

Lab 2: Tabular Dataset

Predicting Heart Disease with Deep Learning

11220IEEM513600

Deep Learning for Industrial Applications

2024/03/07 Taco



Tabular Dataset

Background Survey for Deep Learning — Edited

View Zoom Add Category Pivot Table Insert Table Chart Text Shape Media Comment Collaborate Format Organize

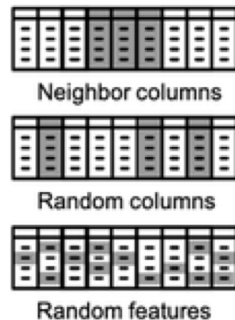
Sheet 1

時間戳記	4. Department and year	5. Your undergraduate major	6. A short Introduction to your research field or research interest
2024/02/22 3:37:01 下午 GMT+8	IEEM master	Statistics	Deep learning
2024/02/22 3:38:38 下午 GMT+8	工工碩一	工工	最佳化
2024/02/22 3:40:57 下午 GMT+8	工工碩一	清大工工	目前在研究分析脈波預測糖尿病
2024/02/22 3:41:04 下午 GMT+8	IEEM 2025	IEEM	AI applied to production line improvement
2024/02/22 3:41:30 下午 GMT+8	工工碩一	運輸與物流管理學系	整數賽局
2024/02/22 3:42:12 下午 GMT+8	工工系碩一	工工系	數學模型建模
2024/02/22 3:42:38 下午 GMT+8	工工所碩一	運輸科學系	賽局理論
2024/02/22 3:44:33 下午 GMT+8	工業工程與工業管理學系 碩一	企業管理學系	破權相關的預測
2024/02/22 3:44:40 下午 GMT+8	工工碩一	工工	Machine Learning, Computer Vision
2024/02/22 3:47:30 下午 GMT+8	112工工	工工	Machine learning
2024/02/22 3:50:06 下午 GMT+8	交大工管所 碩一	交大工工系	Abnomaly detection, RUL prediction, imbalanced learning
2024/02/22 4:38:11 下午 GMT+8	工工碩一	工工	用演算法協助廠商搜索潛在上下游客戶
2024/02/22 4:44:49 下午 GMT+8	工工系碩一	工工系	Operations research, machine learning
2024/02/22 5:13:35 下午 GMT+8	工工碩一	IEEM	Machine learning 、computer vision
2024/02/22 5:21:05 下午 GMT+8	交大工業工程與管理學系 碩士班 一年級	交大 運輸與物流管理學系 輔系 工工系	學習如何利用深度強化學習的相關知識，應用於工業上的排程問題
2024/02/22 5:30:42 下午 GMT+8	iPHD 博二	Engineering Science	Digital Transformation and Sustainability in Taiwan Small & Medium-sized Enterprises
2024/02/22 6:23:27 下午 GMT+8	工工所博士班一年級	工業工程與工程管理	Incremental learning
2024/02/22 6:32:27 下午 GMT+8	工工所人因組碩一	工業設計	目前在實驗室的研究領域是跟調度分配有關。 前一陣子讀了一篇叫車出行的訂單調度分配研究，裡面提到他們是使用深度學習在實現

Features

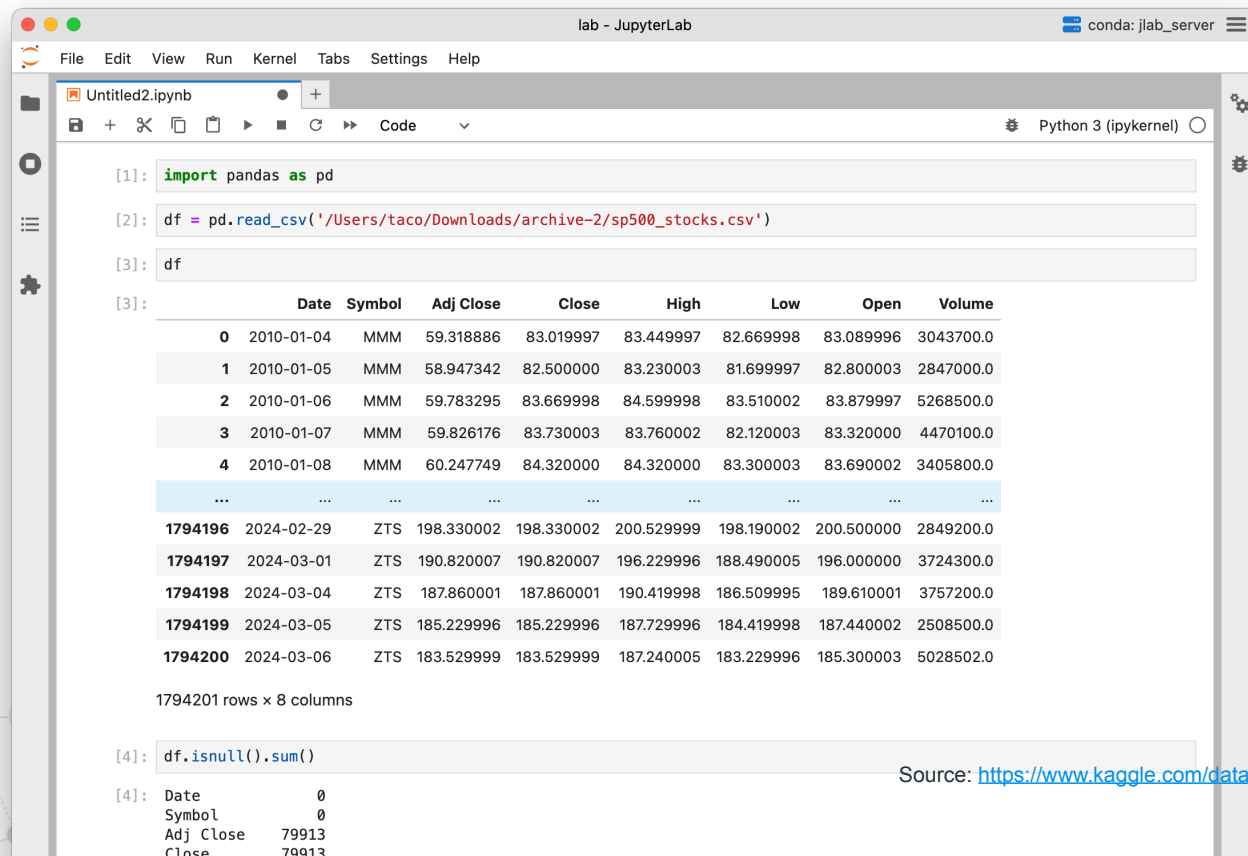
The tabular data commonly used in many fields such as healthcare, advertisement, finance, and law.

- ▶ **Structured Format:** Tabular data is organized into rows and columns, with rows representing records and columns representing variables.
- ▶ **Heterogeneous Data Types:** Columns can contain different data types, including numerical, categorical, datetime, and text.
- ▶ **Feature Relationships:** Features within tabular datasets may exhibit complex relationships and dependencies that are crucial for model accuracy.
- ▶ **Missing Values:** Tabular datasets often contain missing values, necessitating strategies like imputation or omission for effective data analysis.



Source: <https://openreview.net/pdf?id=vrhNQ7aYSdr>

Use Pandas to Process



The screenshot shows a JupyterLab window titled "lab - JupyterLab" with a "conda: jlab_server" environment. The active notebook is "Untitled2.ipynb". The code cells contain the following:

```
[1]: import pandas as pd
```

```
[2]: df = pd.read_csv('/Users/taco/Downloads/archive-2/sp500_stocks.csv')
```

```
[3]: df
```

The output of the third cell is a preview of the DataFrame:

	Date	Symbol	Adj Close	Close	High	Low	Open	Volume
0	2010-01-04	MMM	59.318886	83.019997	83.449997	82.669998	83.089996	3043700.0
1	2010-01-05	MMM	58.947342	82.500000	83.230003	81.699997	82.800003	2847000.0
2	2010-01-06	MMM	59.783295	83.669998	84.599998	83.510002	83.879997	5268500.0
3	2010-01-07	MMM	59.826176	83.730003	83.760002	82.120003	83.320000	4470100.0
4	2010-01-08	MMM	60.247749	84.320000	84.320000	83.300003	83.690002	3405800.0
...
1794196	2024-02-29	ZTS	198.330002	198.330002	200.529999	198.190002	200.500000	2849200.0
1794197	2024-03-01	ZTS	190.820007	190.820007	196.229996	188.490005	196.000000	3724300.0
1794198	2024-03-04	ZTS	187.860001	187.860001	190.419998	186.509995	189.610001	3757200.0
1794199	2024-03-05	ZTS	185.229996	185.229996	187.729996	184.419998	187.440002	2508500.0
1794200	2024-03-06	ZTS	183.529999	183.529999	187.240005	183.229996	185.300003	5028502.0

1794201 rows x 8 columns

```
[4]: df.isnull().sum()
```

The output of the fourth cell is:

```
Date      0
Symbol    0
Adj Close 79913
Close     79913
```

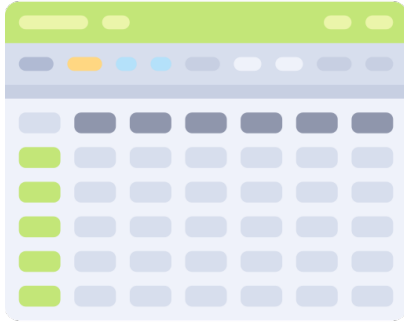
Source: <https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks>

Use Pandas to Process

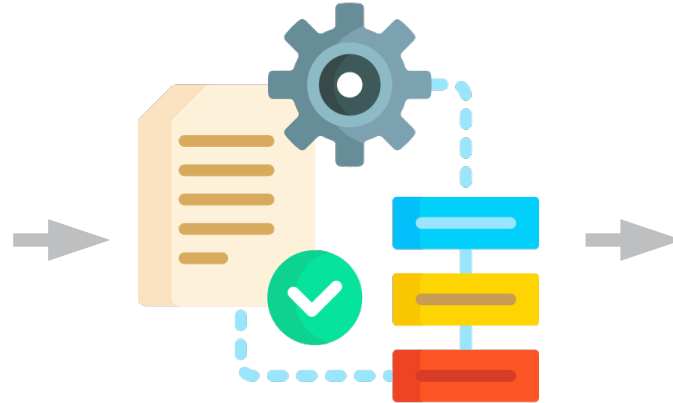
11693	NaN	NaN	NaN	2019	1052407	2	27497030	NaN	b'\x01'	2018-11-26 22:23:04	...
11694	2018-11-26 20:30:00	64.0	Premier League	1295	Burnley FC	513	Newcastle United FC	NaN	NaN	NaN	...
11695	NaN	NaN	NaN	2019	1052407	2	27497030	NaN	b'\x01'	2018-11-26 22:23:05	...
11696	2018-11-26 20:30:00	64.0	Premier League	1295	Burnley FC	513	Newcastle United FC	NaN	NaN	NaN	...
11697	NaN	NaN	NaN	2019	1052407	2	27497030	NaN	b'\x01'	2018-11-26 22:23:06	...

Must be careful for the NaN values!

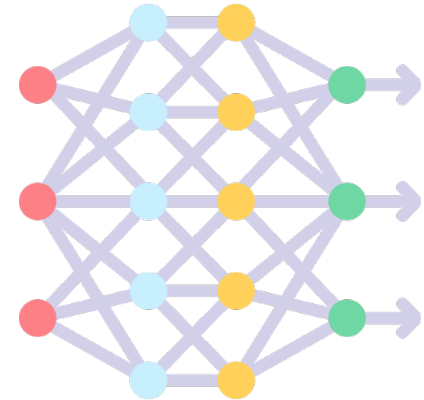
3 Steps for Model Development



Know Your Data




Process Data



Build a Model

Homework 2

- ▶ **Deadline:** 2024.03.21 23:59
- ▶ **Github:** Create a "hw2" folder in your repository, "NTHU_2024_DLIA_HW", containing "hw2.ipynb" and "hw2.pdf". Ensure that you run your code and all outputs are saved within the .ipynb files.
- ▶ **EEclass:** You are required to submit only the GitHub link of your Homework 2. Do not upload files directly to EEclass.
- ▶  **Important:** Make sure your commit is timestamped before the deadline. Late submissions might not be graded or could incur a penalty. Only the GitHub link is required on NTHU EEclass.

CODING TIME!!

