

- [Урок 4](#)
 - [kaggle.com на примере titanic](#)
 - [Описание](#)
 - [Начало работы](#)
 - [1. Регистрация](#)
 - [2. Настройки](#)
 - [3. Конкурсы](#)
 - [4. Работа с данными](#)
 - [Отправка расчетов](#)

Урок 4

Ростислав Ромашин, группа 5128

kaggle.com на примере titanic

Описание

Kaggle — система организации конкурсов по исследованию данных, а также социальная сеть специалистов по обработке данных и машинному обучению.

Принадлежит корпорации Google.

Среда организована как публичная веб-платформа, на которой пользователи и организации могут публиковать наборы данных, исследовать и создавать модели, взаимодействовать с другими специалистами по данным и инженерами по машинному обучению, организовывать конкурсы по исследованию данных и участвовать в них.

В системе размещены наборы открытых данных, предоставляются облачные инструменты для обработки данных и машинного обучения.

Также реализованы обучающие ресурсы, имеется раздел для размещения вакансий работодателями, где тоже возможна организация конкурсов для отбора наилучших кандидатов.

Начало работы

Знакомство с порталом для дата саентистов <https://kaggle.com> рекомендуется осуществлять с помощью учебного курса [Titanic - Machine Learning from Disaster](#).

Конкурс **Titanic - Machine Learning from Disaster** на Kaggle предназначен для того, чтобы помочь новичкам в машинном обучении познакомиться с основами этой области.

В рамках конкурса участники должны создать модель, которая будет предсказывать, выживет ли пассажир на корабле Титаник, используя данные о пассажирах, такие как имя, пол, возраст и т.д.

Чтобы начать работу с Kaggle, нужно зарегистрироваться на сайте и присоединиться к конкурсу Titanic - Machine Learning from Disaster.

После регистрации можно загрузить набор данных, который содержит информацию о пассажирах Титаника, а также примеры кода, которые помогут начать работу с данными.

Для создания модели, которая будет предсказывать, выживет ли пассажир, нужно будет использовать методы машинного обучения, такие как логистическая регрессия, случайный лес или градиентный бустинг.

Можно использовать различные языки программирования (Python, R, SQL, Julia), и любые инструменты машинного обучения.

Для работы с данными, используется Kaggle Notebooks - это облачный инструмент, которые позволяют запускать код на Python, R и других языках программирования, а также визуализировать данные и создавать модели машинного обучения.

Основные этапы для участия в соревнованиях

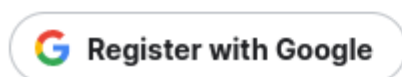
1. Регистрация
2. Настройка личного кабинета
3. Участие в конкурсе.
4. Работа с данными
5. Отправка своих расчетов.

1. Регистрация

Для полноценной регистрации рекомендую это делать через аккаунт Google или по адресу электронной почты с последующей привязкой аккаунта kaggle.com к gmail.com.

Level up with the largest AI & ML community

Join over 16M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies. Discover a huge repository of community-published models, data & code for your next project.

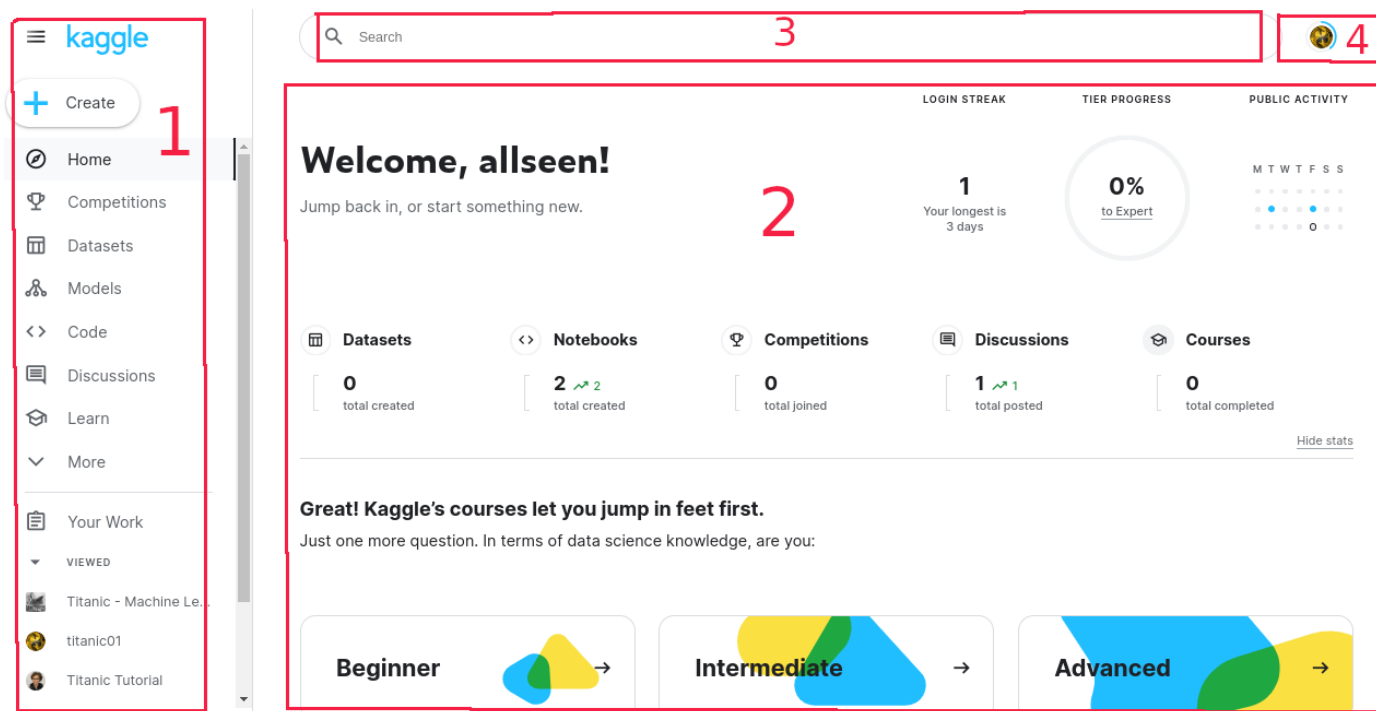


Register with Email

В процессе регистрации придет письмо с шестизначным кодом, который необходимо в вести в поле для подтверждения.

После подтверждения нужно авторизоваться на портале с помощью указанного логина и пароля, либо, в случае использования gmail, авторизация происходит автоматически.

Если процесс регистрации прошел успешно, то попадаем в личный кабинет.



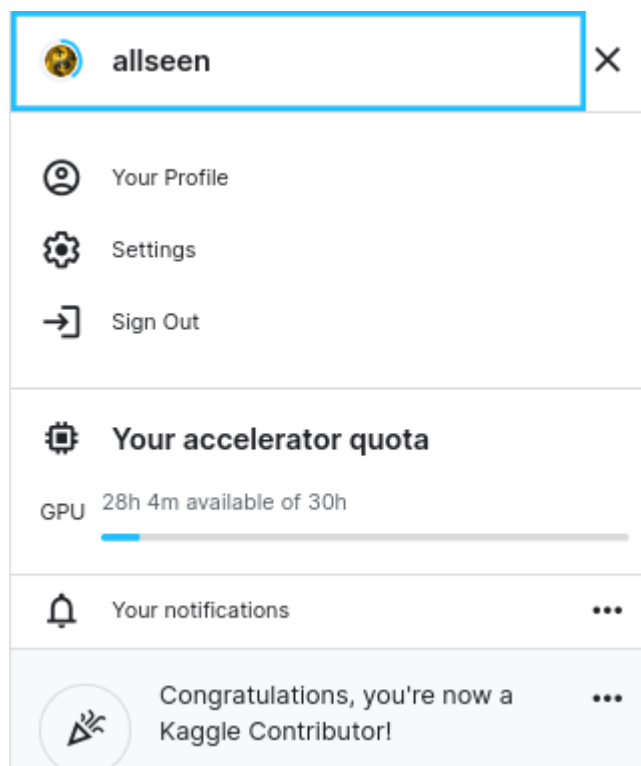
Главная страница личного кабинета (ЛК) состоит из 4 областей:

1. Панель меню
2. Рабочая область
3. Панель поиска
4. Кнопка настроек

На главной странице ЛК элементы панели меню, почти полностью дублируются элементами главной рабочей области. Главное меню неизменно при навигации, но его можно уменьшить, оставив только значки без подписей. Рабочая область меняется в зависимости от навигации. Панель поиска позволяет быстро найти на сайте нужный документ, конкурс либо пользователя.

2. Настройки

При нажатии на кнопку настроек появляется "меню настроек"



Меню настроек содержит следующие пункты:

1. Профиль (Your Profile)
2. Настройки (Settings)
3. Выход (Sign Out)
4. Размер и остаток вычислительной квоты (You accelerator quota)
5. Уведомления (Your notifications)

Рекомендую сразу зайти в пункт "Настройки" (Settings)

Settings

Account

Notifications

Phone verification

Verified

API

Using Kaggle's beta API, you can interact with Competitions and Datasets to download data, make submissions, and more via the command line. [Read the docs](#)

Create New Token

Expire Token

Quotas

Private Datasets

0 B / 107.37 GB

i

Private Models

0 B / 107.37 GB

i

GPU

01:55 / 30 hrs

i

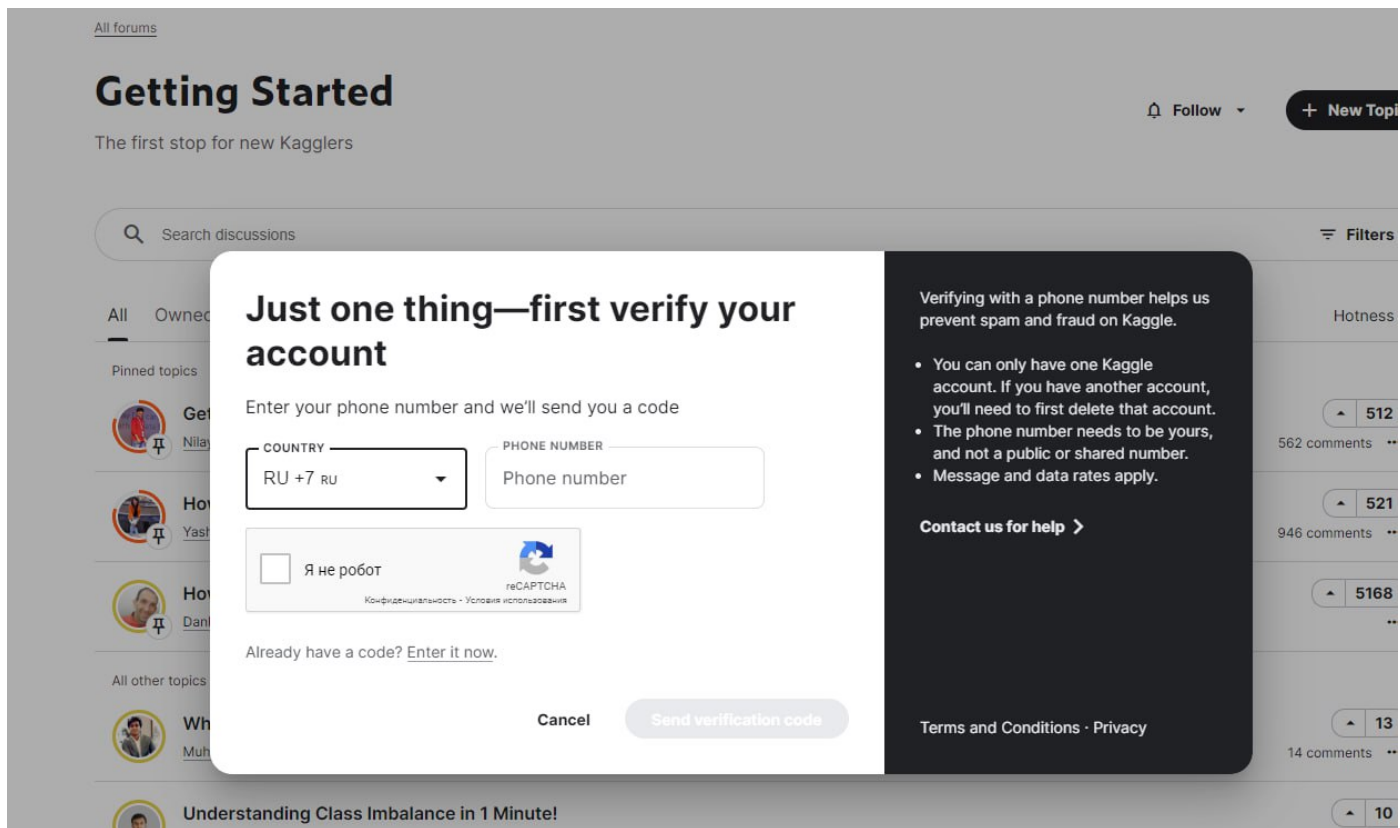
TPU

00:00 / 20 hrs

i

И осуществить верификацию телефонного номера (Phone verification).
Рекомендую верификацию проводить со смартфона, на котором установлена симка с регистрируемым номером, либо с компьютера подключенного к смартфону через раздачу интернета. Т.к. происходит проверка номера именно с того, с какого вы вышли в интернет. Иначе будет возникать ошибка о невозможности проверить номер.

На момент написания моей статьи, начало 2024 года, никаких ограничений на регистрацию с номеров РФ не было.



После успешной верификации номера открываются дополнительные
ВОЗМОЖНОСТИ:

Done! Your account is verified

You're now able to post to the [forums](#), make [public notebooks](#), turn on internet in notebooks, enter [featured competitions](#) and more.

Happy Kaggling!

1. Публикации на форумах kaggle.com
2. Для блокнотов становится доступен интернет
3. Блокноты можно делать публичными
4. Доступны особые соревнования
5. Возможность оставлять комментарии на портале
6. Хранилище для приватных моделей
7. Возможность роста по рангу (Tier)
8. В ноутбуках появляется поддержка GPU

✕ Save version

VERSION NAME

Version 3

VERSION TYPE

✓ Save & Run All (Commit)

Run a fresh copy of your notebook and save the output

SAVE OUTPUT

Never save output when creating a Quick Save

SAVE & RUN ALL WITH AN ACCELERATOR

Run without an accelerator for this session

Run without an accelerator for this session

Run with GPU for this session

Run with GPU for all sessions

Run with TPU for this session

3. Конкурсы

Для участия в конкретном конкурсе, например titanic, необходимо в панели поиска набрать его название

← titanic

C

<> Notebooks 65,926

← Comments 8,483

🗨 Topics 5,756

📁 Datasets 1,675

🏆 Competitions 265

👤 People 96

📖 Tutorials 1

Filter by

82,202 Results

Relevance ▾

DATE

☐ Last 90 days

2,544

☐ This week

208

☐ Today

15


VIEWED BY YOU

☐ Viewed

7

☐ Not Viewed


82,195

 **Titanic Tutorial**
Notebook · 2y ago · by Alexis Cook
Change it to something more descriptive, like ***Getting Started with Titanic***. !

2,544

12694


28,040 comments

 **Exploring Survival on the Titanic**
Notebook · 6y ago · by Meg Risdal
--- title: 'Exploring the Titanic Dataset' author: 'Megan L.'

15

3840

1,055 comments

 **Welcome to the Spaceship Titanic!**
Discussion Topic · 2y ago · by Ryan Holbrook
Welcome to the Spaceship Titanic!

7

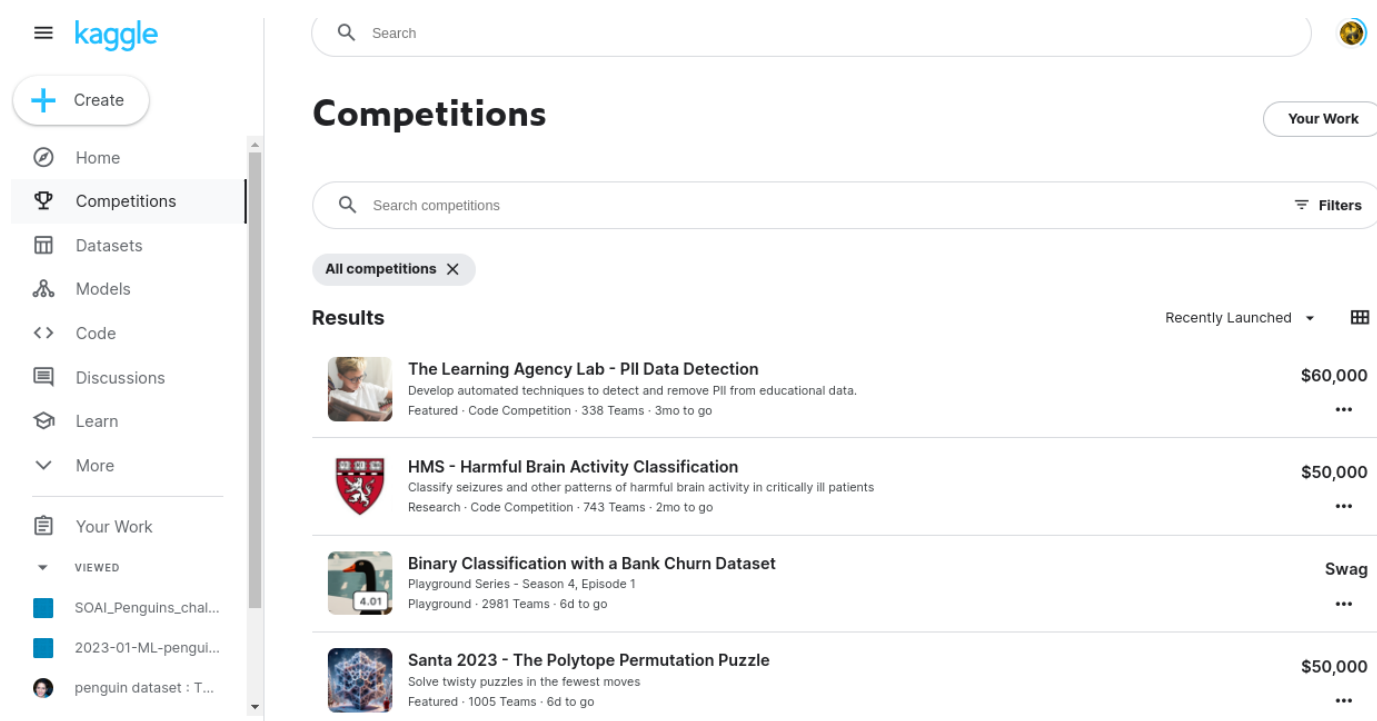
184

223 comments

Titanic Tutorial будет первым в списке. Рекомендую перейти по этой ссылке, т.к. это учебная инструкция, описывающая конкурс Titanic - Machine Learning from Disaster. Само же соревнование расположено по адресу <https://www.kaggle.com/competitions/titanic>

где URL указывает, что мы в разделе competitions (соревнования) - titanic (Титаник)

Чтобы отобразить список всех соревнований, в главном меню необходимо выбрать пункт Competitions (соревнования). Отобразится список всех соревнований, который можно отсортировать по дате, популярности и другим параметрам



The screenshot shows the Kaggle website's 'Competitions' section. On the left is a sidebar with a menu including 'Home', 'Competitions', 'Datasets', 'Models', 'Code', 'Discussions', 'Learn', 'More', and 'Your Work'. The 'Competitions' tab is active. The main content area has a search bar and a 'Filters' button. Below is a 'Results' section with a dropdown menu set to 'Recently Launched'. A list of competitions is shown, each with a thumbnail, title, description, prize pool, and a 'Join Competition' button. The competitions listed are:

Thumbnail	Competition Title	Description	Prize Pool	Join Button
	The Learning Agency Lab - PII Data Detection	Develop automated techniques to detect and remove PII from educational data. Featured · Code Competition · 338 Teams · 3mo to go	\$60,000	Join Competition
	HMS - Harmful Brain Activity Classification	Classify seizures and other patterns of harmful brain activity in critically ill patients. Research · Code Competition · 743 Teams · 2mo to go	\$50,000	Join Competition
	Binary Classification with a Bank Churn Dataset	Playground Series - Season 4, Episode 1. Playground · 2981 Teams · 6d to go	Swag	Join Competition
	Santa 2023 - The Polytope Permutation Puzzle	Solve twisty puzzles in the fewest moves. Featured · 1005 Teams · 6d to go	\$50,000	Join Competition

Для того, чтобы присоединиться к соревнованию, в нашем случае титанику, перейдя по ссылке <https://www.kaggle.com/competitions/titanic>

Необходимо присоединиться к конкурсу, нажав кнопку "Join Competition" справа сверху

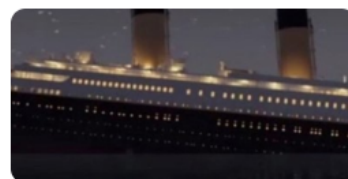


Структура страницы любого соревнования состоит из следующих закладок

1. Обзор (Overview)
2. Данные (Data)
3. Код (Code)
4. Модели (Models)
5. Дискуссии (Discussion)
6. Доска лидеров (Leaderboard)
7. Правила (Rules)
8. Команда (Team)
9. Отправки (Submissions)

Titanic - Machine Learning from Disaster

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



[Overview](#) [Data](#) [Code](#) [Models](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [Submissions](#)

Overview

Competition Host

Kaggle



Prizes & Awards

Knowledge

Does not award Points or Medals

∞ This competition runs indefinitely with a rolling leaderboard. [Learn more.](#)

4. Работа с данными

У всех соревнований во вкладке Данные (Data), находятся как минимум два CSV файла:

- test.csv (Тестовые данные)
- train.csv (Тренировочные данные) Конкретно у Титаника есть третий файл:
- gender_submission.csv

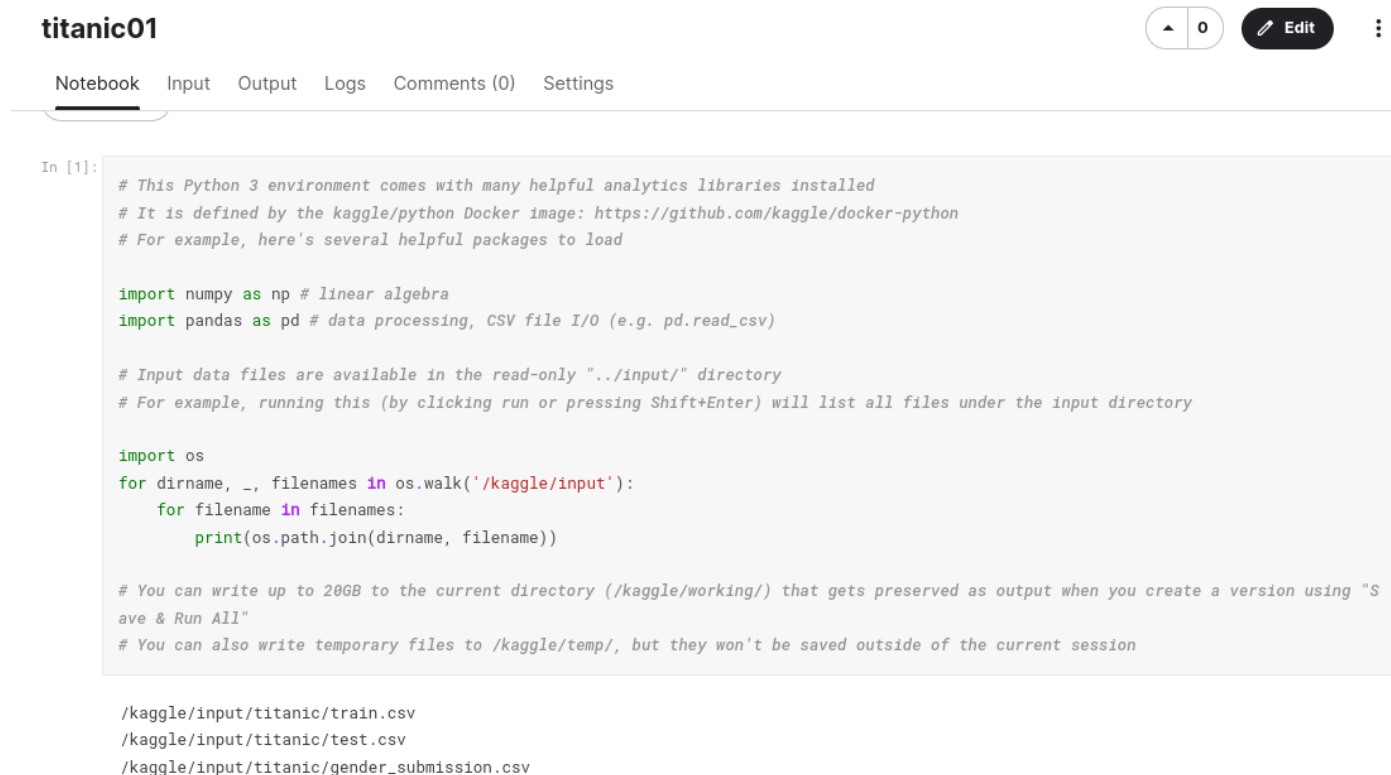
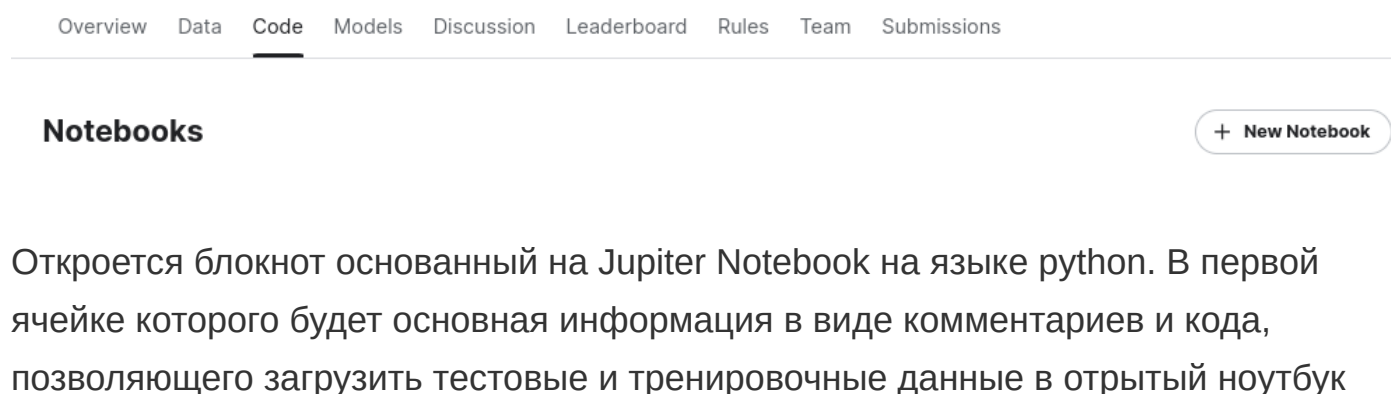
Основная задача скачать эти файлы, обработать и составить выходной файла submission.csv, который будет содержать всего два поля:

- Passengerid
- Survived

Побеждает тот, кто отправит submission с наиболее точными данными. Т.е. задача написать такой код, который на основе имеющихся данных (test.csv и train.csv),

любыми доступными в kaggle notebook средствами, разработает модель, предсказывающую реальные события.

Для написания кода, необходимо перейти во вкладку Код (Code) и нажать кнопку +New Notebook



На данной картинке видно, какие библиотеки используются и какие файлы загружены.

Можно приступить к обработке тестовых и тренировочных данных

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 20GB 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/titanic/train.csv
/kaggle/input/titanic/test.csv
/kaggle/input/titanic/gender_submission.csv
```

Далее подключаем библиотеку `sklearn.ensemble` для создания модели случайного леса (`RandomForestClassifier`). Модель обучается на данных `train_data`, которые содержат информацию о пассажирах Титаника, включая их класс (`Pclass`), пол (`Sex`), количество братьев и сестер/супругов (`SibSp`) и количество родителей/детей (`Parch`).

Далее, данные преобразуются в формат, который может быть использован моделью, с помощью метода `pd.get_dummies()`. Этот метод преобразует категориальные признаки в числовые, чтобы их можно было использовать в модели.

Затем модель обучается на преобразованных данных с использованием гиперпараметров `n_estimators=100`, `max_depth=5` и `random_state=1`. После обучения модель делает прогнозы на данных `test_data` и сохраняет их в файл `submission.csv`.

```
In [2]: train_data = pd.read_csv("/kaggle/input/titanic/train.csv")
train_data.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

```
In [3]: test_data = pd.read_csv("/kaggle/input/titanic/test.csv")
test_data.head()
```

Out[3]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

У ноутбука справа наверху есть кнопка Сохранить Версию (Save Version), конкурс Титаник позволяет сохранить до 50 вариантов ноутбука

При нажатии на кнопку Сохранить версию, появится подменю, в котором можно выбрать вычислительные мощности, как я об этом писал ранее.

Отправка расчетов

После успешного завершения расчетов, создается файл submission.csv и при нажатии на кнопку Submit Prediction (Отправка предсказаний), расчеты отправляются на проверку.

Во вкладке Отправки (Submissions), можно посмотреть свои отправки и их набранный балл, т.е. процент совпадений с оригинальными данными

Titanic - Machine Learning from Disaster

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



Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

All Successful Errors

Recent ▾

Submission and Description

Public Score ⓘ



titanic01 - Version 1

Complete · 10d ago

0.77511

В своем профиле можно увидеть, количество участников и какое место занято в конкретном соревновании

Active Completed Hosted Community Bookmarks

Default ▾



Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started · 16020 Teams · Ongoing

8646/16020



...