1112-量化交易系統-期中作業

1. 題目 ： 鐵達尼號之預測存活率
2. 摘要 ： 利用年齡等特徵預測存活與否。
3. 資料集介紹 :我們所使用 kaggle 網頁上的鐵達尼號資料下載至電腦資料夾 https://www.kaggle.com/c/titanic/data?select=train.csv，裡面有訓練集與測試集。
4. 選擇方法介紹 ： 利用 KNN 分類、決策樹、隨機森林分類、MLP 分類、自適應增強分類進行處理
5. 程式說明 :資料處裡→使用分類進行資料模型預測→模型評估

    先用 Pandas 輸入資料集，將訓練集的標籤印出，確認所有標籤的資料型態，我們先從 Age 標籤進行資料處理，忽略空值，並再次確認 (891rows 處理至 714rows)。

    我們將 X 代入特徵資料，y 代入生還者標籤 ，使 scikit-learn 中的 train_test_split 函數將 X 和 y 變數分割為訓練集和測試集，並將 20%的資料用於測試，再使用 sklearn 的標準化方法將 X_train 和 X_test 資料集中的特徵進行標準化，建立決策樹分類器並且訓練模型使用 fit 方法，測試模型使用 predict 方法，並且計算其在測試集上的精度。

    接著我們用中位數填年齡空值，再使用剛剛的方式計算測試及上面的準確度。船艙的部份我們先按照乘客等級（頭等艙、二等艙、三等艙）的生存率高低排序。

    其中 **train_data** 是一個包含了鐵達尼號乘客資訊的資料集，包括每位乘客的等級、性別、年齡、票價、登船港口等資訊。透過這段程式碼，我們可以知道不同等級的乘客在鐵達尼號撞擊事故中的生存情況，進一步分析生存率和乘客等級之間的關係且劃出 bar 圖表，並計算在繪圖，且繪

製熱力圖。

再來分析性別特徵(男性是 1 女性是 2)。最後再寫一個用來將資料集分為訓練集和測試集，並準備用於後續的機器學習建模。

最後再利用利用 KNN 分類、決策樹、隨機森林分類、MLP 分類、自適應增強分類進行處理，並優化。

6. 模型評估說明：在模型評估這裡，我們使用了 **KNeighborsClassifier**、**DecisionTreeClassifier**、**RandomForestClassifier**、**MLPClassifier** 和 **AdaBoostClassifier** 這些模型，並分別輸出了它們的準確度。其中，隨機森林的準確度最高，為 0.754。其次是 AdaBoostClassifier 模型，準確度為 0.743。而決策樹和 MLPClassifier 模型的表現相對較差，分別為 0.715 和 0.698。K 近鄰的表現則居中，準確度為 0.654
7. 結論：這次資料集的分析主要是針對乘客的基本資料進行了探索性分析，包括乘客的性別、年齡、艙等等。在這些資料的探索中，我們發現女性和頭等艙的乘客有較高的存活率，而老年人和年輕人的存活率相對較低。另外，還使用了不同的機器學習模型對乘客的存活狀況進行預測，並且對這些模型的表現進行了評估。透過這些分析，可以幫助我們更好地理解資料集，並且進一步進行模型設計和優化。