資料探勘專案作業二

開發演算法建模與預測

指導教授：

許中川 教授

成員：

M11021012 林承威

M11021028 劉軒瑋

M11021052 邱守燦

M11021059 李鴻庭

日期：

2021 年 11 月 18 日

摘要

過往都以傳統的統計模型，進行分類預測，但隨著時代的演進，機器學習技術已逐漸取代。透過不同的演算法、參數設定，從大量數據中挖掘分析，得到有用的資訊，這種技術稱為資料探勘。本次研究分別使用 Adult Dataset 及 DryBean Dataset，建立每週工作小時數、豆子種類的預測模型，透過資料前處理抓取特徵及屬性，將資料集分割為 80%訓練資料、20%測試資料，分別使用 K 值鄰近法（KNN）、支持向量機（SVR）、隨機森林（RandomForest）、XGBoost 這些監督式學習的分類預測演算法法，最後透過訓練資料與測試資料的比對，根據各模型之績效評估得知，AdultDataset 使用由 RandomForest 和 XGBoost 建立的模型、DryBean Dataset 使用由 KNN 和 XGBoost 建立的模型，其預測與分析能力較佳，XGBoost 為最優良的分類器。

關鍵字：KNN、SVR、RandomForest、XGBoost

一、 緒論

1.1 動機

1.1.1 Adult Dataset

根據勞動部國際勞動統計，台灣去年就業者平均每年工時為 2021 小時，在 40 個主要國家中排名第 4 名，第一名則是新加坡。工時長短一直是勞動階級所關心的問題，同時也是展現該國家工作文化的一種指標；如若能投入一個人相關資訊，例如職業、收入、國籍等，便能預測出該人之工時長短，從而了解該國整體之工作文化。

1.1.2 Dry Bean Dataset

在種植農作物的田野與自然環境中，總是能發掘各式雜草的蹤影，雜草的危害之大，其中外來種影響尤甚。自從台灣加入世界貿易組織之後，進口的植物種類與數量不斷增加，使之辨識種子種類愈發困難，如若能透過種子外型的各項數據分析，從而辨識種子種類，將減少辨識困難度，進而防範有害外來種植物入侵本土生態。

1.2 研究目的

1.2.1 Adult Dataset

為了預測出該人之工時長短，從而了解該國整體之工作文化，本組選取 Adult 資料集，並使用 KNN、SVR、Random Forest 及 XGBoost 進行分類，投入工作類別、婚姻狀況、教育程度、職業、所屬國籍等 15 種屬性資料，並透過測試資料集驗證上述幾種模型的績效指標(MAE、RMSE、MAPE)，並根據績效指標進行比較，從中選擇較優之模型投入實際使用情境。

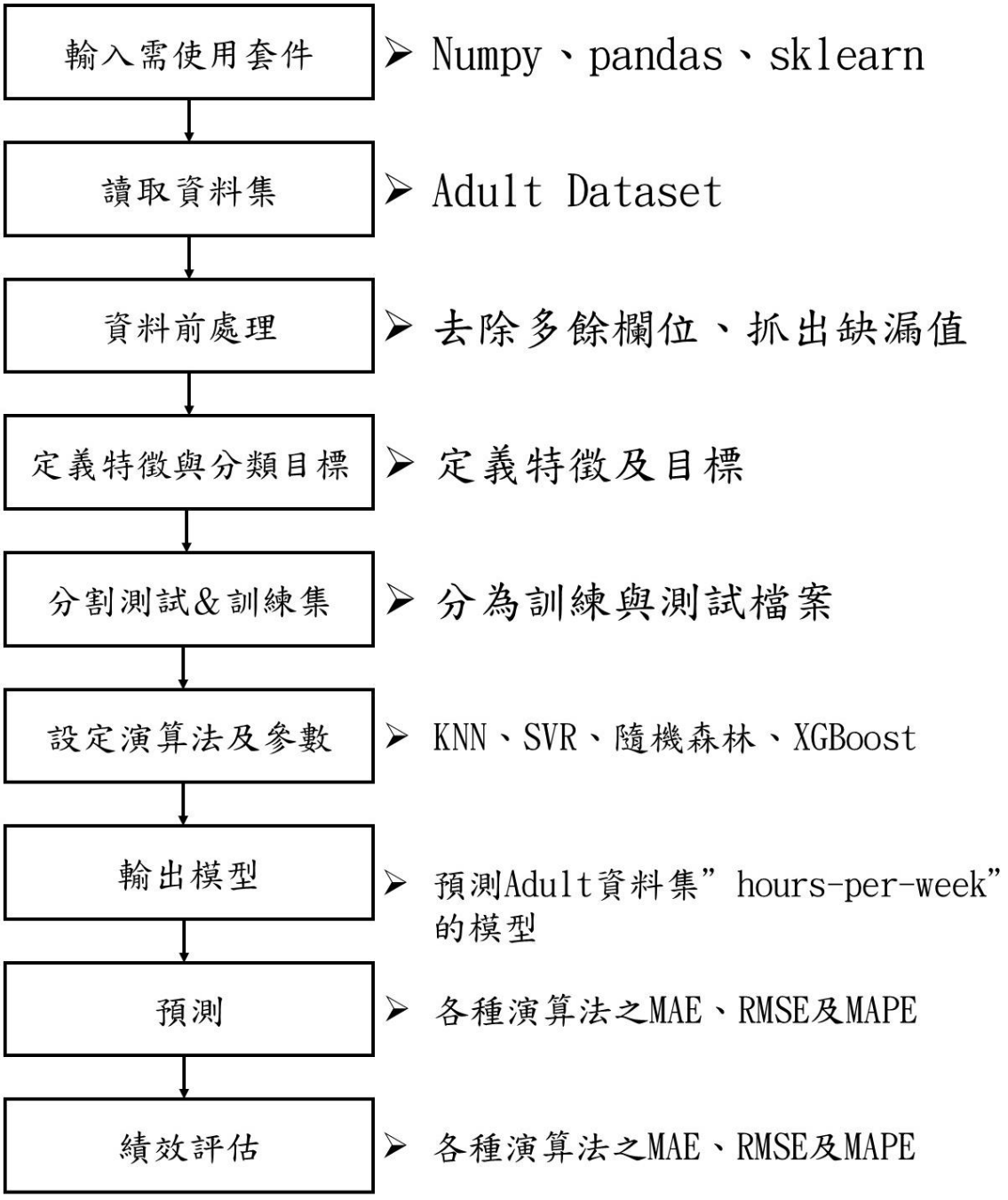
1.2.2 Dry Bean Dataset

為了辨識種子種類，進而防範有害外來種植物入侵本土生態，本組選取 Dry Bean 資料集，並使用 KNN、SVR、Random Forest 及 XGBoost 進行分類，投入種子的外部資料，例如：長軸長度、短軸長度、縱橫比、偏心、真圓度等 17 種屬性資料，並透過測試資料集驗證上述幾種模型的績效指標(MAE、RMSE、MAPE)，並根據績效指標進行比較，從中選擇較優之模型投入實際使用情境。

二、 方法

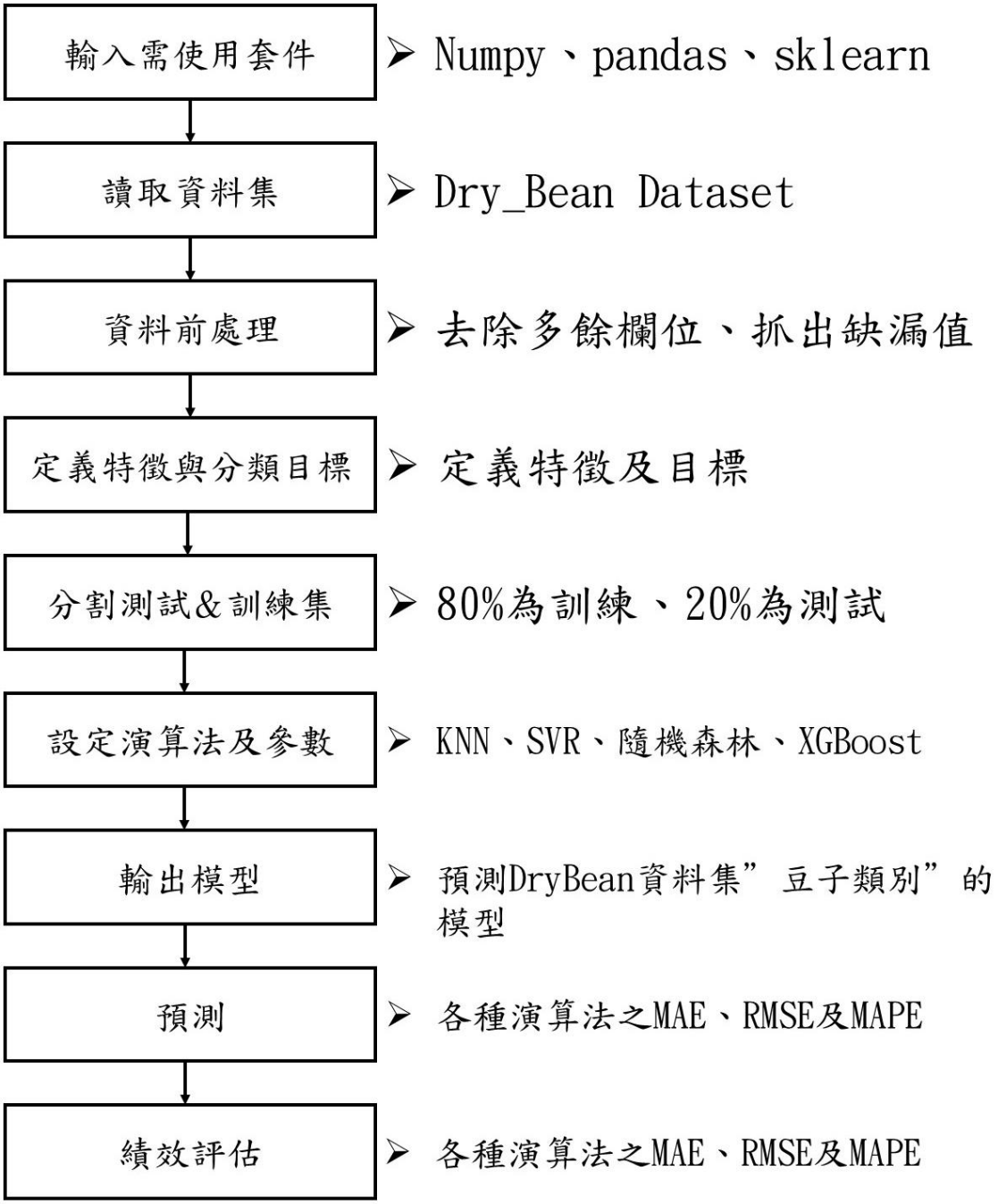
2.1 程式架構

2.1.1 Adult Dataset



圖一 Adult Dataset 之程式架構流程圖及說明

2.1.2 Dry_Bean Dataset



圖二 Dry_Bean Dataset 之程式架構流程圖及說明

2.2 程式執行方法

2.2.1 **KNN（K nearest neighbor）**

KNN（K nearest neighbor）是一種用於分類及回歸的無母數統計方法，此方法採用的是向量空間模型來做分類，會將概念相同或是相似度高的類別，藉由計算後將其歸納為一類，簡單來說就是物以類聚的概念，也是一個非常淺顯易懂的模型，用途很廣泛，在多類別的情況下分類表現也十分優異，但缺點就是其計算量十分龐大。

2.2.2 **SVR（Support Vector Regression/Machine）**

SVR（support vector Regression/Machine）支持向量機，是一種分類演算法但同時他也是迴歸的演算法，能夠根據不同的輸入資料能夠使模型有不同的使用，透過尋求結構化風險最小的方式來提高泛化能力，使在統計樣本量較少的情況下也能夠獲得良好的結果。

2.2.3 **隨機森林（Random Forest）**

隨機森林（Random Forest）是一個包含多個決策樹的分類器，並且其輸出的類別是由個別樹輸出的類別的眾數而定，隨機森林的重點在於抽樣，其樣本抽樣設計是採取後放回，並且在遇到過度配置時，隨機森林能夠解決，因為在大數法則下決策樹的結果會趨向一致，能夠大幅降低錯誤率。

2.2.4 **XGBoost（eXtreme Gradient Boosting)**

XGBoost（eXtreme Gradient Boosting)，是一種 Gradient Boosted Tree（GBDT)，保留原來的模型不變，再加入一個函數至模型中，並修正上一棵樹的錯誤，來提升整體模型的績效，主要應用於監督式學習，不僅適用於分類也可應用於回歸問題。

三、 實驗

3.1 資料集

3.1.1 Adult Dataset 說明

此資料集建立於 1996 年 5 月 1 日共有 48,842 筆，15 個欄位。

表一 Adult Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
Age	年齡
Workclass	工作類別
Fnlwgt	連續數值
Education	教育程度
Education-num	教育人數(連續數值)
Marital-status	婚姻狀況
Occupation	職業
Relationship	關係
Race	種族
Sex	性別
Capital-gain	資本收益
Capital-loss	資本損失
Hours-per-week	小時/周
Native-country	所屬國家
Salary	年收入

3.1.2 Dry Bean Dataset 說明

此資料集建立於 2018 年 9 月 7 日共有 4899 筆，17 個欄位。

表二 Dry Bean Dataset 欄位資料說明彙總表

欄位名稱	欄位說明
Area	地區
Perimeter	周長
MajorAxisLength	長軸長度
MinorAxisLength	短軸長度
AspectRation	縱橫比
Eccentricity	偏心
ConvexArea	凸面面積
EquivDiameter	等校直徑
Extent	程度
Solidity	堅固性
Roundness	真圓度
Compactness	緊湊度
ShapeFactor1	形狀因子 1
ShapeFactor2	形狀因子 2
ShapeFactor3	形狀因子 3
ShapeFactor4	形狀因子 4
Class	種子類別

3.1.3 實驗數據

(1) Adult Train 資料集

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week
0	39	5	77516	9	13	4	0	1	4	1	2174	0	40
1	50	4	83311	9	13	2	3	0	4	1	0	0	13
2	38	2	215646	11	9	0	5	1	4	1	0	0	40
3	53	2	234721	1	7	2	5	0	2	1	0	0	40
4	28	2	338409	9	13	2	9	5	2	0	0	0	40
...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	2	257302	7	12	2	12	5	4	0	0	0	38
32557	40	2	154374	11	9	2	6	0	4	1	0	0	40
32558	58	2	151910	11	9	6	0	4	4	0	0	0	40
32559	22	2	201490	11	9	4	0	3	4	1	0	0	20
32560	52	3	287927	11	9	2	3	5	4	0	15024	0	40

30162 rows × 15 columns

圖一 Adult Train 資料集

(2) Adult Test 資料集

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week
0	25	2	226802	1	7	4	6	3	2	1	0	0	40
1	38	2	89814	11	9	2	4	0	4	1	0	0	50
2	28	1	336951	7	12	2	10	0	4	1	0	0	40
3	44	2	160323	15	10	2	6	0	2	1	7688	0	40
5	34	2	198693	0	6	4	7	1	4	1	0	0	30
...	...	...	...	...	...	...	...	...	...	...	...	...	...
16275	33	2	245211	9	13	4	9	3	4	1	0	0	40
16276	39	2	215419	9	13	0	9	1	4	0	0	0	36
16278	38	2	374983	9	13	2	9	0	4	1	0	0	50
16279	44	2	83891	9	13	0	0	3	1	1	5455	0	40
16280	35	3	182148	9	13	2	3	0	4	1	0	0	60

15060 rows × 15 columns

圖二 Adult Test 資料集

(3) DryBean Dataset 資料集

	Area	Perimeter	MajorAxisLength	MinorAxisLength	AspectRation	Eccentricity	ConvexArea	EquivDiameter
0	28395	610.291	208.178117	173.888747	1.197191	0.549812	28715	190.141097
1	28734	638.018	200.524796	182.734419	1.097356	0.411785	29172	191.272751
2	29380	624.110	212.826130	175.931143	1.209713	0.562727	29690	193.410904
3	30008	645.884	210.557999	182.516516	1.153638	0.498616	30724	195.467062
4	30140	620.134	201.847882	190.279279	1.060798	0.333680	30417	195.896503
...	...	...	...	...	...	...	...	...
13606	42097	759.696	288.721612	185.944705	1.552728	0.765002	42508	231.515799
13607	42101	757.499	281.576392	190.713136	1.476439	0.735702	42494	231.526798
13608	42139	759.321	281.539928	191.187979	1.472582	0.734065	42569	231.631261
13609	42147	763.779	283.382636	190.275731	1.489326	0.741055	42667	231.653247
13610	42159	772.237	295.142741	182.204716	1.619841	0.786693	42600	231.686223

13611 rows × 17 columns

圖三 DryBean Dataset 資料集

3.2 前置處理

3.2.1 Adult Train Dataset

將 Adult Train 資料集，由原來的 txt 檔案格式轉換為 csv 檔案格式，發現在些許的資料欄位中具有不必要的符號資料，進行資料的清洗，利用 lambda 語法將具有'?'資料篩選與過濾，完成資料清洗，接著將 Adult Dataset 裡面文字屬性使用 Labelencoder 進行轉碼，最後讀取印出完成資料的前置處理。

3.2.2 Adult Test Dataset

將 Adult Train 資料集，由原來的 txt 檔案格式轉換為 csv 檔案格式，發現在些許的資料欄位中具有不必要的符號資料，進行資料的清洗，利用 lambda 語法將具有'?'資料篩選與過濾，完成資料清洗，接著將 Adult Dataset 裡面文字屬性使用 Labelencoder 進行轉碼，最後讀取印出完成資料的前置處理。

3.2.3 DryBean Dataset

將 DryBean 資料集，利用 Pandas 套件的 DataFrame 物件讀取 csv 檔，使用 isna 函數檢查資料欄位中是否有缺失值,將資料過濾,完成資料的前置處理。

3.3 實驗設計

3.3.1 Adult Train/Test Dataset

(1)正規化：由 sklearn preprocessing 開源軟體載入 MinMaxScaler 標準化套件，將資料的數值做標準化。

(2)套件載入:由 Scikit-Learn 開源軟體載入 KNN 建模(KNeighborsClassifier)的套件、RandomForest 建模（RandomForestRegressor）的套件、SVR 建模（LinearSVR）的套件、XGBoost 建模（xgboost）的套件。

(3)KNN：找尋最佳 K 值鄰近點，進行後續分類

(4)Randomorest：建立森林樹木（n_estimators）數量為 1000，將處理器設定無限制函式的運作，把 oob_score 默認為 True，使用袋外樣本估計模型約略的準確率，由 bootstrap 參數默認為 True，經由取後放回的隨機抽樣技術完成模型架構，完成參數設定後進行模型建立與訓練。

(5)SVR：設定隨機數的生成，並設定容錯值為:1e-5，完成參數設定後進行模型建立與訓練。

(6)XGBoost ： 定 義 類 別 數 (objective) 為 :"reg:linear" ， 設 列 取 樣（colsample_bytree）參數為 0.3，資料學習率（learning_rate）設 0.1，最大深度（max_depth）為 5，模型樣本數（n_estimators）為 10，完成參數設定後進行模型建立與訓練。

3.3.2 DryBean Dataset

(1)正規化：由 sklearn preprocessing 開源軟體載入 MinMaxScaler 標準化套件，將資料的數值做標準化。

(2)套件載入:由 Scikit-Learn 開源軟體載入 KNN 建模(KNeighborsClassifier)的套件、RandomForest 建模（RandomForestRegressor）的套件、SVR 建模（LinearSVR）的套件、XGBoost 建模（xgboost）的套件。

(3)KNN：找尋最佳 K 值鄰近點，進行後續分類

(4)RandomForest：建立森林樹木（n_estimators）數量為 1000，將處理器設定無限制函式的運作，把 oob_score 默認為 True，使用袋外樣本估計模型約略的準確率，由 bootstrap 參數默認為 True，經由取後放回的隨機抽樣技術完成模型架構，完成參數設定後進行模型建立與訓練。

(5)SVR：設定隨機數的生成，並設定容錯值為:1e-5，完成參數設定後進行模型建立與訓練。

(6)XGBoost ： 定 義 類 別 數 (objective) 為 :"reg:linear" ， 設 列 取 樣（colsample_bytree）參數為 0.3，資料學習率（learning_rate）設 0.1，最大深度（max_depth）為 5，模型樣本數（n_estimators）為 10，完成參數設定後進行模型建立與訓練。

3.4 實驗結果

3.4.1 Adult 績效評估

	KNN	RandomForest	SVR	XGBoost
MAE	7.1593	0.0279	-6.18	0.0633
RMSE	12.0875	0.0406	-3.92	0.0858

表一 Adult 績效評估

3.4.2 DryBean 績效評估

	KNN	RandomForest	SVR	XGBoost
MAE	0.1014	6.4497	-1.5363	0.1124
RMSE	0.4285	0.0003	0.7471	0.1362

表二 DryBean 績效評估

四、 結論

Adult Dataset 資料集以 RandomForest 跟 XGBoost 較為優秀。

DryBean Dataset 資料集以 KNN 跟 XGBoost 較為優秀。

總結以上，XGBoost 是一個最佳的分類器。

五、 參考文獻

[1]機器學習首部曲隨機森林模型簡介 Rondom Forest
https://pyecontech.com/2019/09/17/random_forest/

[2] 臺灣草本植物種子彩色圖鑑 II - 農業藥物毒物試驗所
https://www.tactri.gov.tw/Uploads/Item/bccc7a04-c637-42cd-af50-4af26691e15d.pdf

[3]學習機器學習必知的程序-資料庫知識探索
https://medium.com/marketingdatascience/%E5%AD%B8%E7%BF%92%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E5%BF%85%E7%9F%A5%E7%9A%84%E7%A8%8B%E5%BA%8F-%E8%B3%87%E6%96%99%E5%BA%AB%E7%9F%A5%E8%AD%98%E6%8E%A2%E7%B4%A2-72bd2d73781c

[4] 過勞之島！台灣 2020 總工時全球第 4 比前年減 6 小時
https://udn.com/news/story/7238/5822033

[5]用 Python 自學資料科學與機器學習入門實戰：Scikit Learn 基礎入門
https://blog.techbridge.cc/2017/11/24/python-data-science-and-machine-learning-scikit-learn-basic-tutorial/

[6] Python 機器學習筆記(五)：使用 Scikit-Learn 進行 K-Nearest 演算法
https://yanwei-liu.medium.com/python%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E7%AD%86%E8%A8%98-%E4%BA%94-%E4%BD%BF%E7%94%A8scikit-

learn%E9%80%B2%E8%A1%8Ck-nearest%E6%BC%94%E7%AE%97%E6%B3%95-1191ea94ecaf
[7] 機器學習：KNN 分類演算法
https://ithelp.ithome.com.tw/articles/10197110}
[8] knn – adult
https://www.kaggle.com/kenzok/knn-adult
[9] Python 實作隨機森林模型 Random Forest
https://wreadit.com/@pyecontechcom/post/274052
[10] RANDOM FOREST WITH US ADULT INCOME DATASET
https://meuge.github.io/blog/2019/03/08/us-adult-income
[11] 學習 SVM，這篇文章就夠了！（附詳細程式碼）
https://iter01.com/9679.html
[12] ML 入門（十一）支援向量機(Support Vector Machine,SVM) SVM)
https://medium.com/chung-yi/ml%E5%85%A5%E9%96%80-%E5%8D%81%E4%B8%80-%E6%94%AF%E6%8F%B4%E5%90%91%E9%87%8F%E6%A9%9F-support-vector-machine-svm-c8c1bb1c970f}
[13] Python 實作支援向量 SVM
https://pyecontech.com/2020/04/11/python_svm/
[14] XGBoost——機器學習（理論 圖解 python 程式碼）
https://codertw.com/%E7%A8%8B%E5%BC%8F%E8%AA%9E%E8%A8%80/394270/
[15] [Day 15] 機器學習常勝軍 –XGBoost
https://ithelp.ithome.com.tw/articles/10273094?sc=iThomeR