**作業一 報告**

**資訊114 H44091196 洪茂菘**

**練習 1：改善決策樹分類模型**

這裡我分別採用了兩種方法來試著改善test acc 預測原始分數，以下會個別對改進方法及改進部分作介紹講解:

* **作法一: 使用超參數**

如下圖所示，這邊我仿照助教上課的教學來設定超參數，我將最佳化方法設定為 Gini Index；最大設定深度為 10，最多葉子個數則設定為 5^10(以上數字是我自己實測出來的)。

一張含有 文字, 字型, 螢幕擷取畫面, 設計 的圖片

自動產生的描述

使用後的前後結果差異如下:

一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述一張含有 文字, 字型, 印刷術 的圖片

自動產生的描述

如上兩張圖，第一張圖是使用超參數後的結果，第二張則是原本的範例，可以觀察到test accuracy從原本的0.7262提升至0.7821。

* **作法二: 增加更多的輸入特徵**



在助教提供的範例中，原先只有Sex, Age, Fare三個輸入特徵，我自己則是新增了後三個特徵來試著提升test acc 的預測原始分數，以下會簡單介紹我選擇該參數的原因:

1. **Pclass:**

根據資料解釋，Pclass屬性為鐵達尼號中乘客所搭乘的船艙等級，分為1,2,3，數字越小則代表越高級，因此，我認為待在艙等越高的乘客，其生存率也會越高，實際對測試資料統計下來的結果也是如此，故將其列入輸入特徵中。

1. **SibSp 和 Parch:**

根據資料，**SibSp代表的是「手足和配偶人數」；Parch則代表「父母及子女人數」**，由於這兩個屬性的性質較類似，所以我放在一起解釋。

依據我對測試資料的觀察與實際統計結果，我發現SibSp 和 Parch 人數越少的乘客(即同行家人越少)，其生存率較高，推測其原因可能是因為當有同行家人時，在發生船難的當下第一反應是設法去拯救自己的家人，進而導致自己的存活率降低了。由以上分析，我最終也選擇將這兩個屬性列入至輸入特徵中。

使用後的前後結果差異如下:

一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述

一張含有 文字, 字型, 印刷術 的圖片

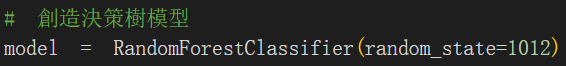
自動產生的描述

如上兩張圖，第一張圖是使用增加輸入特徵後的結果，第二張則是原本的範例，可以觀察到test accuracy從原本的0.7262略微提升至0.7430。

**練習2: 使用不同的模型**

第二小題則是要求我們使用不同的模型來試著超越原本的test accuracy數值。這邊我選了兩個分數表現較原本高的模型來呈現:

1. **ensemble.RandomForestClassifier模型**



一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述

如上兩張圖所示，我第一個選用的是集合學習中的RandomForestClassifier模型，可以看到其結果為0.7486，大於原本的0.7262。

1. **ensemble. ExtraTreesClassifier模型**



一張含有 文字, 字型, 螢幕擷取畫面, 印刷術 的圖片

自動產生的描述

如上兩張圖所示，我第二個選用的也是集合學習中的ExtraTreesClassifier模型，可以看到其結果為0.7430，大於原本的0.7262。