

PYTHON AI WORKSHOP

參與 Kaggle 競賽，讓你從 0 秒變數據高手！

About me



CHRIS 李嘉桓

資料科學 | 機器學習 | 商業分析 | 線上教學

7年開發經驗

- 國泰金控資深資料科學分析師
客群經營團隊領導
數據工程團隊領導
- 銓威數位整合行銷有限公司技術顧問
- 上下新媒體整合行銷有限公司技術顧問
- 凱羅斯健康有限公司資料科學顧問

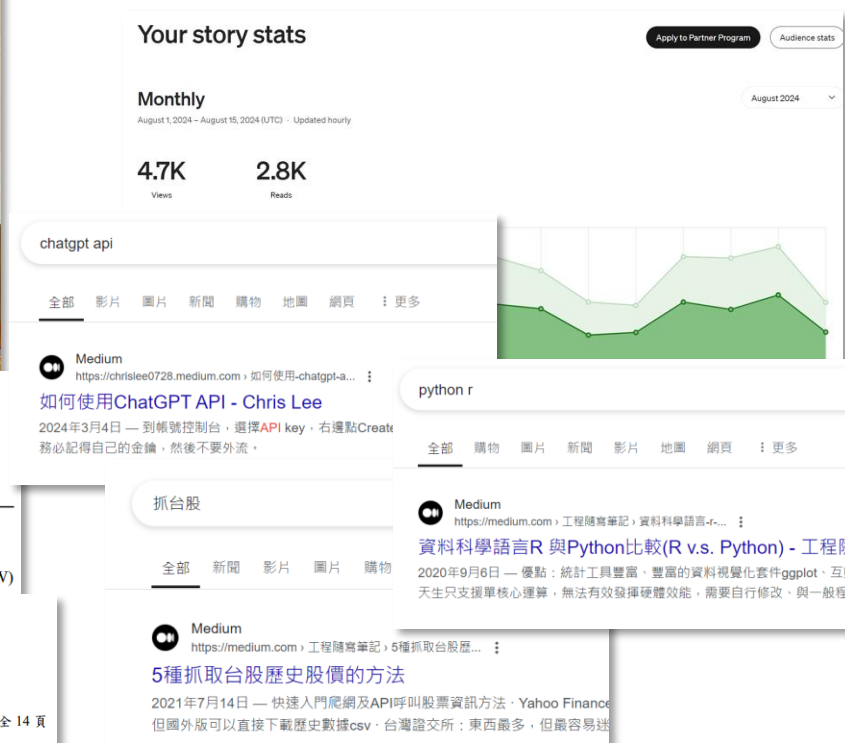


【19】中華民國 【12】發明公開公報 (A)
【11】公開編號：202109430 申請實體審查：有
【43】公開日：中華民國 110 (2021) 年 03 月 01 日
【51】Int. Cl. : G06Q40/08 (2012.01)
【54】發明名稱：核保風險評估系統及方法
【21】申請案號：108129349 【22】申請日：中華民國 108 (2019) 年 08 月 16 日
【72】發明人：李嘉桓 (TW) LI, CHIA-HUAN；劉學汝 (TW) LIU, SHIUE-RU；廖奕翔 (TW)

【19】中華民國 【12】專利公報 (U)
【11】證書號數：M656609
【45】公告日：中華民國 113 (2024) 年 06 月 11 日
【51】Int. Cl. : G06Q40/08 (2012.01) G06Q40/00 (2023.01)
G16H10/00 (2018.01)
新型 全 14 頁

【54】名稱：個人化保險之保障額度推薦系統及遠端伺服器
【21】申請案號：113200530 【22】申請日：中華民國 113 (2024) 年 01 月 16 日
【72】新型創作人：李嘉桓 (TW)；黃三騰 (TW)；李育瑩 (TW)；鄭安庭 (TW)；李奕慧

<https://medium.com/@chrislee0728>



AI 時代來臨 你準備好了嗎

快訊 美股道瓊早盤大漲500點 強勁零售數據減輕衰退恐懼、台積電A... 21:45

udn / 產經 / 財經焦點

聽新聞 0:00 / 0:00

經發會拍板 4年培育20萬AI人才

2024-07-19 00:23 聯合報 / 記者黃婉婷、林海 / 台北報導

+ 保險

分享 2 分享

行政院「經濟發展委員會」昨天舉行首場會議，由行政院長卓榮泰主持，會中拍板兆元投資國家發展計畫、打造台灣成為亞洲資產管理中心、國家人才競爭躍升方案等三項計畫，鎖定三點二七兆壽險資金投資公建設施，目標二〇二八年培育廿萬名AI人才，吸引「數位牧民」來台。

經發會執行秘書、國發會主委劉鏡清表示，全球初估有三千五百萬名數位牧民，希望跟外交部協調規畫推出「數位遊牧簽證」並迅速啟動，把他們拉進台灣，先達到促進觀光、消費的效果，如果他們愛上台灣就會留下來，希望進來十萬人就可以留下一萬人。

搶AI人才！ 53%不限科系、薪水較傳統高15%

張蕙纖 王興堂

2024年6月15日



AI浪潮來襲，企業徵才也搭上車！今天(15日)有人力銀行舉辦徵才博覽會，多家科技大廠都來搶才，有大學生就說，AI是未來趨勢，因此還沒畢業但也要先來看看，或是有本來就在科技業的民眾，想轉職看看有拿到輝達大單的大廠，人力銀行也統計，今年5月釋出3萬個AI相關工作機會，53%不限科系，薪水也比傳統職務高15%。



你的競爭優勢



金字名校學歷

台清交成政...
DS、CS、EE...



豐富實務經驗

AI、ML Project
Data Lake、Cloud...



顯赫競賽成績

Kaggle、T-Brain
AI CUP、Github Star...



強大關係網路

你好 我爸董事長

我該怎麼做 快速提升競爭力



刷學歷

洗血統洗上去
念個在職專班也不錯



Side Project

做幾個很屌的專案
放進履歷給面試官問



自我進修

課程報起來
從基礎開始打底

有工作經驗

- 公司遇到什麼商業問題，怎麼分析
- 使用 AI 或 ML 演算法解決什麼問題
- 替公司帶來多少價值及收益

無工作經驗

- 想辦法呈現技術專業及分析能力

kaggle

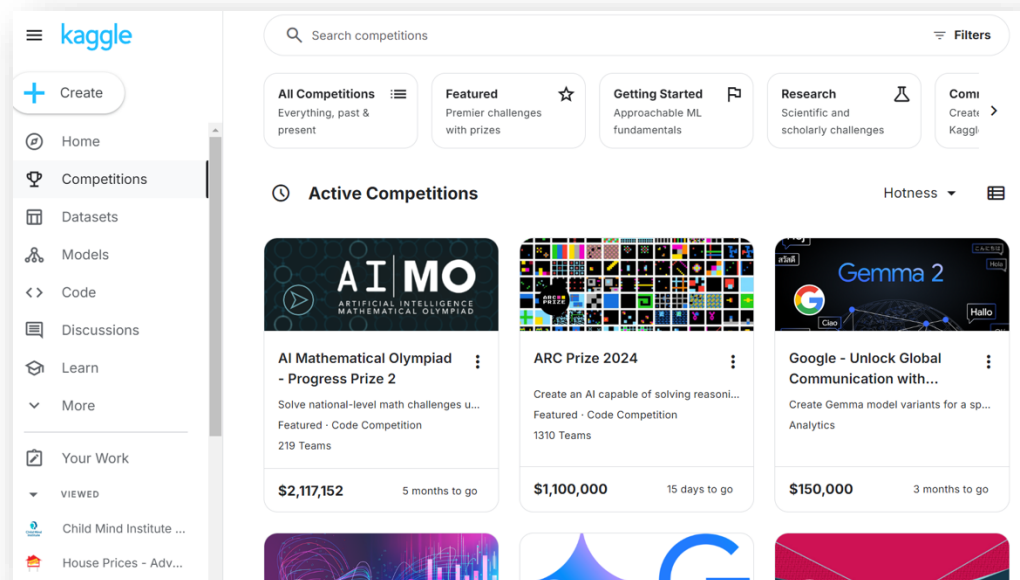
TBrain AI實戰吧



Google 專門提供數據學習及競賽的平台
提供大量免費的 Dataset，同時也有許多學習資源



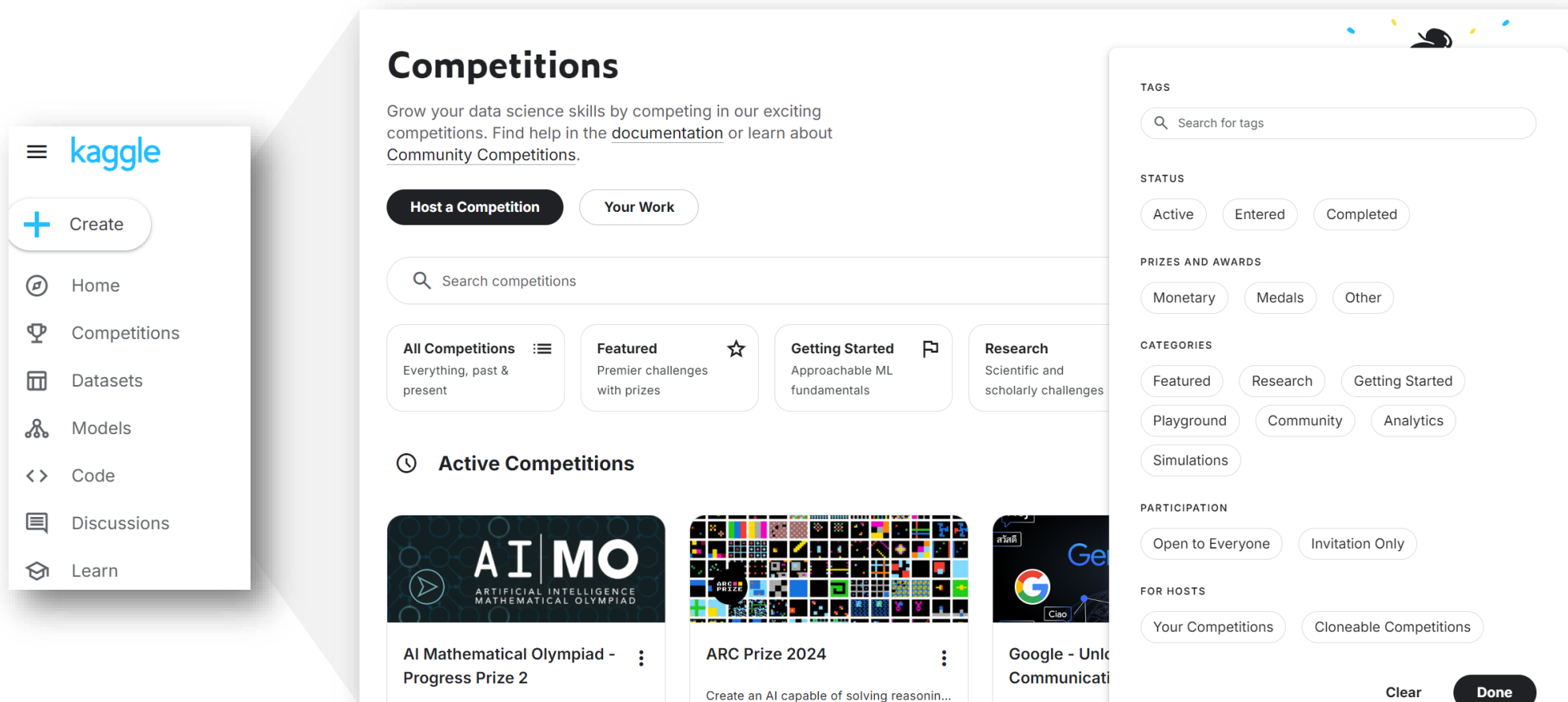
趨勢科技提供專門數據競賽的平台
每年中研院及玉山金控都舉辦數據競賽



一起來玩 Kaggle 吧！

競賽 (Competitions)

包含過去及進行中的競賽，每個競賽會提供訓練資料、測試資料。提交測試資料預測結果，可以在排行榜(Leaderboard)看到排名
在排行榜拿到前幾名是有獎金的



數據集 (Datasets)

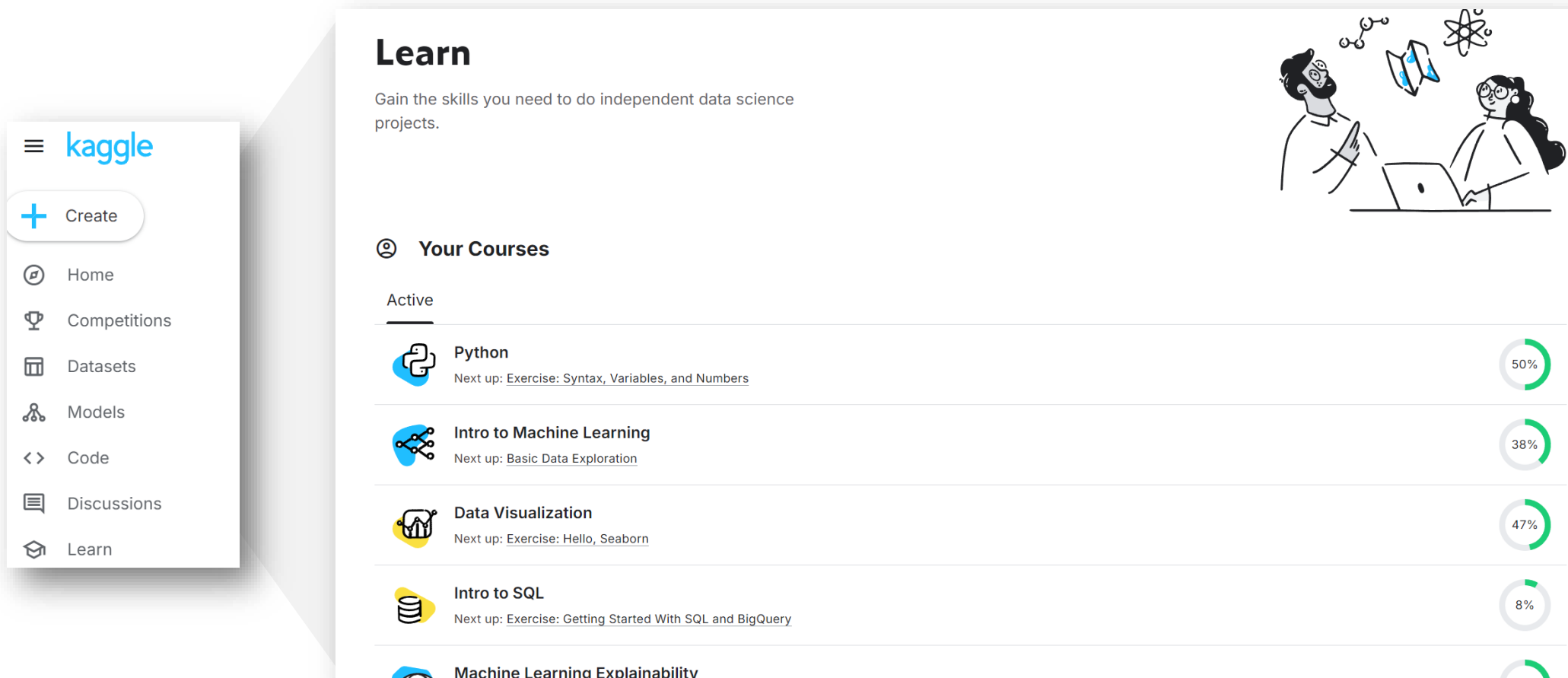
包含各式各樣的資料，如金融、生醫、文本、影像等結構化與非結構化資料

台灣部分學校老師喜歡把數據上傳Kaggle，讓學生來交作業，因此練技術苦無資料都可以來這邊挖寶

The image shows the Kaggle Datasets homepage. On the left is a sidebar with the Kaggle logo and navigation links: Home, Competitions, Datasets, Models, Code, Discussions, and Learn. The main content area is titled "Datasets" and includes a search bar, a "New Dataset" button, and a "Your Work" button. Below these are category filters like "All datasets", "Computer Science", "Education", "Classification", "Computer Vision", "NLP", and "Data Visualization". A "Trending Datasets" section displays three featured datasets: "Big Mart Sales Prediction" by Elaheh Kazemian, "League of Legends Worlds 2024 - Swiss Stage Stats" by lunovian, and "Indian Personal Finance Spending Habits" by Shriyash Jagtap. On the right, a filter overlay is visible, allowing users to refine their search by tags, file size, file types (CSV, JSON, SQLite, BigQuery), licenses (Creative Commons, GPL, Open Database, Other), usability rating (8.00 or higher, 9.00 or higher, 10.00), and highly voted for categories (Learning, Research, Application, Well-documented, Well-maintained, Clean data, Original, High-quality notebooks, LLM Fine-Tuning).

學習 (Learn)

系統化的教學，讓你從數據小白變成資料科學家，唯一缺點是全英文，可以使用瀏覽器翻譯功能來輔助

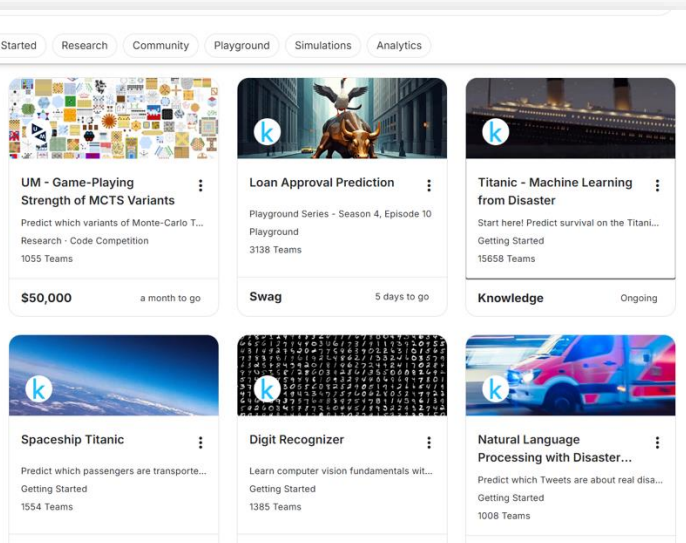



The screenshot displays the Kaggle Learn interface. On the left is a sidebar menu with the Kaggle logo and navigation options: Create, Home, Competitions, Datasets, Models, Code, Discussions, and Learn. The main content area is titled 'Learn' and includes the text 'Gain the skills you need to do independent data science projects.' Below this is a section for 'Your Courses' with a filter for 'Active' courses. A list of five courses is shown, each with an icon, title, next step, and a progress indicator.

Course	Next up	Progress
Python	Exercise: Syntax, Variables, and Numbers	50%
Intro to Machine Learning	Basic Data Exploration	38%
Data Visualization	Exercise: Hello, Seaborn	47%
Intro to SQL	Exercise: Getting Started With SQL and BigQuery	8%
Machine Learning Explainability		

選擇 Active Competitions

活躍競賽內的資料集，通常都有豐富的程式碼範例及討論，建議初學者可以從這邊著手，首推鐵達尼號資料集在 code 分頁可以找到其他人上傳的完整程式碼



 KAGGLE · GETTING STARTED PREDICTION COMPETITION · ONGOING

Submit Prediction ...

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


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

Description

  **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.

Competition Host

Kaggle 

Prizes & Awards

Knowledge
Does not award Points or Medals

Participation

1,328,216 Entrants
15,745 Participants
15,656 Teams
57,109 Submissions

Tags

Binary Classification Tabular

為什麼說 Kaggle 最適合新手入門

實務上的數據分析 SOP 都能在 Kaggle code 上顯現



數據工程

原始資料的清洗及加工
特徵工程及數據結構處理



商業分析

資料探勘及視覺化工程
通常稱探索性分析(EDA)



資料科學

模型建置及演算法優化
AI 專家的核心技能



重要觀念

即使 Kaggle 排名再高，與實務上的分析還是有不小差距

Kaggle 畢竟是競賽使用的數據，多數資料品質是好的

鐵達尼號資料集

	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

實務上可能的資料集 (充滿各種缺失值及離群值)

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

前置作業

1. 一台能上網的電腦

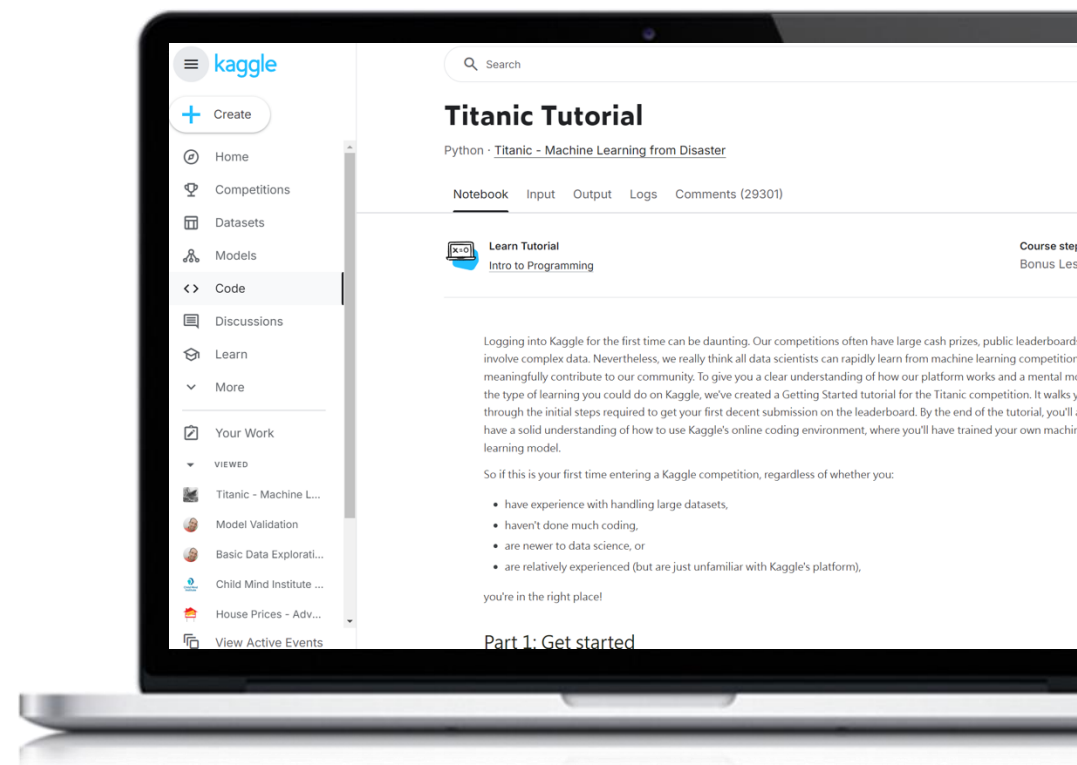
推薦使用 Chrome 瀏覽器，在網路良好的環境下一起作業

2. 使用 Colab 線上開發

課程使用 Colab 作為示範，無需在本地安裝 Python 環境

3. 註冊 Kaggle 帳號

需要註冊才能上傳模型預測結果，可以用 Google 帳號綁定



本課程的所有程式碼可以在Github瀏覽
https://github.com/chrisleeelearning/20240901_kaggle





Colab 可讓你在瀏覽器編寫及執行 Python 程式碼

不必進行任何設定 # 免付費使用 GPU # 輕鬆共用

讓我們開始吧！

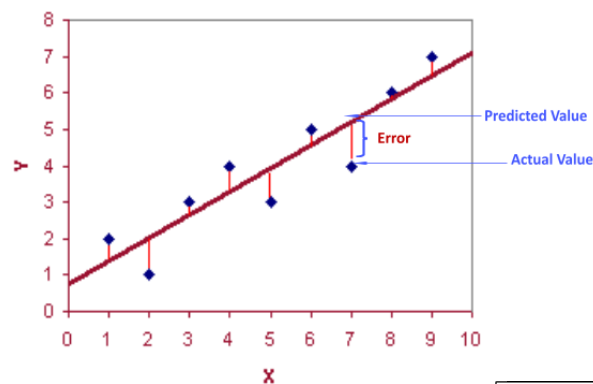
House Prices - Advanced Regression Techniques

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



場景貼近實務

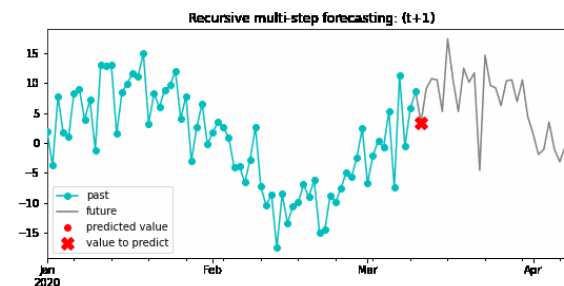
每人都想知道房價走勢



$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

衡量指標易懂

RMSE 即使數學不好也能理解

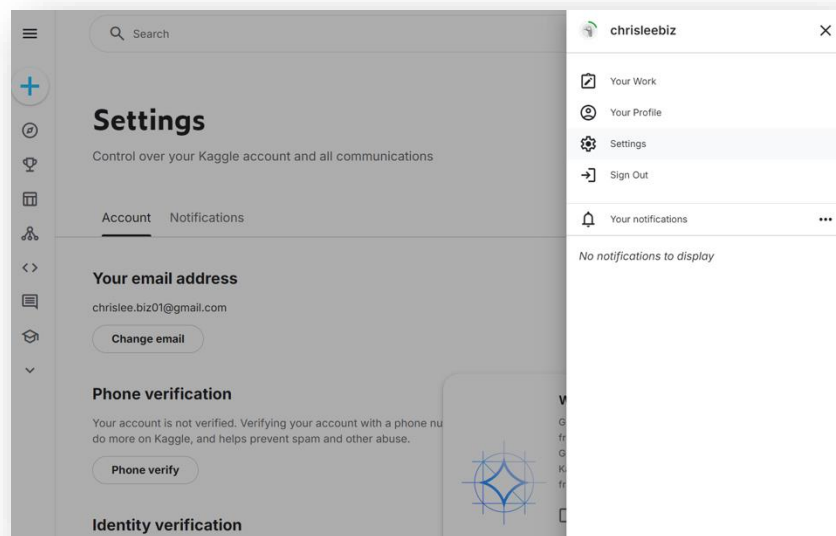


技能延續性

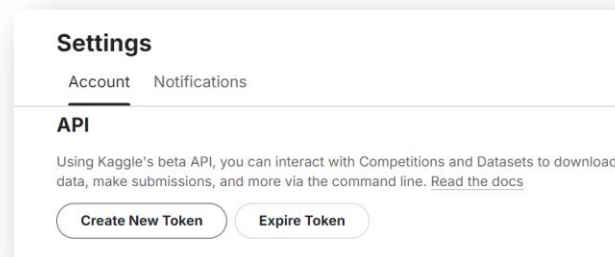
數值預測可以用在其他多元場景
如匯率、股價等

下載資料到 Colab (方法一)

使用 Kaggle 的 API 將資料直接抓到 Colab



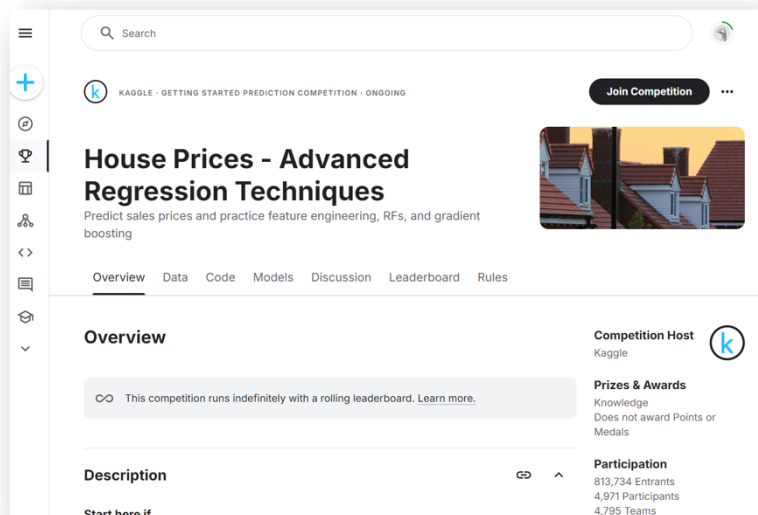
Settings 建立API token



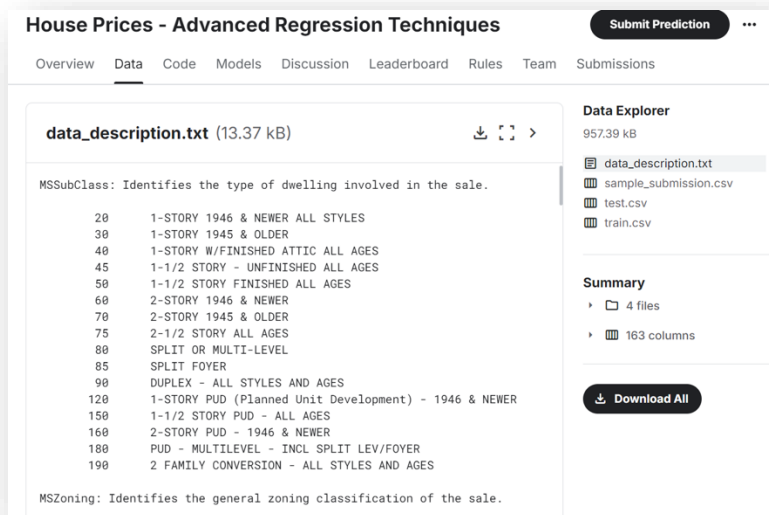
將 json 丟到 Colab
使用語法下載檔案 (詳見官方文檔)

下載資料到 Colab (方法二)

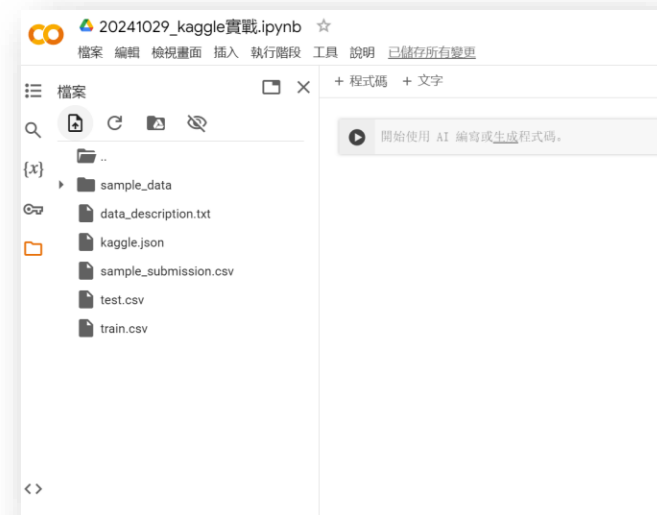
在網站上下載所有資料，自行上傳到 Colab



Join Competition 加入競賽

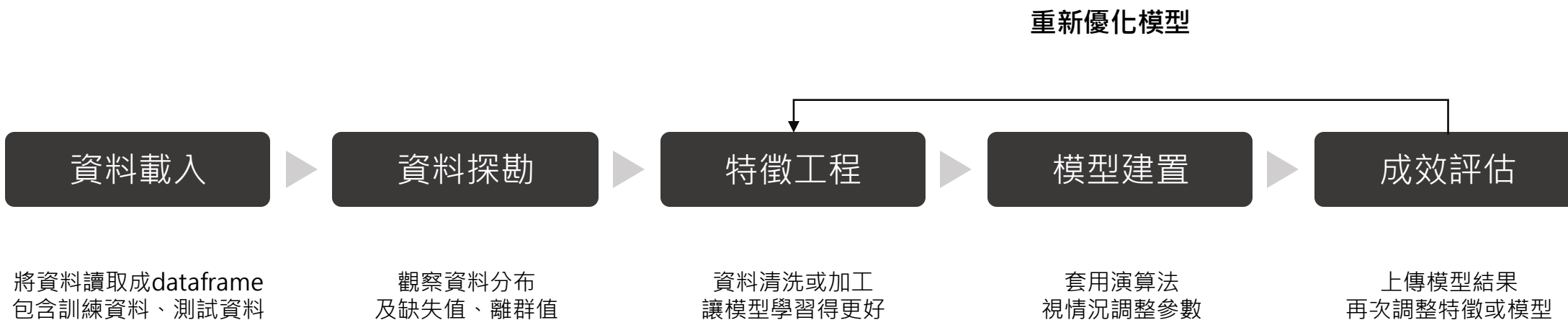


下載資料到本機



上傳到 Colab

數據競賽標準 SOP



Thank You