巨量資料分析期末報告

01057051 陳俞君

一. 目的:這個主題想要達成什麼成果

主要目的是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，來識別哪些個體的收入可能會超過此門檻。

二. 文獻回顧:

這個問題以前的人怎麼做，目前有沒有處理的方法

在研究分析中一般利用統計方法和機器學習技術來進行收入預測。比較早期的分析多基於迴歸分析，通過分析人口統計學特徵來預測收入水平。隨著機器學習技術的發展，越來越多研究開始使用新的分類算法，如決策樹、隨機森林、支持向量機、K 近鄰算法和梯度提升等方法，來提高預測的準確性和效率。

本次由機器學習平台 Kaggle 入門(主要是鐵達尼號生存預測)，使用多種 Python lib 撰寫程式來進行數據處理、機器學習建模和數據可視化。。

三. 分析過程及方法:

資料從哪裡取得，我拿裡面那些資料來做，怎麼做，有沒有遇到困難，怎麼克服

資料來源

Income Predictor Dataset- US Adult

使用的數據集來自於 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。**

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。

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。

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。

統計分析

對於每個特徵進行了基本的統計描述，包括平均值、中位數、標準差、最小值、最大值等。這有助於了解數據的基本特性，例如：年齡分佈、每週工作小時數的分佈、各類工作類別和教育程度的分佈。

此外，繪製各個特徵的分佈圖，如直方圖和盒鬚圖，以便更直觀了解數據的分佈情況。

模型訓練與評估

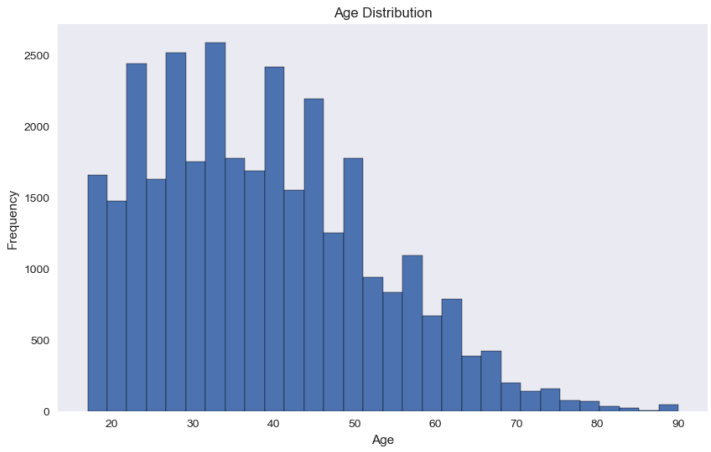
1. 使用多種機器學習算法來訓練模型，包括隨機森林、決策樹、支持向量機、K 近鄰算法和

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。

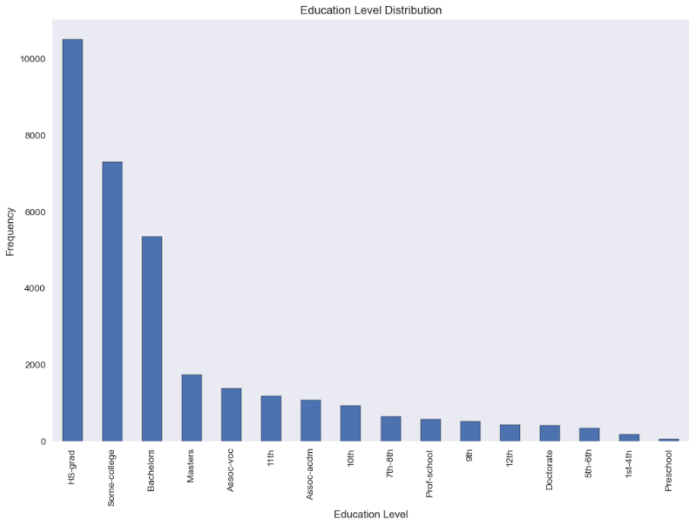
描述性統計分析：

繪製數據分佈圖，如直方圖和盒鬚圖。



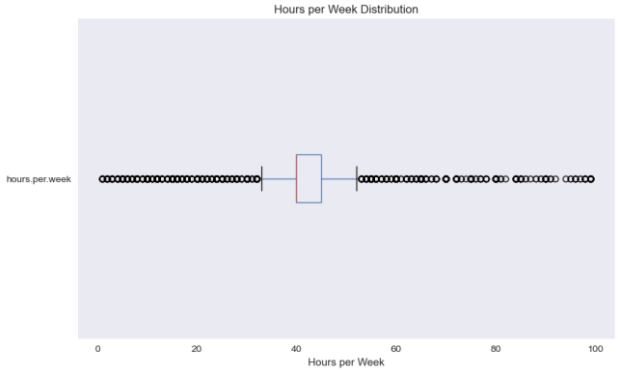
年齡分佈直方圖：展示了不同年齡段的數據分佈情況。

主要分布在 20~45 歲



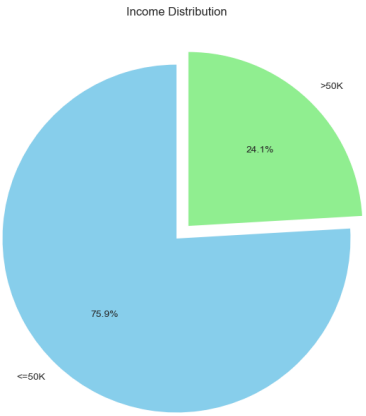
教育程度分佈條形圖：展示了不同教育程度的人數分佈。

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鬚圖：展示了每周工作小時數的分佈和異常值。

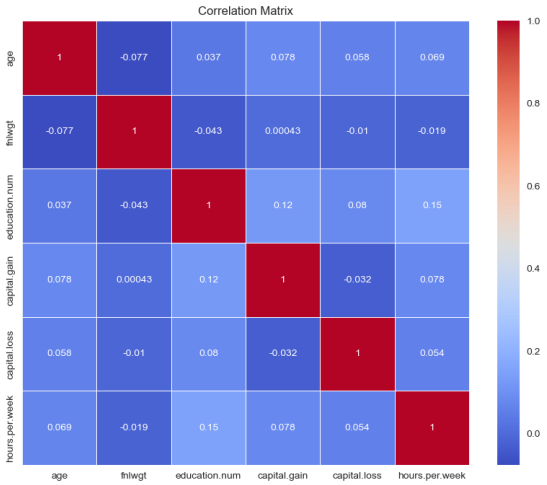
平均在每周 42 小時左右



收入分佈圓餅圖：展示了收入在兩個範圍（<=50K 和 >50K）中的比例。

<=50K 的人佔 75% >50K 的人佔約 25%

相關性分析：計算不同變量之間的相關性。



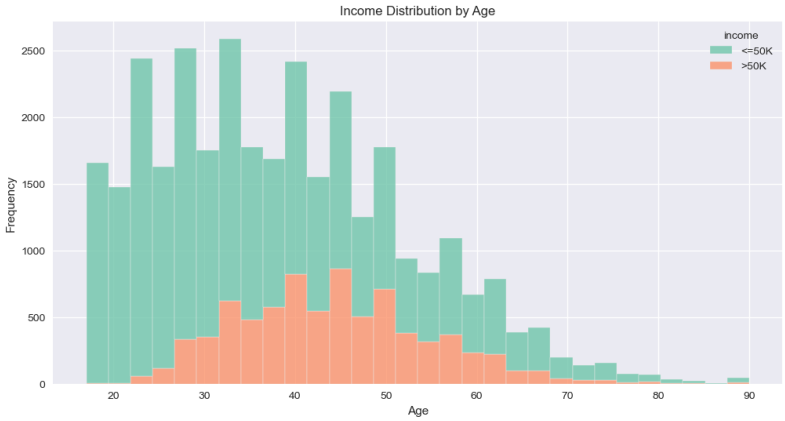
熱力圖展示了數據集中主要數值變量之間的相關性。從圖中可以看出：

變量之間沒有明顯的線性關係。

分群分析 根據收入（<=50K 或 >50K）進行分群，分析各群體的特徵

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。

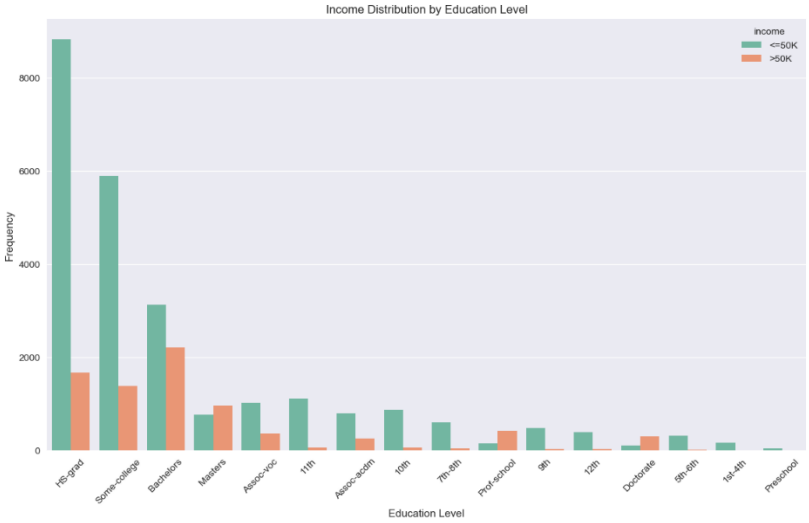
年齡較大的群體中，高收入（>50K）的人數比例相對較高。



2. 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。

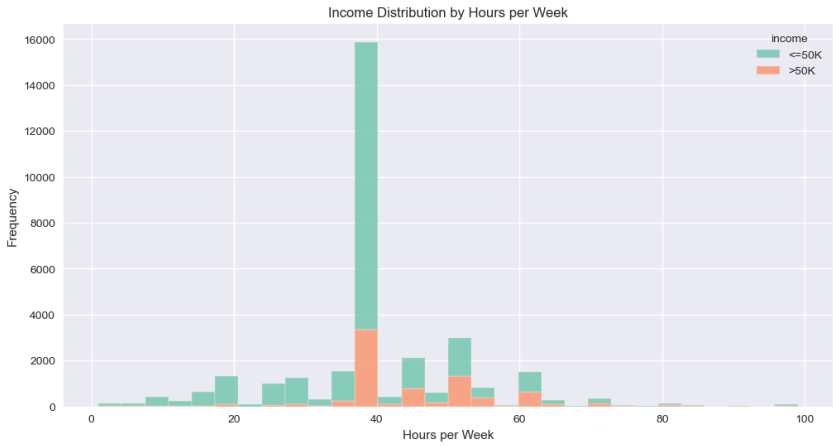
i. 擁有較高教育程度（如學士、碩士及以上學位）的人群中，高收入者比例較高。

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。



3. 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。



四. 結果:

我得到什麼樣的結果，跟我的預期有沒有差距

遇到的困難與克服方法

數據缺失：數據集中存在部分缺失值，刪除包含缺失值的行來清理數據。

特徵處理：分類變量需要進行 one-hot encoding，這會導致特徵數量增加。使用了 pd.get_dummies 方法來自動完成這一過程。

模型過擬合：一些模型（如決策樹）容易過擬合。透過調整模型參數（如最大深度）和使用集成方法（如隨機森林和 XGBoost）來減少過擬合的影響。

相關性分析

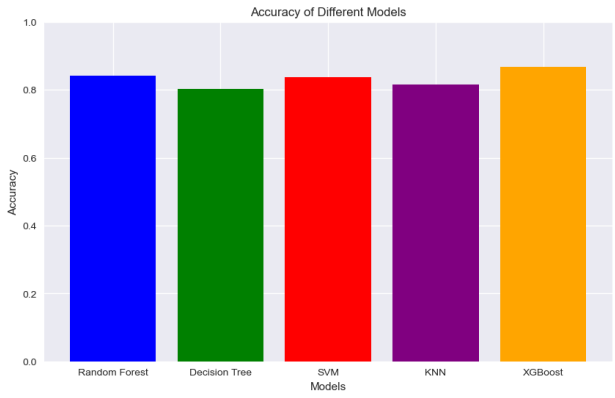
計算各個數值變量之間的相關性，特別是與收入水平的相關性。使用皮爾森相關係數來衡量變量之間的線性關係。

結果如下：

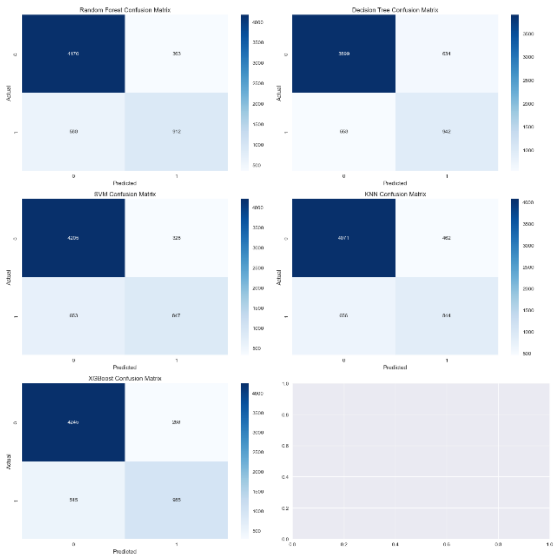
1. 教育年數與收入有較高的正相關性，表明受教育程度越高，收入越有可能超過 50K。
2. 每週工作小時數與收入也有一定的正相關性，表明工作時間越長，收入越高的可能性越大。
3. 年齡與收入有中等程度的相關性。

機器學習預測

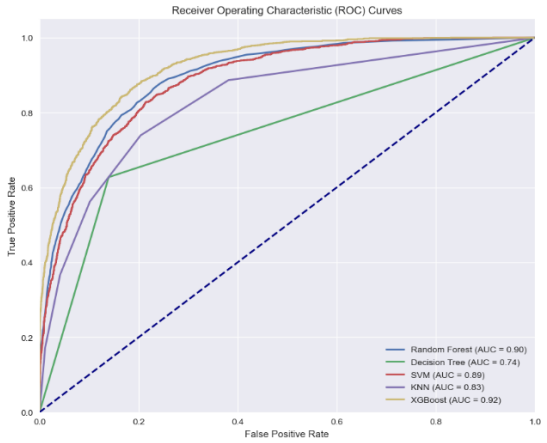
隨機森林和 XGBoost 模型在準確率和穩定性方面表現較好



準確率



混淆矩陣



ROC 曲線

五. 結論:整體的收穫及後續可能的延伸方向

本次試著使用 Python 撰寫程式碼透過以上方法，建構多個收入預測模型，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。

六. 參考資料

Income Predictor Dataset- US Adult Predict whether income exceeds $50K/yr based on census data : https://www.kaggle.com/datasets/jainaru/adult-income-census-dataset/code