

# 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 2017년 구글에 인수
    - 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 해결하고 싶었다.
- **기업**에서 훌륭한 데이터사이언스를 채용하고 싶었다.
- **정부기관/단체**에서 데이터사이언스를 양성하고 싶었다.
- **개인**은 데이터사이언스로 성장하고 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
  - 개인의실력향상을위한툴로사용하는것이가장 좋음
- 캐글 내 활동 가능 분야
  - Competition: 대회순위에따른메달
  - Notebook: 좋은설명,좋은코드에따른메달
  - Dataset: 좋은데이터셋
  - Discussion: 댓글및좋은토론
- 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 → 메달 획득

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

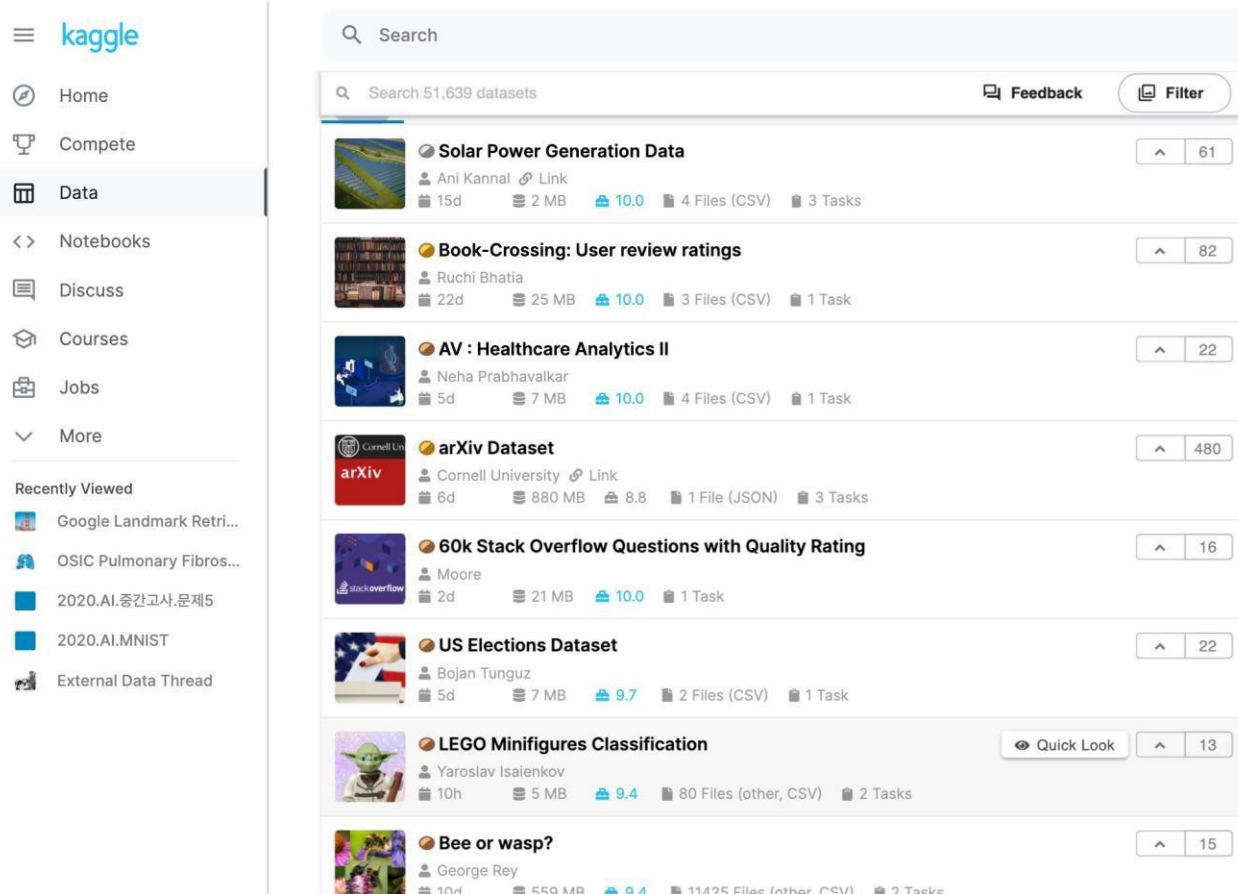
The screenshot shows the Kaggle website interface. On the left is a sidebar with navigation links: Home, Compete, Data, Notebooks, Discuss, Courses, Jobs, and More. Below these are 'Recently Viewed' items. The main content area displays the 'Google Landmark Retrieval 2020' competition page. It includes a search bar, a competition banner with the title 'Google Landmark Retrieval 2020' and a prize of '\$25,000', and a description: 'Given an image, can you find all of the same landmarks in a dataset?'. Below the banner are tabs for Overview, Data, Notebooks, Discussion, Leaderboard (selected), and Rules. A 'Late Submission' button is also present. The 'Leaderboard' section shows a 'Public Leaderboard' and a 'Private Leaderboard' (which is currently empty). A note states: 'The private leaderboard is calculated with approximately 66% of the test data. This competition has completed. This leaderboard reflects the final standings.' Below this is a legend for medal colors: In the money (green), Gold (yellow), Silver (grey), and Bronze (brown). The main table lists the top teams with columns for rank, change in public score, team name, notebook, team members, score, entries, and last update.

#	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/회사의데이터셋공유,가치있는데이터셋공개및가공
- 데이터셋을통한커뮤니티기여



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, 2020 AI Intermediate Exam Questions, 2020 AI MNIST, and an External Data Thread.

The main content area shows a search bar at the top with the text 'Search 51,639 datasets'. Below the search bar is a list of datasets, each with a thumbnail, title, creator, upload time, size, rating, file format, and number of tasks. The datasets listed are:

- Solar Power Generation Data** by Ani Kannal, 15d, 2 MB, 10.0 rating, 4 Files (CSV), 3 Tasks. 61 votes.
- Book-Crossing: User review ratings** by Ruchi Bhatia, 22d, 25 MB, 10.0 rating, 3 Files (CSV), 1 Task. 82 votes.
- AV : Healthcare Analytics II** by Neha Prabhavalkar, 5d, 7 MB, 10.0 rating, 4 Files (CSV), 1 Task. 22 votes.
- arXiv Dataset** by Cornell University, 6d, 880 MB, 8.8 rating, 1 File (JSON), 3 Tasks. 480 votes.
- 60k Stack Overflow Questions with Quality Rating** by Moore, 2d, 21 MB, 10.0 rating, 1 Task. 16 votes.
- US Elections Dataset** by Bojan Tunguz, 5d, 7 MB, 9.7 rating, 2 Files (CSV), 1 Task. 22 votes.
- LEGO Minifigures Classification** by Yaroslav Isaienkov, 10h, 5 MB, 9.4 rating, 80 Files (other, CSV), 2 Tasks. 13 votes. Includes a 'Quick Look' button.
- Bee or wasp?** by George Rey, 10d, 559 MB, 9.4 rating, 11425 Files (other, CSV), 2 Tasks. 15 votes.



## 2. 캐글소개

### ■ Notebook

- 커뮤니티내소통의창구,설명과시각화에노력
- Jupyter Notebook의 캐글판

≡ kaggle

🏠 Home

🏆 Compete

📊 Data

<> Notebooks

💬 Discuss

🎓 Courses

💼 Jobs

⌵ More

Recently Viewed

🖼️ Google Landmark Retri...

👤 OSIC Pulmonary Fibros...

📄 2020.AI.중간고사.문제5

📄 2020.AI.MNIST

👤 External Data Thread

🔍 Search

## Notebooks

Explore and run machine learning code with Kaggle Notebooks! Find help in the [Documentation](#).

+ New Notebook

GPU quota: 41h remaining

Public

Your Work

Shared With You

Favorites

Sort by Hotness

Categories

Outputs

Languages

Tags

Search notebooks

🔍

51



Heart Failure Prediction & Visualization

1h ago in Heart Failure Prediction 🔖 beginner, exploratory data analysis, data visualization, classifi...

📄

Py

💬 43

13



You're In!

4h ago in Campus Recruitment 🔖 exploratory data analysis, random forest, logistic regression

📄

Py

💬 0

55



Top 10%. Efficient ensembling in few lines of code

4h ago in Titanic: Machine Learning from Disaster 📊 0.79425 🔖 ensembling, model comparison, t...

📄

Py

💬 28

125



Amazon Alexa Reviews

8h ago in Amazon Alexa Reviews 🔖 nlp, data visualization, classification, spaCy

📄

Py

💬 42

7



BBC News Categorization using Embedding

4h ago in BBC News Archive 🔖 ensembling, dnn, keras

📄

Py

💬 0

10



Mall Customer Segmentation Using K-Means

7h ago in Mall Customer Segmentation Data 🔖 exploratory data analysis, data visualization, cluste...

📄

Py

💬 0

## 2. 캐글소개

### ■ Notebook

- 커뮤니티내소통의창구
- Jupyter Notebook의 캐글판
  - <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

데이터분석및 시각화

코드설명

# 3. 캐글사용법

- 캐글 사용법 예시: 2023년도 인공지능 7주차 실습 문제
  - 참가를 위해 **[Join Competition]**을 클릭

The screenshot shows the Kaggle interface for a competition titled "[2023-2][AI][W7P1] MNIST". The subtitle is "2023학년도 2학기 인공지능 7주차 MNIST 숫자 분류 문제". A banner at the top indicates "7 days to go". Below the banner is a navigation bar with tabs: Overview, Data, Code, Models, Discussion, Leaderboard, and Rules. A red box highlights the "Join Competition" button in the top right corner, with a yellow arrow pointing to it. The main content area is divided into two columns. The left column has an "Overview" section with a timeline showing the competition started "3 days ago" and will "Close" in "7 days to go". The right column contains information about the "Competition Host" (hwkim), "Prizes & Awards" (Kudos, no points or medals), "Participation" (0 Competitors, 0 Teams, 0 Entries), and a "Table of Contents" (Overview).

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

**Join Competition**

### Overview

2023학년도 2학기 인공지능 7주차 MNIST 숫자 분류 문제

**Start**  
3 days ago

**Close**  
7 days to go

**Competition Host**  
hwkim

**Prizes & Awards**  
Kudos  
Does not award Points or Medals

**Participation**  
0 Competitors  
0 Teams  
0 Entries

**Description**

**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


OverviewDataCodeModelsDiscussionLeaderboardRulesTeamSubmissionsSubmit 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

런타임시드는 반드시 고정하여, 매번 실행할 때 마다 성능이 변경되지 않도록 주의해야 합니다.

코멘트

과제 진행에 어려움이 있다면, 담당 조교와 상의하세요.

Table of Contents

Overview

Description

Evaluation

Citation

### 3. 캐글사용법

- **Data** 탭에는문제해결을위한학습/테스트 데이터그리고정답 제출 템플릿 파일이있음
- **Description** 탭에는제공되는제공된데이터의 설명이있음
- 데이터분석후정답제출템플릿에정답을적어파일을리더보드에제출

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 )
- 1×1 ~ 28×28: 철현명이 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

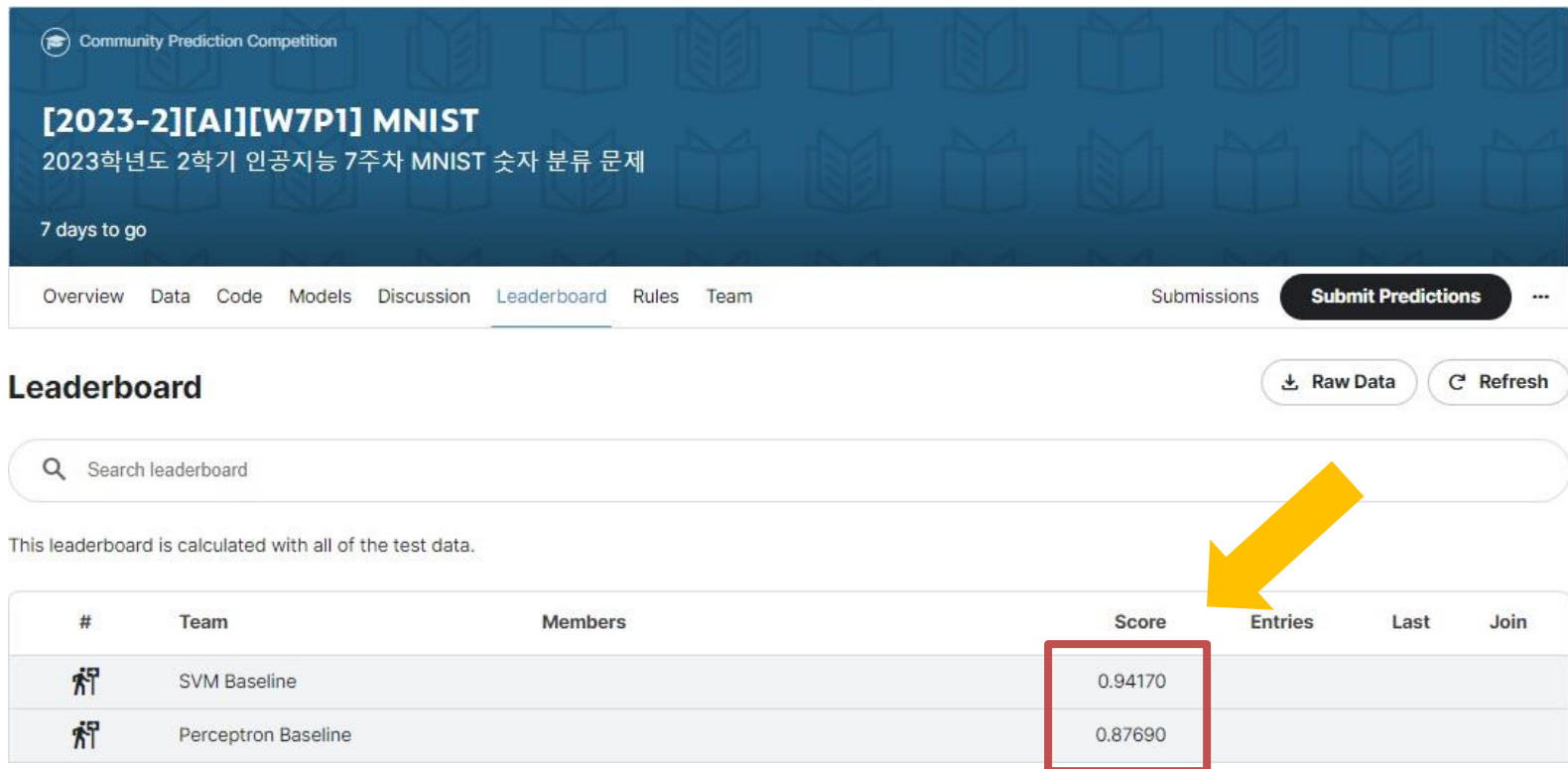
id label

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



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 실습문제를 해결하기 위한 실습 노트북을 작성할 수 있음

