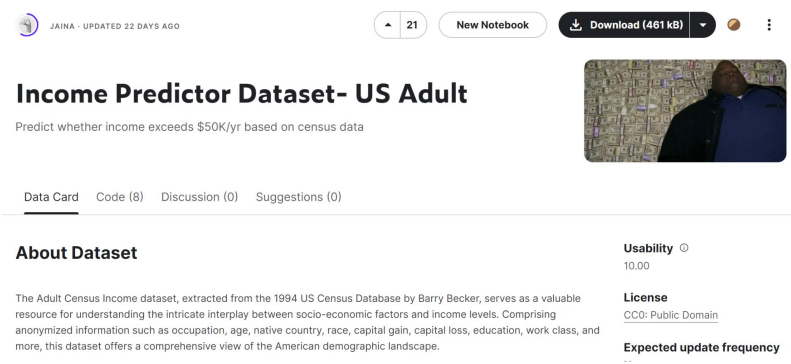
巨量資料分析期末報告

01057051 陳俞君

資料集&目的

使用的數據集來自UCI機器學習資料庫中的Adult數據集。
該數據集包含了許多個體的人口統計信息和收入水平標籤。具體特徵包括年齡、工作類別、教育程度、婚姻狀況、職業、種族、性別、每週工作時數等。
主要目的: 利用機器學習技術來預測個體的收入水平是否超過50K美元。
通過分析人口統計數據和其他特徵，建構分類模型，識別哪些個體的收入可能會超過此門檻。



資料集

資料集具32561筆資料及 15 項特徵，特徵如下:

工作類別(workclass):

私營企業(Private)
自營但不成立公司(Self-emp-not-inc)
自營且成立公司(Self-emp-inc)
聯邦政府(Federal-gov)
地方政府(Local-gov)
州政府(State-gov)
無薪工作(Without-pay)
從未工作(Never-worked)

Subtitle here

教育程度(education):

學士學位(Bachelors)
大專(Some-college)
11年級(11th)
高中畢業(HS-grad)
專業學校(Prof-school)
學術副學士(Assoc-acdm)
職業副學士(Assoc-voc)
9年級(9th)
7至8年級(7th-8th)
12年級(12th)
碩士學位(Masters)
1至4年級(1st-4th)
10年級(10th)
博士學位(Doctorate)
5至6年級(5th-6th)
幼稚園(Preschool)

婚姻狀況(marital.status):

已婚-配偶在場
(Married-civ-spouse)
離婚(Divorced)
未婚(Never-married)
分居(Separated)
喪偶(Widowed)
已婚-配偶不在場
(Married-spouse-absent)
已婚-軍人配偶
(Married-AF-spouse)

資料集

關係(relationship):

妻子(Wife)
孩子(Own-child)
丈夫(Husband)
非家庭成員(Not-in-family)
其他親屬(Other-relative)
未婚(Unmarried)

種族(race):

白人(White)
亞洲-太平洋島民
(Asian-Pac-Islander)
美洲印第安人-愛斯基摩人
(Amer-Indian-Eskimo)
其他(Other)
黑人(Black)

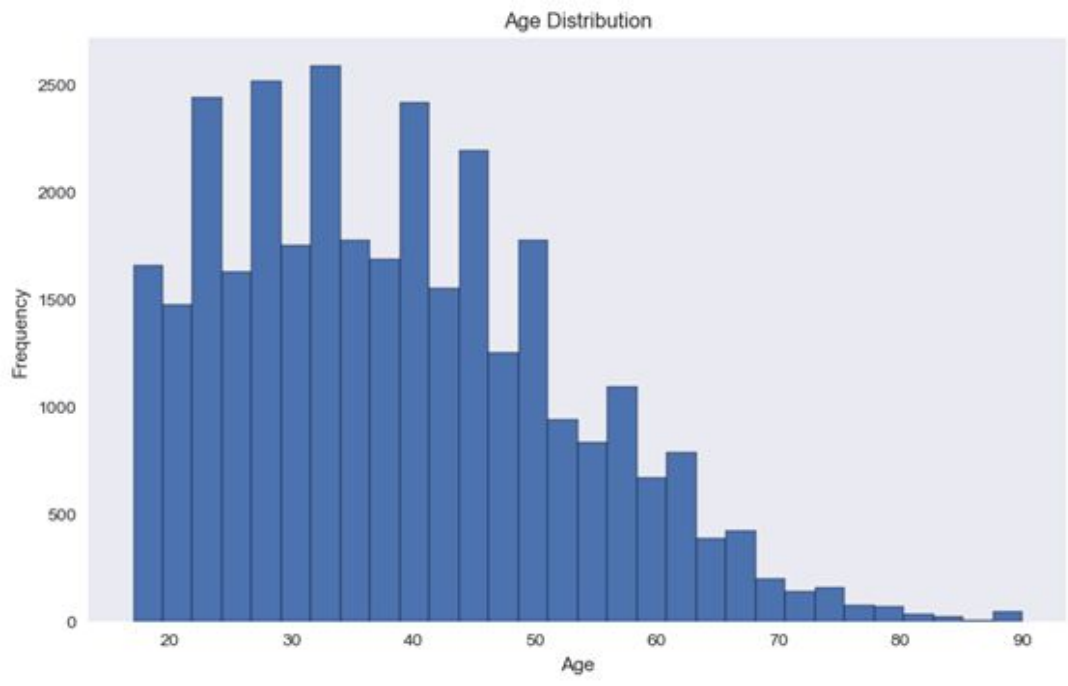
性別(sex):

男性(Male)
女性(Female)

國籍(native.country):

美國(United-States)
柬埔寨(Cambodia)
英國(England)
波多黎各(Puerto-Rico)
加拿大(Canada)
德國(Germany)
美國邊疆(Outlying-US(Guam-USVI-etc))
印度(India)
日本(Japan)
希臘(Greece)
南部(South)
中國(China)
古巴(Cuba)
伊朗(Iran)
洪都拉斯(Honduras)
菲律賓(Philippines)
意大利(Italy)
波蘭(Poland)
牙買加(Jamaica)
越南(Vietnam)
墨西哥(Mexico)
葡萄牙(Portugal)
愛爾蘭(Ireland)
法國(France)
多明尼加共和國(Dominican-Republic)
老撾(Laos)
厄瓜多爾(Ecuador)
台灣(Taiwan)
海地(Haiti)
哥倫比亞(Columbia)
匈牙利(Hungary)
危地馬拉(Guatemala)
尼加拉瓜(Nicaragua)
蘇格蘭(Scotland)
泰國(Thailand)
南斯拉夫(Yugoslavia)
薩爾瓦多(El-Salvador)
特立尼達和多巴哥(Trinadad&Tobago)
秘魯(Peru)
香港(Hong)
荷蘭(Holand-Netherlands)

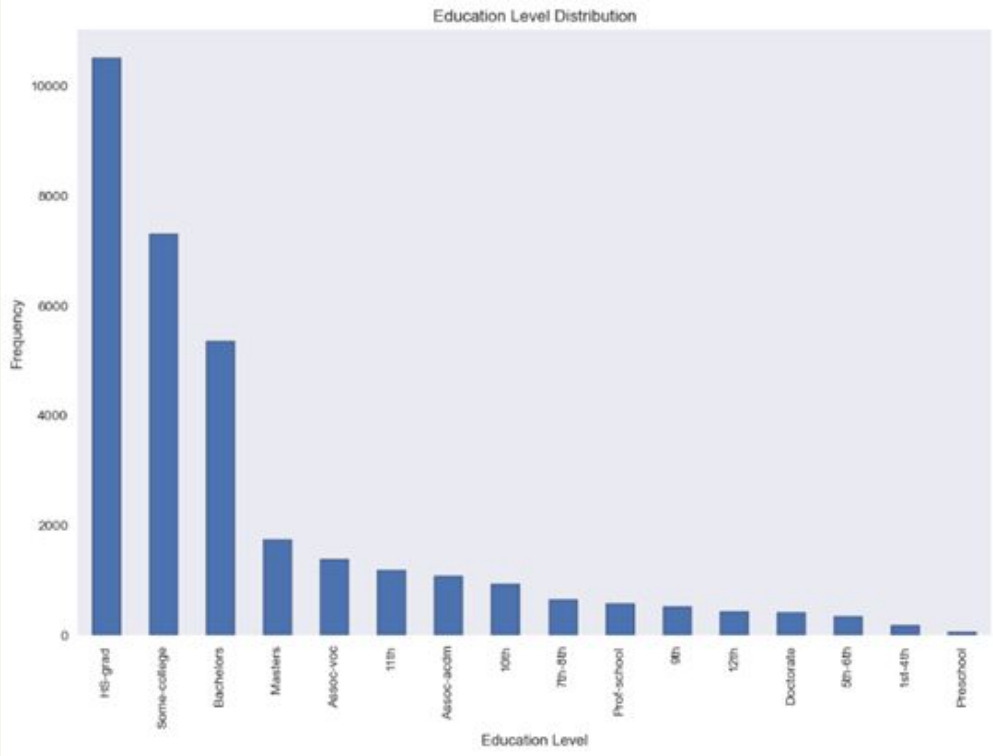
統計分析



Subtitle here

年齡分佈直方圖：展示了不同年齡段的數據分佈情況。

主要分布在20~45歲



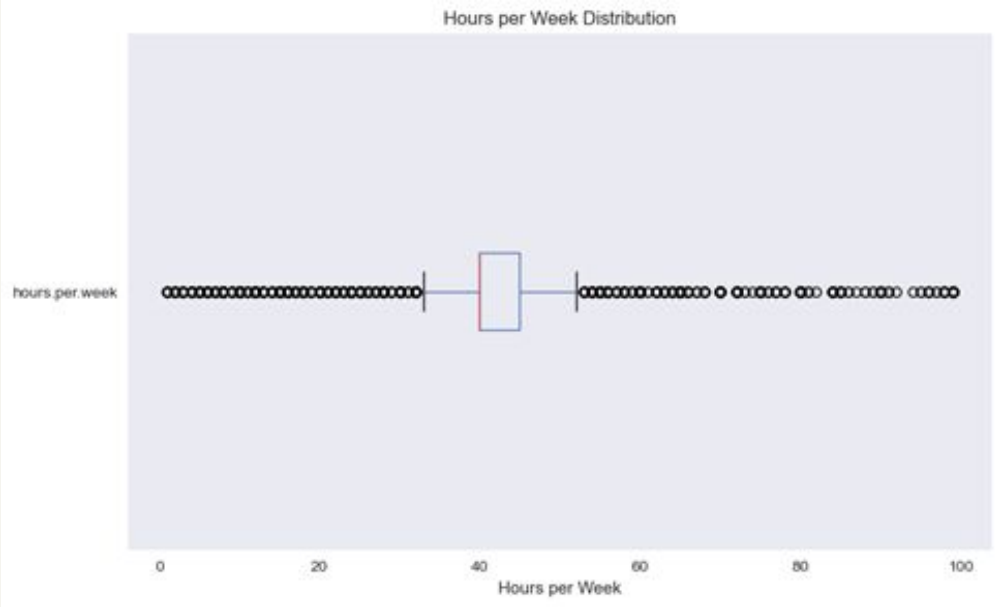
Subtitle here

教育程度分佈條形圖：展示了不同教育程度的人數分佈。

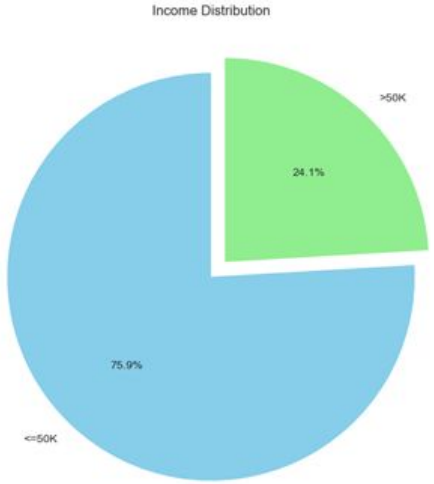
教育程度（education）：由多至少

高中畢業（HS-grad）
學院（Some-college）
學士學位（Bachelors）
碩士學位（Masters）
職業副學士（Assoc-voc）
11年級（11th）
學術副學士（Assoc-acdm）
10年級（10th）
7至8年級（7th-8th）
專業學校（Prof-school）
9年級（9th）
12年級（12th）
博士學位（Doctorate）
5至6年級（5th-6th）
1至4年級（1st-4th）
幼稚園（Preschool）

最高的前三類別由高到低為高中畢業、大學學院)、學士



每周工作小時數的盒圖：展示了每周工作小時數的分佈和異常值。



平均在每周42小時左右

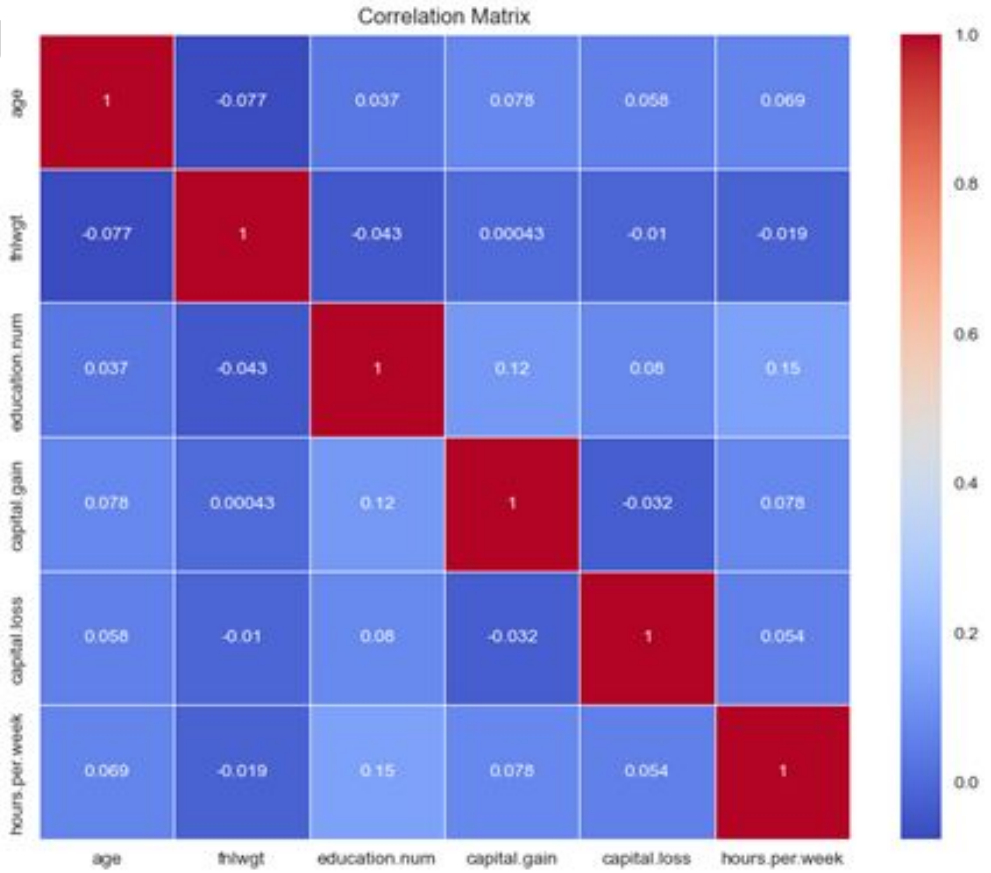
收入分佈圓餅圖：展示了收入在兩個範圍（<=50K 和 >50K）中的比例。

熱力圖展示了數據集中主要數值變量之間的相關性。從圖中可以看出：

變量之間的相關性不明顯。

(越接近1相關性越高)

Subtitle here

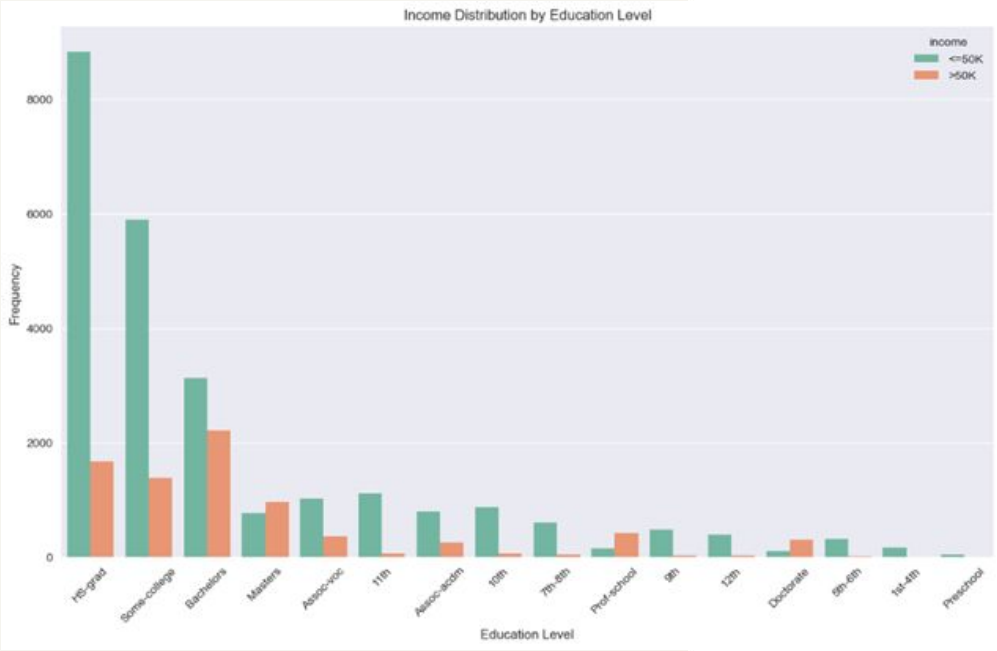


相關性分析

1. 收入與年齡的關係: 將年齡分段，並分析每個年齡段的收入分佈情況。

Subtitle here

年齡較大的群體中，高收入（>50K）的人數比例相對較高。



教育程度（education）：由多至少

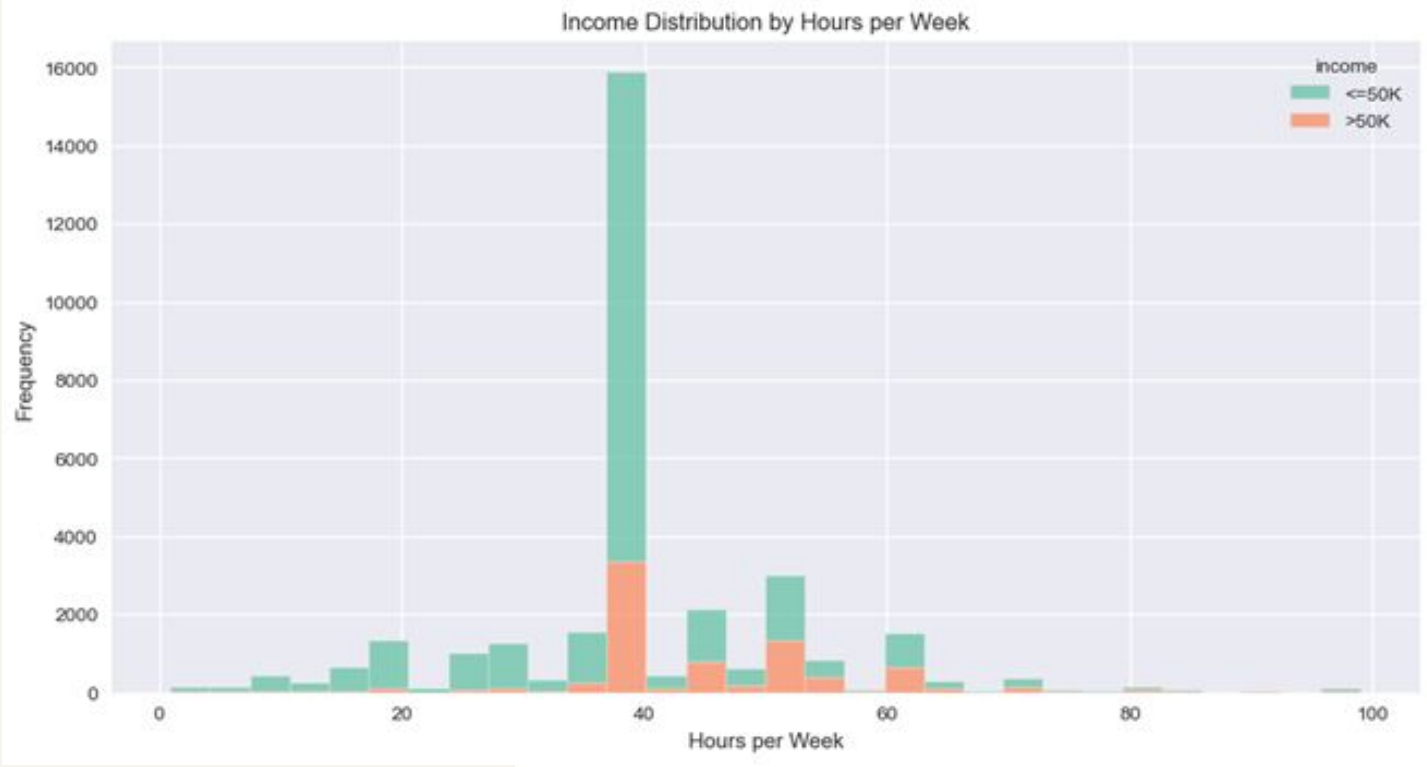
高中畢業（HS-grad）
學院（Some-college）
學士學位（Bachelors）
碩士學位（Masters）
職業副學士（Assoc-voc）
11年級（11th）
學術副學士（Assoc-acdm）
10年級（10th）
7至8年級（7th-8th）
專業學校（Prof-school）
9年級（9th）
12年級（12th）
博士學位（Doctorate）
5至6年級（5th-6th）
1至4年級（1st-4th）
幼稚園（Preschool）

1. **收入與教育程度的關係 分析不同教育程度的收入分佈情況。**

Subtitle here

i. **擁有較高教育程度（如學士、碩士及以上學位）的人群中，高收入者比例較高。**

ii. **教育程度較低（如未完成高中教育）的群體中，低收入（<=50K）的人數比例較高。**



1. 收入與每周工作小時數的關係 每周工作小時數與收入的分佈圖。

每周工作小時數較多的人群中，高收入者比例較高。

建立機器學習模型

pandas、seaborn、matplotlib

一、預處理過程

數據清洗：首先，對數據中的缺失值進行處理，將包含缺失值的行刪除並對一些文本數據進行清理，去除多餘的空格。

二、特徵工程：使用 pd.get_dummies 方法將分類變量轉換為多個二元變量（one-hot encoding），以便機器學習算法可以處理這些數據。

三、特徵縮放：使用 StandardScaler 對數值特徵進行標準化，以確保各特徵的值在同一範圍內，有助於模型的收斂。

建立機器學習模型

四、模型訓練與評估

scikit-learn

1. 使用多種機器學習算法來訓練模型，包括隨機森林、決策樹、支持向量機、K近鄰算法和XGBoost。

2. 使用準確率、混淆矩陣和ROC曲線等指標評估這些模型的性能。

Subtitle here

總結: 相關性分析

計算各個數值變量之間的相關性，使用皮爾森相關係數來衡量變量之間的線性關係。

1. 教育年數與收入有較高的正相關性，表明受教育程度越高，收入越有可能超過50K。

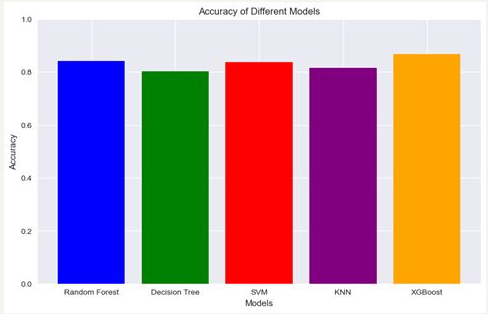
2. 每週工作小時數與收入也有一定的正相關性，表明工作時間越長，收入越高的可能性越大。

3. 年齡與收入有中等程度的相關性。

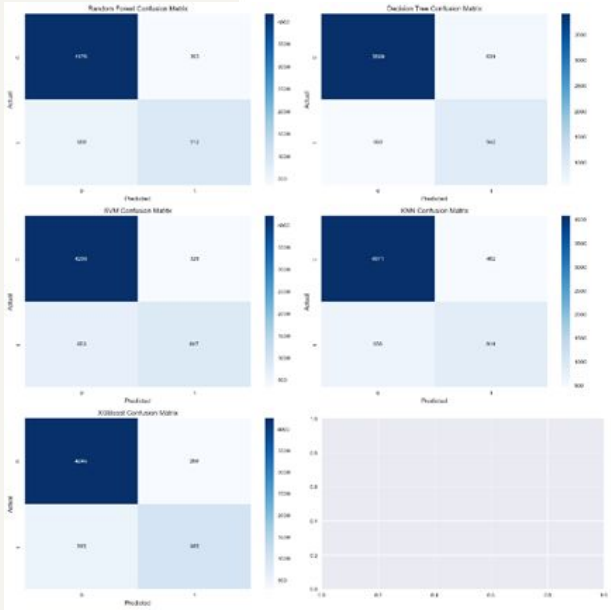
Subtitle here

總結: 機器學習預測

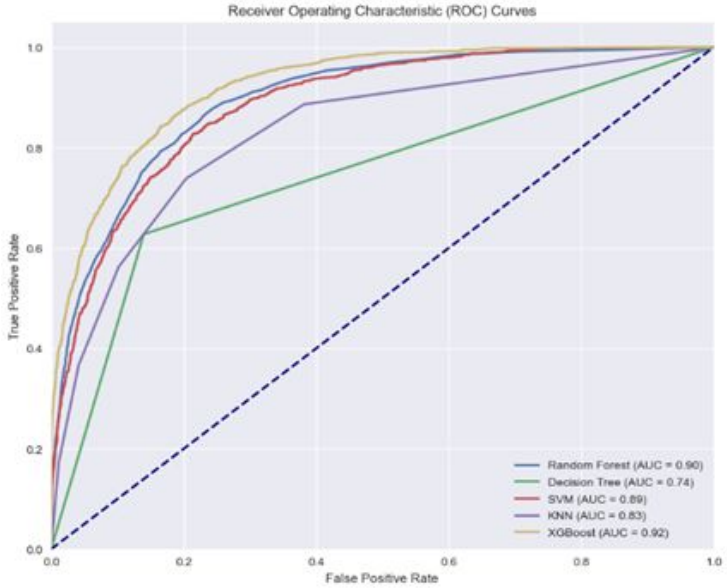
隨機森林和XGBoost模型在準確率和穩定性方面表現較好



準確率



混淆矩陣



ROC曲線

總結

本次試著使用Python撰寫程式碼透過以上方法，建構多個機器學習模型，並比較模型的性能。隨機森林和XGBoost模型在準確率和穩定性方面表現較好，在相關性方面，教育年數與收入有較高的正相關性，每週工作小時數與收入也有一定的正相關性。

本次程式碼:
https://drive.google.com/file/d/11kLy7iNrIlZ2bqOiDuqSFL4zt4hQs-ax/view?usp=sharing

Subtitle here

參考

Income Predictor Dataset- US Adult (使用的資料集 FROM Kaggle)

Titanic - Machine Learning from Disaster鐵達尼號生存預測 (參考模型程式碼寫法)

Subtitle here