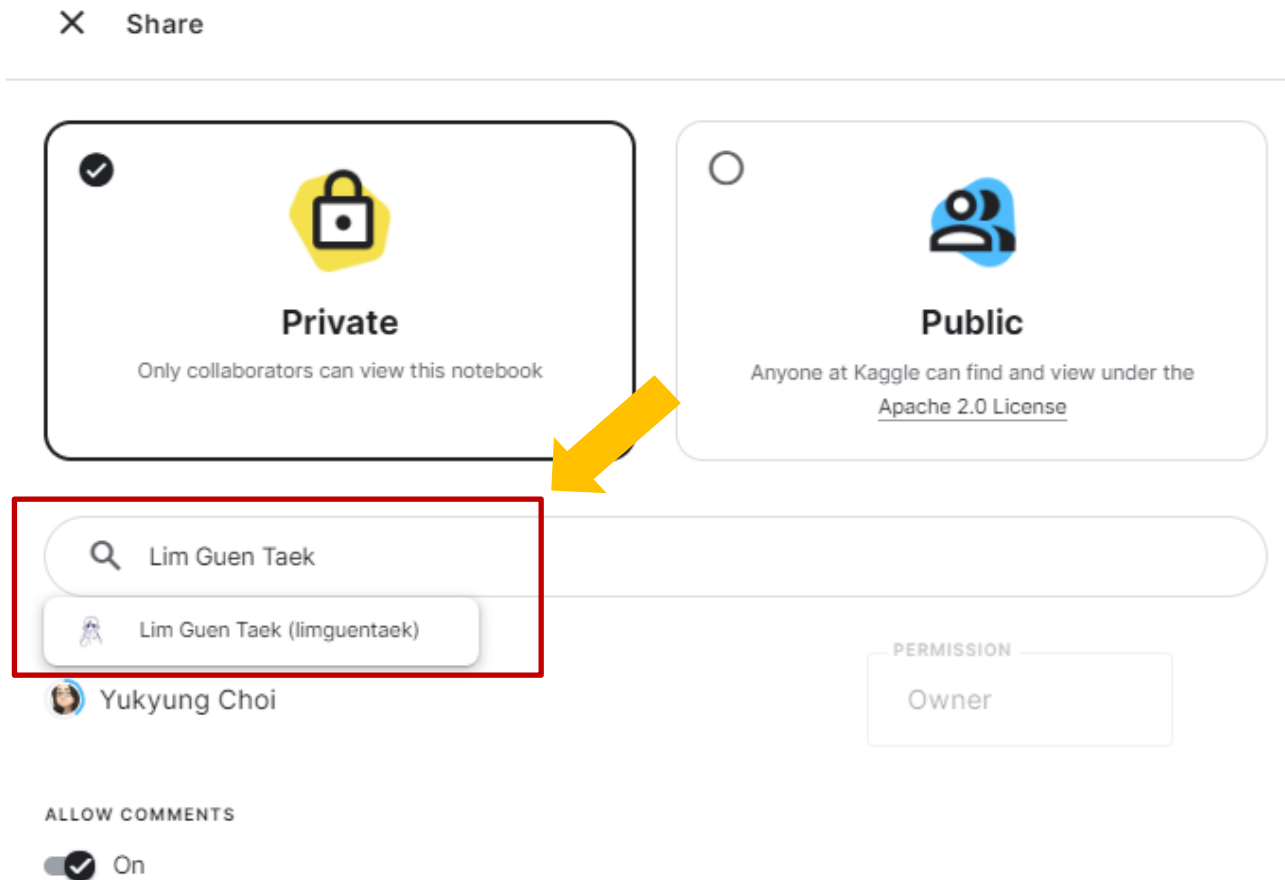
# 3. 캐글사용법

- 조교들에게 공유되지 않은 노트북은 채점할 수 없음
- 공유를 위해 [Share] 클릭



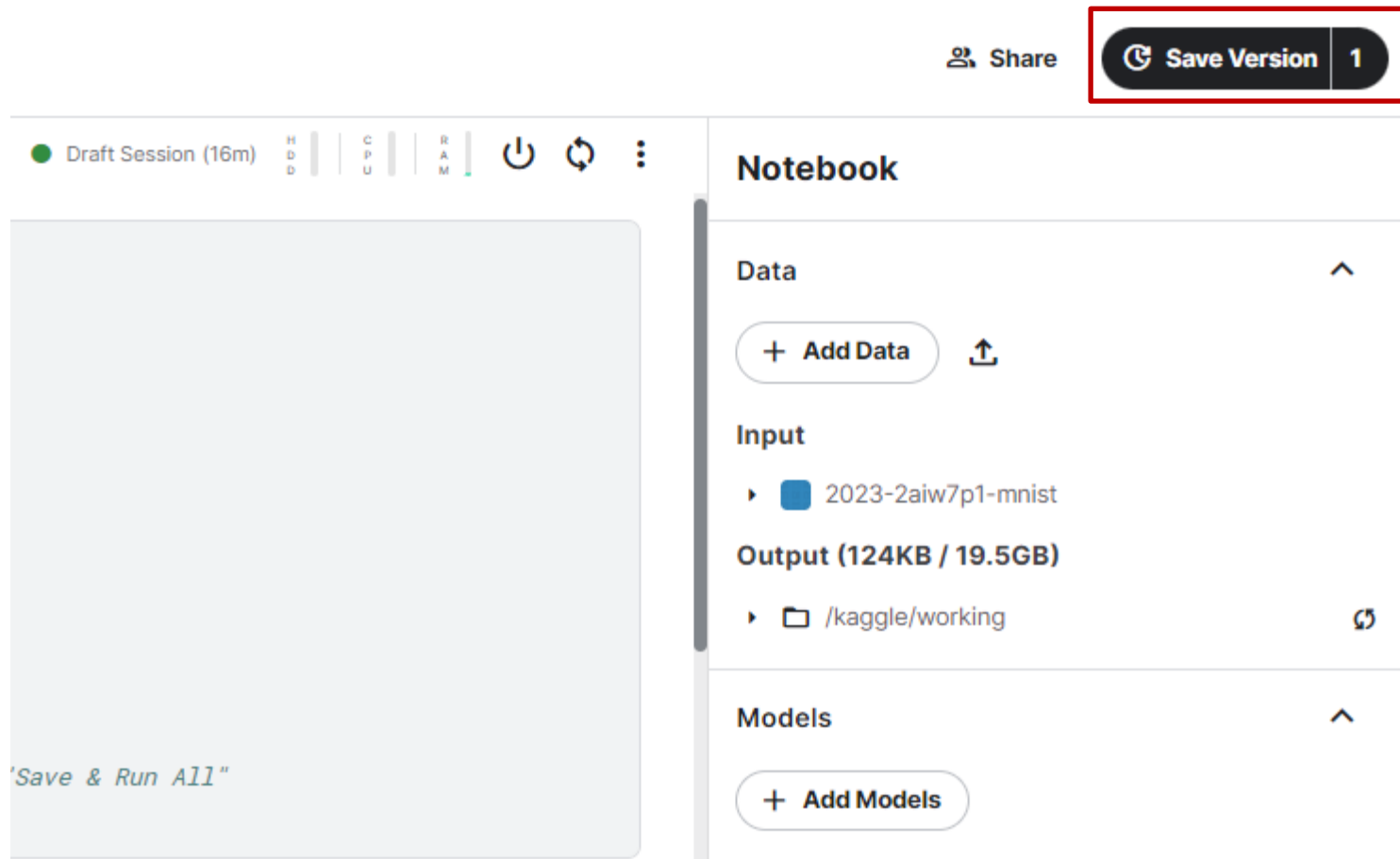
### 3. 캐글사용법

- 조교들 캐글 아이디를 입력하여 반드시 공유할 것
- 조교들 아이디를 입력하고 [Save] 클릭 하면 완료




### 3. 캐글사용법




- 코드 작성이 완료되면 노트북 저장 필수, 저장을 위해 [Save Version] 클릭
- [Save & Run All]을 클릭해서 노트북 저장과 실행을 동시에 진행



### 3. 캐글사용법

- [Save & Run All (Commit)] 이 끝난 노트북은 [Output]란에 들어가서 제출 필요
- Output Data를 확인하고 [Submit]을 눌러서 답안제출

 YUKYUNG CHOI +1 · 2M AGO · PRIVATE

 0  Edit 

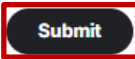


## [2023-AI][W7P1]2000000\_홍길동

Python · [\[2023-2\]\[AI\]\[W7P1\] MNIST](#)


Notebook Input **Output** Logs Comments (0) Settings

### Output Data

my\_submission.csv (68.9 kB)


id	label
1	6




**Output :**  
 my\_submission.csv



### 3. 캐글사용법

- 답안 제출 후 공유된 노트북에서 아래와 같이 점수를 확인할 수 있어야 채점 진행
- 해당 점수를 바탕으로 베이스라인 통과 여부 확인



 YUKYUNG CHOI +1 · 4M AGO · 1 VIEW · PRIVATE

 0  

## [2023-AI][W7P1]2000000\_홍길동

Python · [\[2023-2\]\[AI\]\[W7P1\] MNIST](#)

Notebook Input Output Logs Comments (0) Settings

	Competition Notebook <a href="#">[2023-2][AI][W7P1] MNIST</a>	Run 32.0s	Public Score 0.87690	Best Score <u>0.8769 V1</u>	 Version 1 of 1
---	--	--------------	-------------------------	--------------------------------	--

Add Tags

In [1]:

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load
```