

みんなのKaggle講座

Section 1



イントロダクション



講師紹介



SAI-Lab株式会社 代表取締役

AI関連の教育、研究開発に従事

理学博士（物理学）

Udemyで数万人を指導 / 有名企業でAI研修を担当

著書に「はじめてのディープラーニング」など

我妻 幸長

Yukinaga Azuma

@yuky_az



コースの特徴

- **Kaggleを初歩から学ぶ**
 - Kaggleの始め方を、要点をおさえてコンパクトに学びます
- **体験を重視**
 - 理論よりも体験を重視し、Kaggleに親しみます
- **講座の対象**
 - Kaggleを通して機械学習、データ分析を学びたい方
 - ゲーム感覚でKaggleを楽しみたい方
 - データを扱った実績がほしい方

コースの注意点

- **初心者向けの講座**
 - 深い理論の解説や、Kaggle中上級者向けの講義は行いません
- **Pythonの解説はしません**
 - Python自体の解説はしませんが、Pythonを学ぶための教材を配布します
- **機械学習、データサイエンスの解説は最低限です**
 - これらを学びたい方には「みんなのデータサイエンス講座」をお勧めします



Udemyコース

みんなのデータサイエンス講座

-Python、Colab、Kaggleで基礎から学び親しむ
「データ」の世界-

講座の内容

Section1. Kaggleの概要

Section2. 機械学習とKaggle

Section3. 精度向上のためのテクニック

Section4. Titanicの先へ

今回の内容

1. イントロダクション
2. 講座の概要
3. Kaggleの概要
4. Kaggleの設定
5. 開発環境について
6. 演習

教材の紹介

- **Pythonの基礎:**
python_basic

講座の概要



Section1. Kaggleの概要



- Kaggleの概要と設定、開発環境について学びます

Section2. 機械学習とKaggle



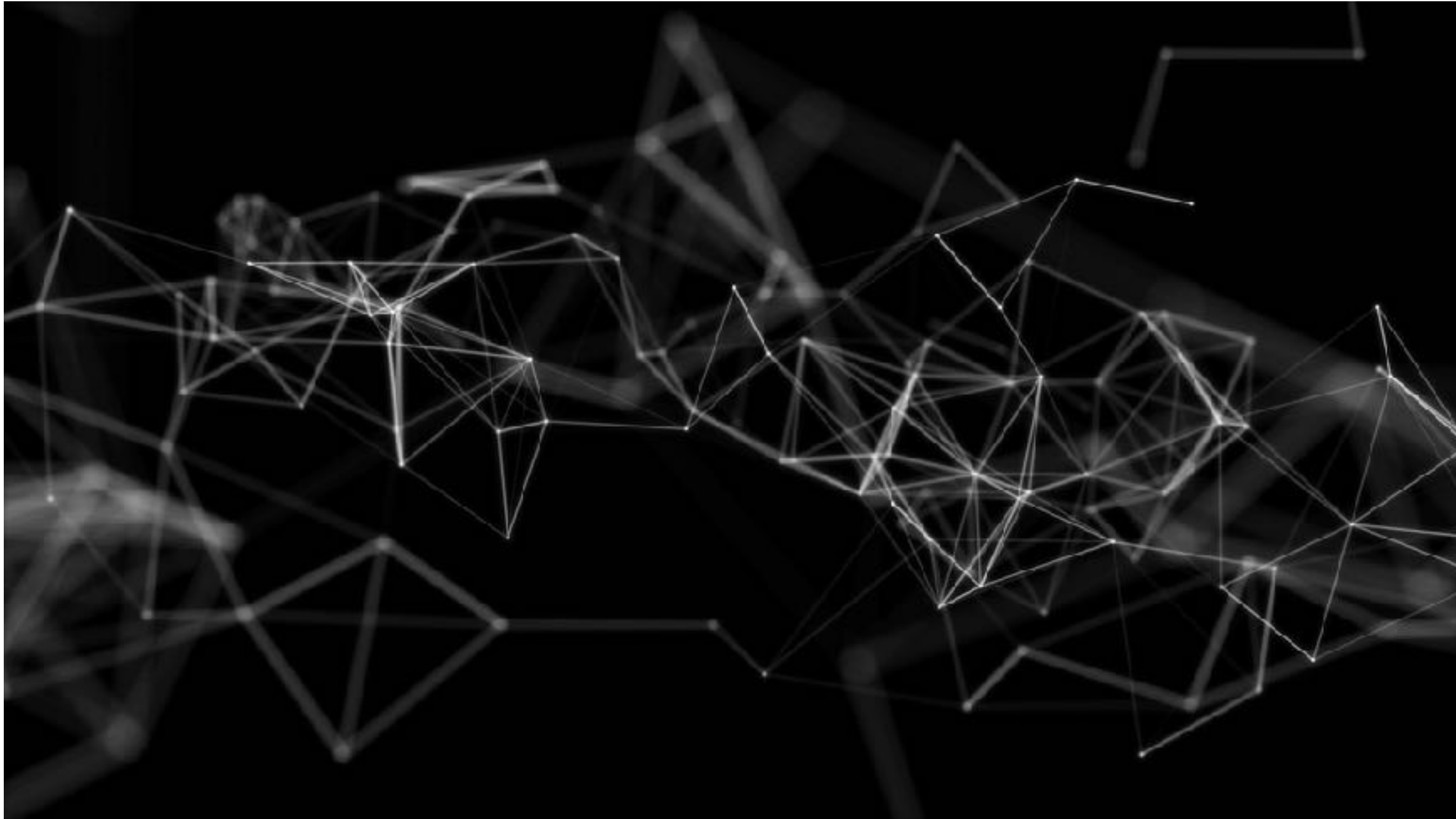
- 機械学習の概要と、Kaggleの課題への適用について学びます

Section3. 精度向上のためのテクニック



- 機械学習モデルを改善し、スコアを向上させるためのテクニックについて学びます

Section4. Titanicの先へ



- さらに複雑な課題に取り組み、様々な課題に取り組むための指針を学びます

Kaggleの概要



Kaggleとは？

- **Kaggle**

- 企業や研究機関などが提供するデータ分析の実践的な課題に対して、世界中の参加者が精度を競うプラットフォーム
- データ収集や環境構築の手間がかからない
- 優秀なスコアを残した参加者には、賞金やスコアに応じた称号が与えられる
- 多くの企業がKaggleのスコアをリクルーティングに利用

Kaggleのメリット

- 生のビッグデータ

→ 50万件のクレジットカード等の決済履歴、
6000万件の米小売大手の売上履歴、 etc...

- スキルの証明

→ 機械学習エンジニア、データサイエンティストの求人要件によく掲載

- 楽しくて実力がつく

→ 自分の順位がリアルタイムに表示、実績に応じた称号

Kaggleの人気

- Googleトレンド「Kaggle」

→ <https://trends.google.co.jp/trends/explore?date=today%205-y&q=kaggle>

DeepL翻訳の活用

- DeepL翻訳
 - DeepL社が提供する機械学習を利用した翻訳サービス
 - <https://www.deepl.com/translator>

Kaggleの課題の例

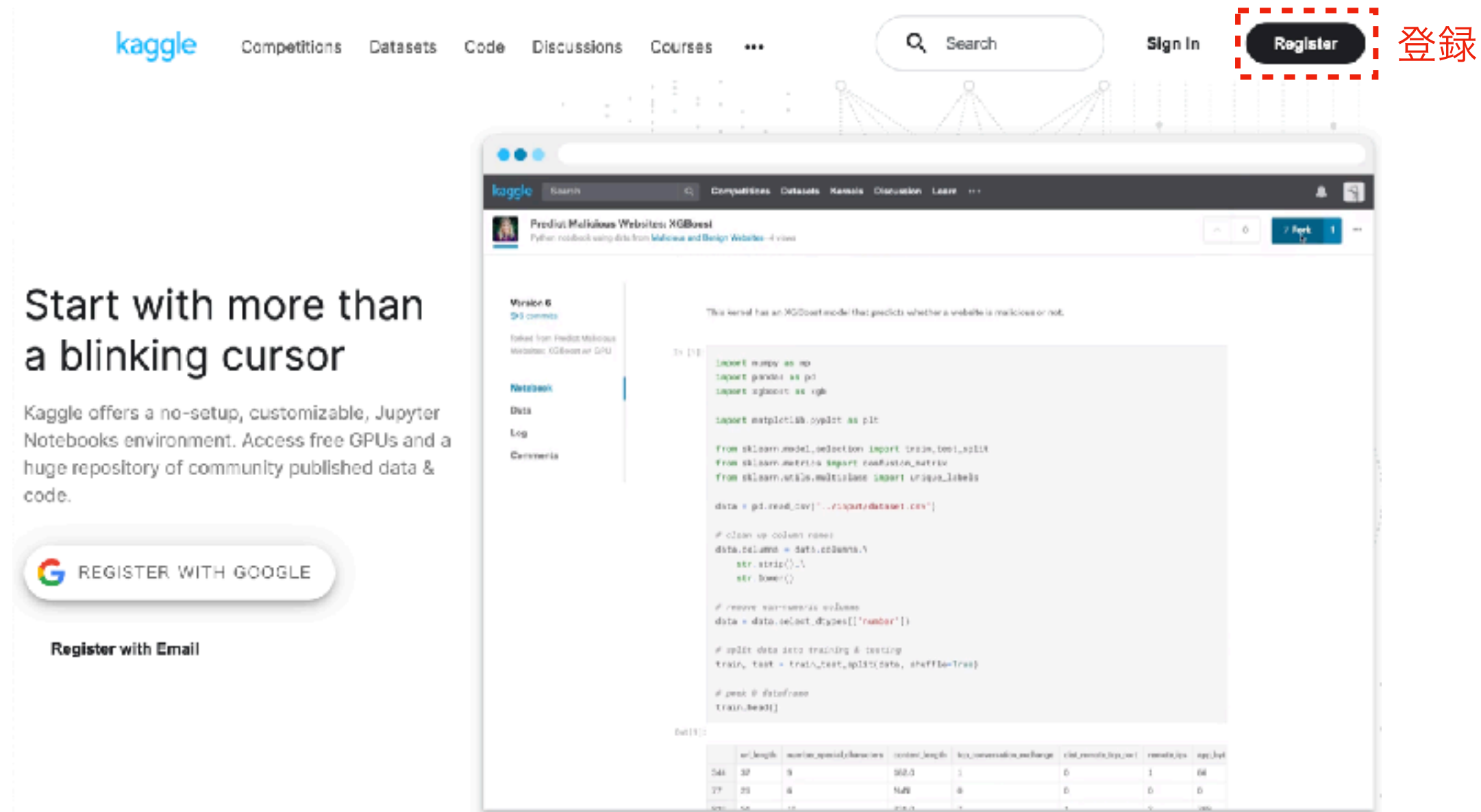
- **COVID-19 Open Research Dataset Challenge (CORD-19)**
→ <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>
- **RSNA STR Pulmonary Embolism Detection | Kaggle**
→ <https://www.kaggle.com/c/rsna-str-pulmonary-embolism-detection>
- **Cornell Birdcall Identification | Kaggle**
→ <https://www.kaggle.com/c/birdsong-recognition>

<https://www.kaggle.com/competitions>

Kaggleの設定




Kaggleのサイト



The screenshot displays the Kaggle website interface. At the top, the navigation bar includes links for Competitions, Datasets, Code, Discussions, and Courses, along with a search bar and a 'Register' button highlighted with a red dashed border and the Japanese text '登録' (Dokuroku). The main content area features a notebook titled 'Predict Malicious Websites XGBBoost' by a user named 'Predator Malicious Websites XGBBoost'. The notebook is in 'Version 6' and shows a Jupyter Notebook environment with Python code for data loading and preprocessing. The code includes imports for numpy, pandas, sklearn, and matplotlib, followed by data loading from a CSV file and initial data inspection using head().

Start with more than a blinking cursor

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.

 REGISTER WITH GOOGLE

Register with Email

```
In [1]:
import numpy as np
import pandas as pd
import sklearn as skl

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

data = pd.read_csv("../input/dataset.csv")

# clean up column names
data.columns = data.columns.str.strip().str.lower()

# remove unnecessary columns
data = data.select_dtypes(['number'])

# split data into training & testing
train, test = train_test_split(data, shuffle=True)

# peek @ dataframe
train.head()
```


	url_length	url_has_special_characters	content_length	host_consistency_percentage	url_protocol_https	response_size	app_type
548	32	9	552.0	1	0	1	66
77	25	4	NaN	0	0	0	0
300	54	10	200.0	1	1	0	66


<https://www.kaggle.com/>


Kaggleの登録


Sign In

Register

 Sign in with Google

 Sign in with your email

 Sign in with Facebook

 Sign in with Yahoo

No Account? [Create one.](#)

Kaggleのホーム画面

The screenshot shows the Kaggle homepage for user **yuky_az**. The interface includes a left sidebar with navigation icons, a top search bar, and a main content area. The main content area features a welcome message, a notebook feed, and a 'Novice' checklist.

Search Bar: Search

Welcome yuky_az


This is your personal newsfeed. As we learn what you like, we'll update you on cool Kaggle stuff that matches your interests. You can also choose to follow topics, notebooks, and people you want to keep up with.

TPS06 NN w/ discrete and continuous features
Python Notebook on [Tabular Playground Series - Jun 2021](#)
⌚ 25m to run | 📄 272 lines | 👁 16 views | 📊 6 visualizations

Novice

- ☐ Add a bio to your profile
- ☐ Add your location
- ☐ Add your occupation
- ☐ Add your organization
- ☐ SMS verify your account
- ☐ Run 1 notebook or script
- ☐ Make 1 competition or task submission
- ☐ Make 1 comment
- ☐ Cast 1 upvote

コンペの一覧

kaggle

+

Create

Home

Competitions

Datasets


Code


Discussions


Courses


More


Recently Viewed


Titanic - Machine Lear...


notebook1aaa229ef9


House Prices - Advanc...

Bike Sharing Demand

notebookf95abe5f90

View Active Events


Search



Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the [documentation](#) or learn about [Community competitions](#).

Host a CompetitionYour Work


Search competitions


Filters

All competitionsEnteredHostedFeaturedResearchGetting StartedPlaygroundAnalyticsAnalyticsCommunity

🕒Active Competitions

Hotness






TensorFlow - Help Protect the Great Barrier Reef

Detect crown-of-thorns starfish in under...

Research

Code Competition · 462 Teams




G-Research Crypto Forecasting

Use your ML expertise to predict real cry...

Featured


Code Competition · 969 Teams



NFL Big Data Bowl 2022

Help evaluate special teams performance

Analytics



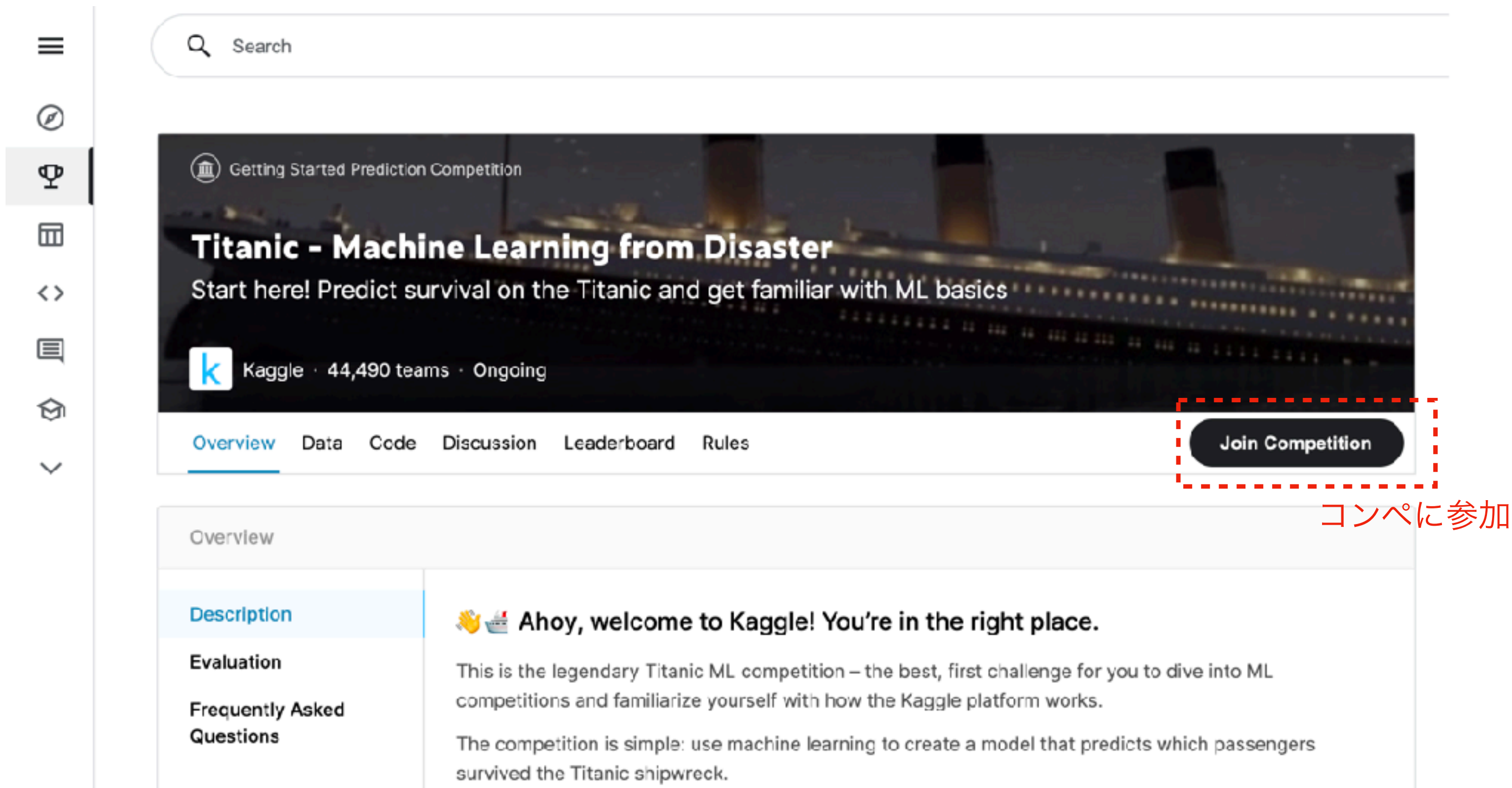
Sartorius - Cell Instance Segmentation

Detect single neuronal cells in microscopy...

Featured

Code Competition · 1318 Teams

Titanic - Machine Learning from Disaster




The image is a screenshot of the Kaggle website's page for the "Titanic - Machine Learning from Disaster" competition. On the left is a vertical navigation bar with icons for menu, home, competition, datasets, notebooks, discussion, and profile. The main content area has a search bar at the top. Below it is a large banner for the competition, featuring a night-time image of the Titanic ship. The banner text includes "Getting Started Prediction Competition", "Titanic - Machine Learning from Disaster", "Start here! Predict survival on the Titanic and get familiar with ML basics", the Kaggle logo, and "44,490 teams · Ongoing". Below the banner is a horizontal menu with "Overview", "Data", "Code", "Discussion", "Leaderboard", and "Rules". A "Join Competition" button is located in the bottom right of the banner area, highlighted with a red dashed box. Below the menu, the "Overview" section is active, showing a "Description" tab. The description text reads: "👋🚢 Ahoy, welcome to Kaggle! You're in the right place. This is the legendary Titanic ML competition – the best, first challenge for you to dive into ML competitions and familiarize yourself with how the Kaggle platform works. The competition is simple: use machine learning to create a model that predicts which passengers survived the Titanic shipwreck."

Search

Getting Started Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

 Kaggle · 44,490 teams · Ongoing

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#)

[Join Competition](#)

Overview

[Description](#)

Evaluation

Frequently Asked Questions

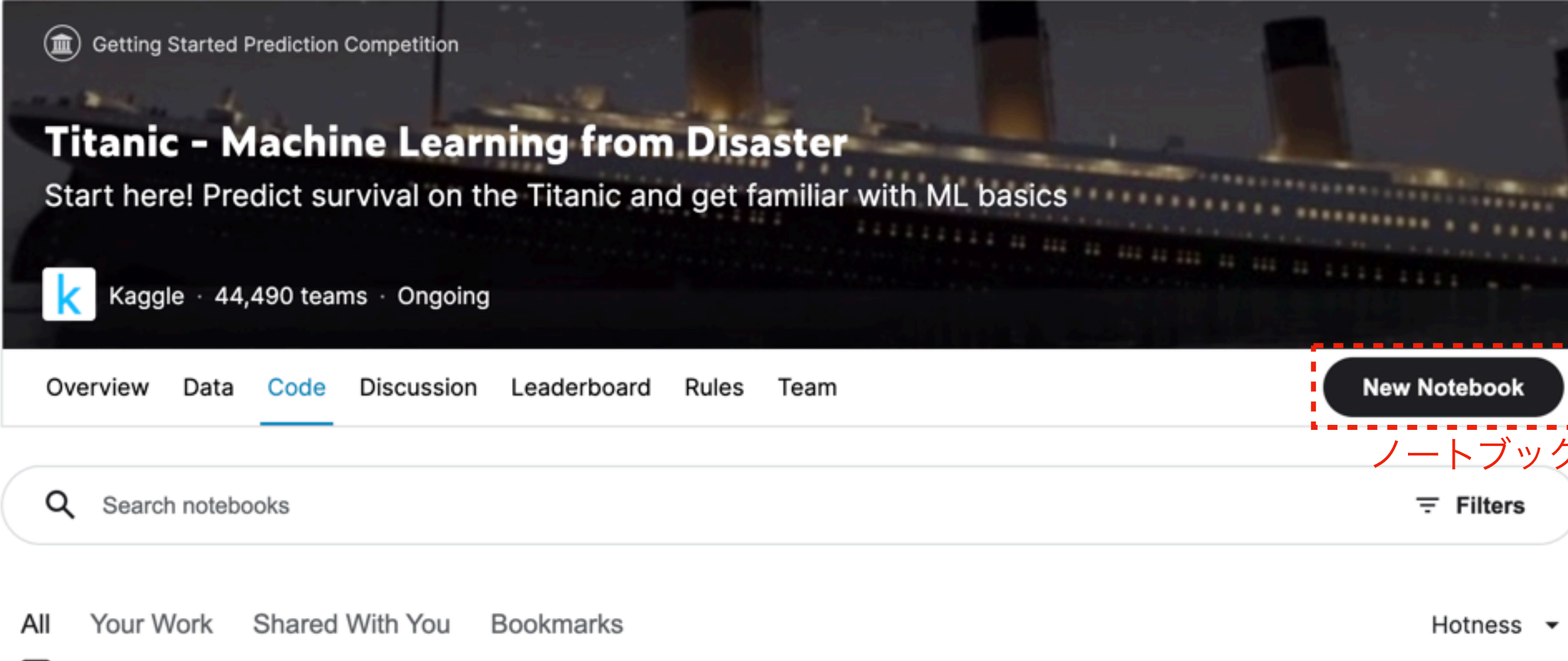
👋🚢 Ahoy, welcome to Kaggle! You're in the right place.

This is the legendary Titanic ML competition – the best, first challenge for you to dive into ML competitions and familiarize yourself with how the Kaggle platform works.

The competition is simple: use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.

コンペに参加

ノートブックの新規作成



The image shows the Kaggle interface for the 'Titanic - Machine Learning from Disaster' competition. The header includes the competition title and a description: 'Start here! Predict survival on the Titanic and get familiar with ML basics'. Below this, the Kaggle logo and statistics '44,490 teams · Ongoing' are displayed. A navigation bar contains links for 'Overview', 'Data', 'Code' (which is underlined), 'Discussion', 'Leaderboard', 'Rules', and 'Team'. On the right side of this bar, a 'New Notebook' button is highlighted with a red dashed border. Below the navigation bar is a search bar labeled 'Search notebooks' and a 'Filters' button. At the bottom, there are tabs for 'All', 'Your Work', 'Shared With You', and 'Bookmarks', along with a 'Hotness' dropdown menu.

Getting Started Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 44,490 teams · Ongoing

Overview Data Code Discussion Leaderboard Rules Team

New Notebook

Search notebooks Filters

All Your Work Shared With You Bookmarks Hotness ▼

ノートブックの新規作成

Titanicのデータ

OverviewDataCodeDiscussionLeaderboardRulesTeamMy SubmissionsSubmit Predictions...

Data Explorer93.08 kB

- gender_submission.csv
- test.csv
- train.csv

< gender_submission.csv (3.26 kB)

DetailCompactColumn2 of 2 columns

About this file

An example of what a submission file should look like.
These predictions assume only female passengers survive.

PassengerId# Survived

8921309

8920

8931

8940

0.00 - 0.10

Count: 266

提出データの例

予測結果の提出

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 14,507 teams · Ongoing

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#) [...](#)


```
> kaggle competitions submit -c titanic -f submission.csv -m "Message"
```

Make a submission for [yuky_az](#)

You have 10 submissions remaining today. This resets 18 hours from now (00:00 UTC).

Step 1

Upload submission file



File Format

Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions

We expect the solution file to have 418 prediction rows. This file should have a header row. Please see sample submission file on the [data page](#).

予測結果を提出

環境について

Google Colaboratoryとは？

- **Google Colaboratory**

- Googleが提供する、ブラウザでPythonを実行できる環境
- Googleアカウントで利用可能
- 基本的に無料
- 環境構築が簡単
- 共有が簡単
- etc...

<https://colab.research.google.com/>

コードセルとテキストセル

- コードセル
 - Pythonのコードを記述し、実行する
- テキストセル
 - 文章や数式を記述する

Google Colaboratoryの練習



Pythonの基礎



「python_basic」 フォルダ

演習



演習

• House Prices - Advanced Regression Techniques

結果の提出にトライしよう！

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Data Explorer

957.39 kB

- data_description.txt
- sample_submission.csv
- test.csv
- train.csv

提出データの例

< data_description.txt (13.37 kB)



70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

次回の内容

Section1. Kaggleの概要

 **Section2. 機械学習とKaggle**

Section3. 精度向上のためのテクニック

Section4. Titanicの先へ