

퍼블릭 클라우드 개발환경

캐글 (Kaggle)

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)

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

캐글 소개편

- 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 데이터 셋
 - 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

캐글 소개편

- 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

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

캐글 소개편

Competition

- 대회에서 좋은 결과를 얻는 것을 목표로 함 → 메달 획득

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

Google Landmark Retrieval 2020
Given an image, can you find all of the same landmarks in a dataset?
Google · 541 teams · 16 days ago
\$25,000 Prize Money

Overview Data Notebooks Discussion **Leaderboard** Rules

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

캐글 소개편

■ Dataset

- 개인/단체/회사의 데이터 셋 공유, 가치 있는 데이터 셋 공개 및 가공
- 데이터 셋을 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, 2020 AI Intermediate 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. Upvotes: 61.
- Book-Crossing: User review ratings** by Ruchi Bhatia, 22d, 25 MB, 10.0 rating, 3 Files (CSV), 1 Task. Upvotes: 82.
- AV : Healthcare Analytics II** by Neha Prabhavalkar, 5d, 7 MB, 10.0 rating, 4 Files (CSV), 1 Task. Upvotes: 22.
- arXiv Dataset** by Cornell University, 6d, 880 MB, 8.8 rating, 1 File (JSON), 3 Tasks. Upvotes: 480.
- 60k Stack Overflow Questions with Quality Rating** by Moore, 2d, 21 MB, 10.0 rating, 1 Task. Upvotes: 16.
- US Elections Dataset** by Bojan Tunguz, 5d, 7 MB, 9.7 rating, 2 Files (CSV), 1 Task. Upvotes: 22.
- LEGO Minifigures Classification** by Yaroslav Isaienkov, 10h, 5 MB, 9.4 rating, 80 Files (other, CSV), 2 Tasks. Upvotes: 13. Includes a 'Quick Look' button.
- Bee or wasp?** by George Rey, 10d, 559 MB, 9.4 rating, 11425 Files (other, CSV), 2 Tasks. Upvotes: 15.

캐글 소개편

- 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 description: '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, user profile, title, description, tags, and interaction buttons (view, code, comments). The first notebook, 'Heart Failure Prediction & Visualization', is highlighted in yellow.

Rank	User	Title	Description	Tags	Interactions
51	[Profile]	Heart Failure Prediction & Visualization	1h ago in Heart Failure Prediction	beginner, exploratory data analysis, data visualization, classifi...	View, Py, 43
13	[Profile]	You're In!	4h ago in Campus Recruitment	exploratory data analysis, random forest, logistic regression	View, 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...	View, Py, 28
125	[Profile]	Amazon Alexa Reviews	8h ago in Amazon Alexa Reviews	nlp, data visualization, classification, spaCy	View, Py, 42
7	[Profile]	BBC News Categorization using Embedding	4h ago in BBC News Archive	ensembling, dnn, keras	View, Py, 0
10	[Profile]	Mall Customer Segmentation Using K-Means	7h ago in Mall Customer Segmentation Data	exploratory data analysis, data visualization, cluste...	View, Py, 0

캐글 소개편

■ 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	26500
1	55.0	0	7861	0	38	0	26335
2	65.0	0	146	0	20	0	16200
3	50.0	1	111	0	20	0	21000
4	65.0	1	160	1	20	0	32700

데이터 분석 및 시각화

코드 설명

참가 방법과 리더보드 사용 방법

- 캐글 사용법 예시: 인공지능 중간고사 문제
 - <https://www.kaggle.com/c/2020-ai-exam-cancer/>
- 참가를 위해 [Join Competition] 을 클릭

The screenshot shows the Kaggle website interface. On the left is a navigation menu with links to Home, Compete, Data, Notebooks, Discuss, Courses, Jobs, and More. The main content area displays the 'InClass Prediction Competition' for '2020.AI.중간고사.문제5' (Breast Cancer Diagnosis). The competition title is in Korean: '유방암 악성/양성 종양 예측 문제' (Breast Cancer Malignant/Benign Tumor Prediction Problem). It shows 31 teams and was created 4 months ago. Below the title are tabs for Overview, Data (selected), Notebooks, Discussion, Leaderboard, and Rules. A prominent 'Join Competition' button is located at the bottom right of the competition header. A large yellow arrow with a red outline points directly to this button. Below the tabs is a 'Data Description' section titled '제공되는 데이터의 설명' (Description of provided data), which lists three CSV files: '2020.AI.cancer-train.csv' (exam results and diagnosis), '2020.AI.cancer-test.csv' (diagnosis results), and '2020.AI.cancer-submission.csv' (submission file example). At the bottom, there is a terminal-like box with the command 'kaggle competitions download -c 2020-ai-exam-cancer' and icons for file and help.

참가 방법과 리더보드 사용 방법

- 참가를 위해 [I Understand and Accept] 클릭

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. The main content area displays the competition details for '2020.AI.중간고사.문제5' (2020 AI Intermediate Exam Question 5), which is an 'InClass Prediction Competition' with 31 teams and 4 months ago. A yellow arrow points to the 'I Understand and Accept' button. Below this button is a section titled 'Rules' containing a paragraph and a list of rules: 'Don't cheat!', 'Apply yourself!', and 'Have fun!'.

Search

Sign In Register

InClass Prediction Competition

2020.AI.중간고사.문제5
유방암 악성/양성 종양 예측 문제

31 teams · 4 months ago

Overview Data Notebooks Discussion Leaderboard **Rules** Join Competition

By clicking on the "I understand and accept" button, you indicate that you agree to be bound with the rules outlined below.

I Understand and Accept

Rules

This is a page where you can include rules that participants must accept before joining. You may wish to include rules like:

- Don't cheat!
- Apply yourself!
- Have fun!

Data 사용법

- Data 탭에는 문제 해결을 위한 학습/테스트 데이터 그리고 정답 제출 템플릿 파일이 있음
- Description에는 제공되는 제공된 데이터의 설명이 있음
- 데이터 분석 후 정답 제출 템플릿에 정답을 적어 파일을 리더보드에 제출
<https://www.kaggle.com/c/2020-ai-exam-cancer/data> (둘러보기)

The screenshot shows the Kaggle competition interface for '2020-ai-exam-cancer'. The 'Data' tab is selected in the top navigation bar. Below it, the 'Data Description' section provides information about the datasets: '2020.AI.cancer-train.csv' (inspection results and malignancy), '2020.AI.cancer-test.csv' (diagnosis results), and '2020.AI.cancer-submission.csv' (submission template). A terminal window shows the command 'kaggle competitions download -c 2020-ai-exam-cancer'. The 'Data Explorer' on the left lists the files: '2020.AI.cancer-sample-sub...', '2020.AI.cancer-test.csv', and '2020.AI.cancer-train.csv'. The main panel displays the '2020.AI.cancer-sample-submission.csv' file (929 B) in 'Detail' view, showing a preview of the submission template with columns 'id' and 'diagnosis'.

Overview Data Notebooks Discussion Leaderboard Rules Team Host My Submissions Late Submission

Data Description Edit

제공되는 데이터의 설명

- 2020.AI.cancer-train.csv : 검사 결과와 악성(Benign tumor = 0) / 양성(Malignant tumor = 1) 여부
- 2020.AI.cancer-test.csv : diagnosis (진단결과: 악성/양성 여부) 를 제외한 나머지 train과 동일
- 2020.AI.cancer-submission.csv : submission 파일의 예시

>_ kaggle competitions download -c 2020-ai-exam-cancer

Data Explorer
145.24 KB

- 2020.AI.cancer-sample-sub...
- 2020.AI.cancer-test.csv
- 2020.AI.cancer-train.csv

< 2020.AI.cancer-sample-submission.csv (929 B)

Detail Compact Column 2 of 2 columns

About this file



- id: 환자 번호
- diagnosis: 악성/양성 여부

id	diagnosis
0	1
0	1
1	1
2	1
3	1

Notebook 사용법

- 일반 Jupyter Notebook과 동일
- Notebook을 통해 리더 보드에 결과 파일 제출 연동
- 학습용 GPU 사용 가능, 주 40시간 (추사 사용시 구글 클라우드 결제 필요)

<https://www.kaggle.com/hyeongjun0117/kernel301998fa4c> (둘러보기)

 **kernel301998fa4c**
Python notebook using data from 2020.AI.중간고사.문제5 · 17 views · 2mo ago ·  gpu

Best Submission ✓ Successful Submitted by Hyeongjun0117 2 months ago	Private Score 0.61403	Public Score 0.61403
---	--------------------------	-------------------------

Version 1 of 1

Notebook

Input (1)

Output

Execution Info

Log

Comments (0)

```
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

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files
# under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 5GB to the current directory (/kaggle/working/) that gets preserved
# as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of
# the current session

/kaggle/input/2020-ai-exam-cancer/2020.AI.cancer-sample-submission.csv
/kaggle/input/2020-ai-exam-cancer/2020.AI.cancer-test.csv
/kaggle/input/2020-ai-exam-cancer/2020.AI.cancer-train.csv
```