 國立雲林科技大學資訊管理系

機器學習-作業三

Department of Information Management

National Yunlin University of Science & Technology Assignment

香蕉資料集和尺寸三資料集分析

Banana Datasets and Sizes3 Dataset Analysis

楊欣蓓、陳怡君、鄭皓名、陳郁云

指導老師：許中川 博士

Advisor: Chung-Chian Hsu, Ph.D.

 中華民國112年12月

December 2023

**摘要**

在成人資料集中我們主要的預測目標是hours-per-week這項數值，也設定了幾個不同的模型參數組合，來比較哪個參數組合可以讓我們績效最好，在經過不同的測試之後，我們發現增加模型的深度確實可以讓模型學得更好，但是如果batch size設的太小的話，就算模型深度較深，你的績效也會跟淺層的神經網路差不多，另外我們也發現early stopping 設定的patient也是一個關鍵，如果你設定的太大，雖然模型在訓練和驗證集上看起來是最好的，但是在測試集資料上反而是績效最差的，所以可能是有overfitting的情況發生，在learning rate方面也可以觀察到當你設的越小loss就會很大，一旦你一直往上加到超過了某個數值之後，loss又開始慢慢往上升了，藉由不同的嘗試來調整參數，進而達到更好的績效。

本研究想要透過葡萄酒品質資料集，來分別預測酒的品質(類別預測)以及酒成分中總二氧化硫含量多寡(數值預測)，實驗過程嘗試了不同模型參數組合並觀察各自的績效優劣。在這兩個實驗中，發現增加模型深度確實可以縮小generalization gap，但當深度較深時，可能會降低泛化能力，以本實驗為例，加入BatchNormalization後，確實有提高泛化能力，並提高訓練績效。

關鍵字：hours-per-week、early stopping、patient、overfitting、learning rate、generalization gap、BatchNormalization

**一、緒論**

1. **研究動機**

在本課程中學習了許多分群演算法，本研究主要是在眾多的群聚分析中進行分析，因其資料集的資料雜亂，所以需要預先進行整理才能進行分析，而選擇Banana、size的主要原因，這個資料集包含不同的群集，使得本研究可以使用不同的分群算法（例如 K-means、階層式分群、DBSCAN）來進行實驗，通過比較這些分群算法表現，可以評估它們在處理不同形狀和密度的分群時的績效。本研究發現 Banana、size這兩個資料集的二維結構使得群集可視化相對容易，並且可以使用不同的分群演算法將數據點劃分到群集中，以及可視化這些群集的分佈，這有助於理解不同算法的群集結果，由於這個資料集包含不同形狀和 密度的群集，所以可以嘗試調整不同群集算法的參數，以查看參數對結果的影響如何。 如果這個資料集與某個實際應用相關，例如顧客購買行為、產品銷售模式等，使用 K-means、階層式分群、DBSCAN 可以幫助本研究識別和理解樣本之間的相似性和差異性，並且可以為業務決策提供有價值的信息。 總上所述，本研究進行 K-means、階層式分群、DBSCAN 群集算法分析是為了從數據中挖掘信息，理解潛在的結構和模式，並且比較不同算法在這個場景下的效能表現。

1. **研究目的**

本研究想以 Banana 資料集和 size 資料集以各種分群來進行比較 K-means、 階層式分群、DBSCAN 三種分群所花費的時間，將此分群以 SSE、Accuracy、 Entropy 為此衡量指標，並劃分出這三種分群所呈現的結果，其中本研究想以 DBSCAN 分群來進行不同的參數設定，此研究透過各種參數設定來進行比較後， 藉由比較後的參數設定上分析出以 Banana 資料集和 size 資料集中的試驗以達成 Banana 資料集和 size 資料集這兩個資料集的最佳設定值，也了解每種分群在不 同情況下的優缺點及此限制，在碰到資料集中存在雜訊時，該如何使用各種分群 來解決資料集的問題，並透過 K-means、階層式分群、DBSCAN 分群來分析資料 集的 SSE、Accuracy、Entropy 的結果。

**二、實驗方法**

1. **實作說明**

本研究使用K-means、階層式分群、DBSCAN三種分群演算法進行

本研究對成人資料集使用KNN、SVM、RandomForest、XGBoost進行工作時長的預測分析；波士頓房價資料集則是使用XGBoost、K-fold cross validation進行房價預測。在進行模型訓練前，做了資料前處理，其中包括刪除重複資料、將名目資料做轉換和數值資料做正規化等，也嘗試的修改模型裡面的架構，像是調整超參數，觀察修改前後的績效有何變化。

1. **操作說明**

本研究執行環境皆採用Python3.10.10，以Visual Studio Code作為開發工具，利用Tenserflow及Keras建構前饋式神經網路，並使用Pandas、Numpy、Scikit-learn、Matplotlib等函式庫來讀取資料、資料前處理以及將模型訓練的績效以視覺化的形式呈現，透過資料正規化、One-hot encoder來得到乾淨的資料，在訓練神經網路時，使用Early Stopping等正則化技術來降低模型過度擬合，以利模型訓練時可以得到較佳的泛化程度。

**三、實驗設計**

1. **資料集**

名稱: banana資料集

原始資料集: 32561筆(訓練資料)+16281筆(測試資料)=48842筆

資料前處理後: 32537筆(訓練資料)+16276筆(測試資料)=48813筆

**表1** 成人資料集欄位介紹

|  |  |  |
| --- | --- | --- |
| 欄位 | 屬性 | 內容 |
| 0 | age | continuous |
| 1 | workclass | Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. |
| 2 | fnlwgt | continuous |
| 3 | education | Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. |
| 4 | education-num | continuous |
| 5 | marital-status | Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse |
| 6 | occupation | Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces |
| 7 | relationship | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried |
| 8 | race | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black |
| 9 | sex | Female, Male |
| 10 | capital-gain | continuous |
| 11 | capital-loss | continuous |
| 12 | hours-per-week | continuous |
| 13 | native-country | United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands |

名稱: size資料集

原始資料筆數: 6497

前處理後的資料筆數: 6497

**表2** 波士頓房價資料集欄位介紹

|  |  |  |
| --- | --- | --- |
| 欄位 | 屬性 | 內容 |
| 0 | crim | per capita crime rate by town |
| 1 | zn | proportion of residential land zoned for lots over 25,000 sq.ft. |
| 2 | indus | proportion of non-retail business acres per town. |
| 3 | chas | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). |
| 4 | ***nox*** | nitrogen oxides concentration (parts per 10 million). |
| 5 | rm | average number of rooms per dwelling. |
| 6 | age | proportion of owner-occupied units built prior to 1940. |
| 7 | dis | weighted mean of distances to five Boston employment centres. |
| 8 | rad | index of accessibility to radial highways. |
| 9 | tax | full-value property-tax rate per \$10,000. |
| 10 | ptratio | pupil-teacher ratio by town |
| 11 | black | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town |
| 12 | lstat | lower status of the population (percent) |
| 13 | medv | median value of owner-occupied homes in \$1000s |

1. **資料前處理**
2. **波士頓房價資料集**

* 資料前處理：
* 在預測波士頓房價實驗中，將欄位quality作為category label，在本研究中quality視為無大小之分，並對該欄位進行One-Hot Encoder(獨熱編碼)，將七個類別分別轉換成七個維度呈現。
* 數值資料的欄位

('fixed acidity', 'volatile acidity', 'citric acid','residual sugar','chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol')透過高斯分布將這些數值型態的資料做標準化。

* 資料分割：

首先，將資料集切割成85%訓練資料、15%測試資料，其中**訓練資料**約18%作為驗證資料，剩下的作為訓練資料。

**表3**顯示部分資料處理後的波士頓房價資料集

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 資料  特徵 | No.1 | No.2 | No.3 | No.4 | No.5 | No.6 |
| **CRIM** | -0.42233 | -0.41987 | -0.41987 | -0.41928 | -0.41928 | -0.41957 |
| **ZN** | 0.296443 | -0.48964 | -0.48964 | -0.48964 | -0.48964 | -0.48964 |
| **INDUS** | -1.31101 | -0.59977 | -0.59977 | -1.33043 | -1.33043 | -1.33043 |
| .  .  .  . | | | | | | |
| **B** | 0.441052 | 0.441052 | 0.396427 | 0.416163 | 0.441052 | 0.410571 |
| **LSTAT** | -1.10415 | -0.51035 | -1.23975 | -1.39533 | 0 | -1.07132 |
| **MEDV** | 24 | 21.6 | 34.7 | 33.4 | 36.2 | 28.7 |

1. **成人資料集**

* 資料前處理：
  + 將資料欄位為「?」的部分，替代成「Nan」，判斷「Nan」的欄位為名目資料或數值資料，前者取眾數做替補，而後者則使用平均值做填補。
  + 刪除意思相近的特徵欄位，如：'education'、'education-num' 取其一。
  + 刪除對於預測年收入較無高度關聯性的欄位 ' fnlwgt '。
  + 刪除資料集空白的筆數。
  + 名目資料的欄位，如：'workclass','education','maritalstatus','occupation',

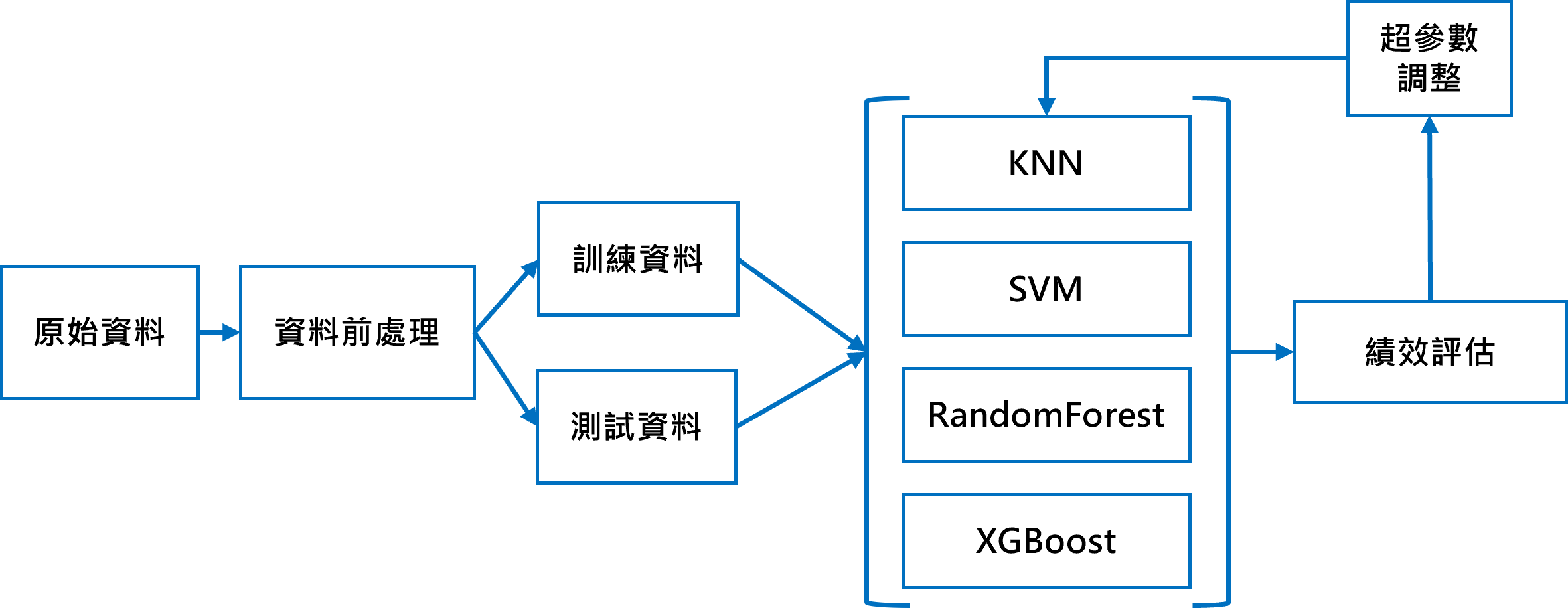
'relationship','race','sex','native-country','class'，透過One hot encoding轉成數值資料。

* + 將名目資料採獨熱編碼方式處理，經One-hot encoding後，訓練集的欄位比測試集多出了'native-country\_Holand-Netherlands'欄位，故在測試集新增該欄位，並將其值都設為0，讓兩個資料集欄位數相同。
  + 將有順序性的欄位資料採用Label Encoding技術，本次資料針對欄位Income做處理，將>50K設為1；<=50K設為0。
  + 刪除資料集中資料重複的筆數。
  + 數值資料使用Normalization技術(z-score)，將欄位 'age','education-num',capital-gain','capital-loss'做處理，降低模型發生Overfitting的狀況。

**表4** 部分資料處理後的成人資料集

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 資料  特徵 | No.0 | No.1 | No.2 | No.3 | No.4 | No.5 |
| **Age** | 0.783411 | -0.09358 | 1.002659 | -0.8244 | -0.16666 | 0.710329 |
| **capital-**  **gain** | 0.148292 | -0.14598 | -0.14598 | -0.14598 | -0.14598 | -0.14598 |
| **capital-loss** | -0.22971 | -0.22971 | -0.22971 | -0.22971 | -0.22971 | -0.22971 |
| **.**  **.**  **.**  **.** | | | | | | |
| **sex\_**  **Female** | 0 | 0 | 0 | 0 | 1 | 1 |
| **sex\_**  **Male** | 1 | 1 | 1 | 1 | 0 | 0 |
| **hours-per-week** | 13 | 40 | 40 | 40 | 40 | 16 |

1. **實驗設計**
2. **成人資料集**

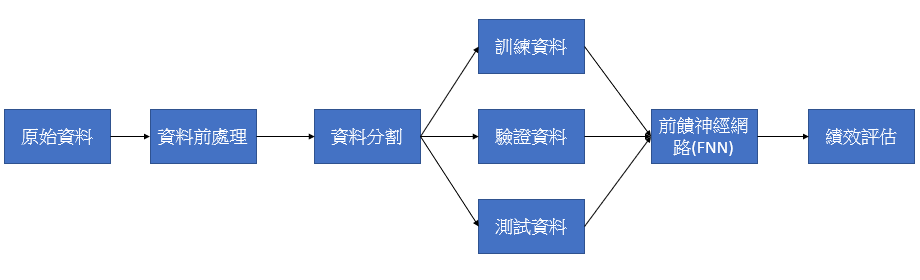


**圖 1**

*成人資料集實驗設計流程圖*

本研究的實驗設計如圖1。首先將成人資料集做資料前處理，得到乾淨的資料集後，再分成訓練資料和測試資料，分別使用KNN、SVM、RandomForest、XGBoost進行訓練，接著藉由測試資料來進行績效評估，最後根據績效的結果，對各模型進行超參數調整，再進行一次績效評估。

**3.2波士頓房價資料集**



**圖 1** 實驗設計流程圖

本研究實驗設計如上圖1，將原始資料透過由上一節所提到的資料前處理進行資料轉換，並將資料分成訓練資料、驗證資料和測試資料，在模型訓練的時候雖然我們會計算訓練和驗證的Loss，但是主要還是關注在驗證loss，另外我們使用early stopping技術當驗證loss不再有所提升時，我們就會停止訓練，將驗證最低的loss當作我們的模型來去測試在資料集上，最後就是調整模型參數和超參數來評估模型績效。

1. **實驗結果**
2. **葡萄酒品質資料集**

葡萄酒品質資料集的實驗分為兩種：類別預測與數值預測。

1. **成人資料集實驗結果**

進行模型訓練時，調整各個演算法的超參數，調整後的各個最好績效分別R2:RandomForest、RMSE:RandomForest、MAPE:RandomForest、RunTime(s):KNN。透過上述綜合結果RandomForest為表現最好的模型。

**表12**  KNN模型設定參數

|  |  |  |
| --- | --- | --- |
| **模型** | **n\_neighbors** | **leaf\_size** |
| **KNN** | 25 | 35 |

於

**表12**  RandomForest模型設定參數

|  |  |  |  |
| --- | --- | --- | --- |
| **模型** | **min\_samples\_split** | **max\_depth** | **max\_features** |
| **Random**  **Forest** | 25 | 30 | ‘sqrt’ |

**表12**  SVR模型設定參數

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **C** | **max\_iter** | **cache\_size** | **gamma** |
| **SVR** | 200 | 1000000 | 1000 | ‘auto’ |

**表12** 各模型績效表現

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **R2** | **RMSE** | **MAPE** | **Run Time(s)** |
| **KNN** | 0.24 | 10.87 | 0.31 | 0.0179 |
| **Random**  **Forest** | 0.25 | 10.83 | 0.31 | 1.79 |
| **XGBoost** | 0.22 | 11.21 | 0.32 | 133.11 |
| **SVR** | 0.20 | 11.15 | 0.33 | 2.45 |

1. **結論**

葡萄酒品質資料集中，分別預測了酒的品質(類別預測)以及酒成分中總二氧化硫含量多寡(數值預測)，實驗過程嘗試了不同模型參數組合並觀察各自的績效優劣。在預測葡萄酒品質時，發現了當資料集各類別資料分布有些懸殊時，容易使測試資料在預測上的績效較差，例如切割訓練資料時，第1、7類的資料只有少數兩筆，因此，測試資料的預測績效就會較差，而為了解決上述的問題，在未來的應用上可以使用Oversampling的技術，使各類別資料分布可趨於平均。此外，在這兩個實驗中，發現增加模型深度確實可以縮小generalization gap，但當深度較深時，可能會降低泛化能力，以本實驗為例，加入BatchNormalization後，確實有提高泛化能力。另外，值得我們注意的是early stopping 設定的patient以及batch\_size的設定，在兩個實驗中，不約而同，當patient設定較大時，模型在訓練與驗證集上的績效是組合中最好的，但在測試資料上的表現卻差強人意，可能模型產生過度擬合，以至於績效較差；而在batch\_size設定較小時，權重更新較為頻繁，模型在訓練以及用驗證資料評估的過程，雖然績效稍微差了一點，但在測試資料上反而績效為最佳，意味著當batch\_size設定太大的時候，權重更新較不頻繁，使模型在測試資料上的績效較為不好，而較小的batch\_size，能讓模型在預測陌生的資料上能夠更準確。

在成人資料集中，主要是透過KNN、RandomForest、XGBoost、SVR四種演算法去各別預測hours-per-week的績效為何，在調整各個演算法的超參數後，本研究的結果發現RandomForest是表現最好的模型，其中R2、 RMSE、MAPE績效分別達到0.25、10.83、0.31。

**參考文獻**

Ronny Kohavi、Barry Becker (1996)。成人數據集。

<https://archive.ics.uci.edu/ml/datasets/adult>

Paulo Cortez (2009)。葡萄酒質量數據集。

<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>

Dario Radečić (2021)。如何使用 TensorFlow 優化學習率——比你想像的要容易。

[How to Optimize Learning Rate with TensorFlow — It’s Easier Than You Think | by Dario Radečić | Towards Data Science](https://towardsdatascience.com/how-to-optimize-learning-rate-with-tensorflow-its-easier-than-you-think-164f980a7c7b)

Ryan Lu (2018)。Preprocessing Data : 類別型特徵\_OneHotEncoder & LabelEncoder 介紹與實作。

[https://medium.com/ai%E5%8F%8D%E6%96%97%E5%9F%8E/preprocessing-data-onehotencoder-labelencoder-%E5%AF%A6%E4%BD%9C-968936124d59](https://medium.com/ai反斗城/preprocessing-data-onehotencoder-labelencoder-實作-968936124d59)