FDA hw2 decision tree

F74046030 李昕

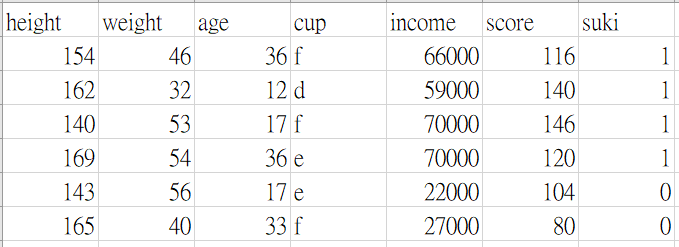
1. 目標

符合自身審美的個人擇偶標準預測器

1. 動機

因為長年寂寞，所以這次作業決定利用desision tree來嘗試實作一個根據自己的標準來評選女孩的個人擇偶標準預測器。首先使用create\_data檔案來隨機產生範圍內的2000名女孩簡單個人資料與數據，再利用這些資料分別給每一個女孩一個分數，最後根據這個分數來判斷擁有這些條件的女生會不會在我的守備範圍之內。建立起這樣的資料之後，再把女生的資料丟進decision tree，預測看看結果和一開始建立的結果會不會差很多。

1. Data



總共選擇了五種資料當作特徵，分別是身高、體重、年齡、月收入，還有cup。

身高的範圍選定在140~170公分。

體重的範圍是30~70公斤。

月收入的範圍在20000~80000之間，以千元為單位隨機產生。

年齡選在12~40歲，為了避免犯罪所以12歲以下就暫時先不考慮了。

cup根據出現的機率，決定把a~f都納入資料。f以後的大小出現的機率就比較小一些，所以在這裡決定先不考慮。

原本各項數據出現的機率應該要是不太一樣的，例如一個女生體重只有30公斤的機率應該會比體重40公斤出現的機率還要小。應該。但是這次為了簡化問題，就決定先把資料都設計成隨機產生。因此同一個特徵之中每一種數據出現的機率都會是一樣的。

分數的計算是將各項特徵再分成幾個區間，並根據我個人的喜好來為各個區間的數據給分數。畢竟是在做我的擇偶標準。例如將身高的資料分為140~150公分5分、150~160公分10分、160~170公分15分。另外再將每一個特徵的分數乘上一個權重，用這個權重來簡單的代表這一項特徵對於我來說的重要程度。例如身高和體重的權重都只有1，表示我並不是一個外貌協會的傢伙。而月收入的權重則給到3。這是一定要的啦。

五項特徵的權重分別是身高 \* 1，體重 \* 1，cup \* 2，月收 \* 3，年齡 \* 3。看看最後descition tree訓練出來的特徵重要程度會不會符合我們一開始這樣的設定。

最後把所有分數再分類成兩類。分數在一半的109分以上分類到喜歡，也就是suki欄位的1，而109分以下的就歸類到不喜歡，用suki欄位的0來代表。

1. Data preprocessing

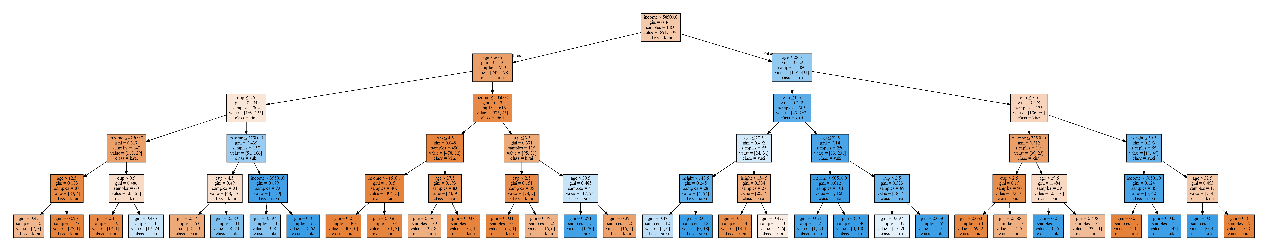
因為是自己產生的資料，所以還算是乾淨，比較不會有缺漏之類的狀況發生。只要把需要和不需要的資料區分開來，再把資料型態改好就行。

* 把score欄位拿掉
* 把資料型態是字元的cup改為用數字表示
* 把資料分成訓練資料和測試資料
* 分別從訓練資料和測試資料取出suki欄位的資料作為答案的y

1. 建立decision tree

改變max\_depth參數，測試結果

* Depth = 5

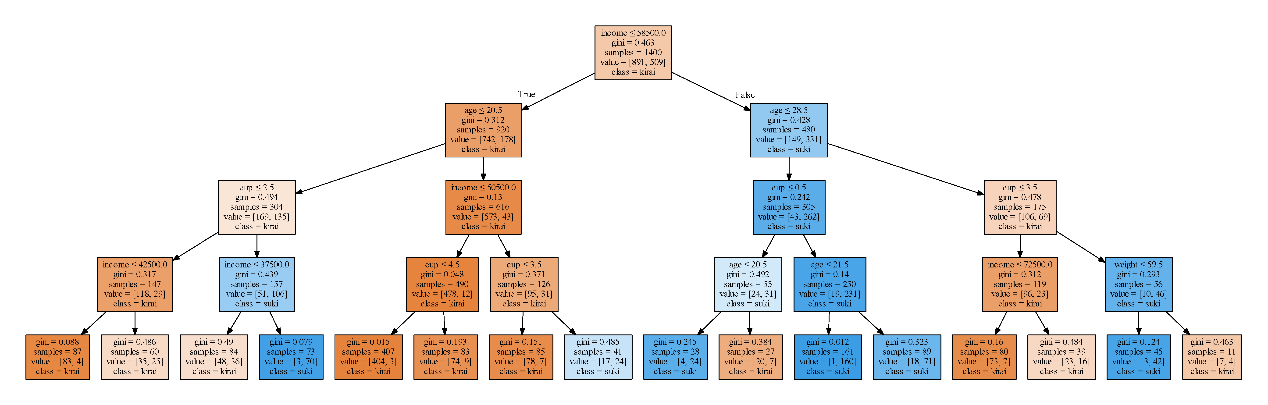


‧每一項特徵都用到

‧準確率0.903

‧準確率最高

* Depth = 4

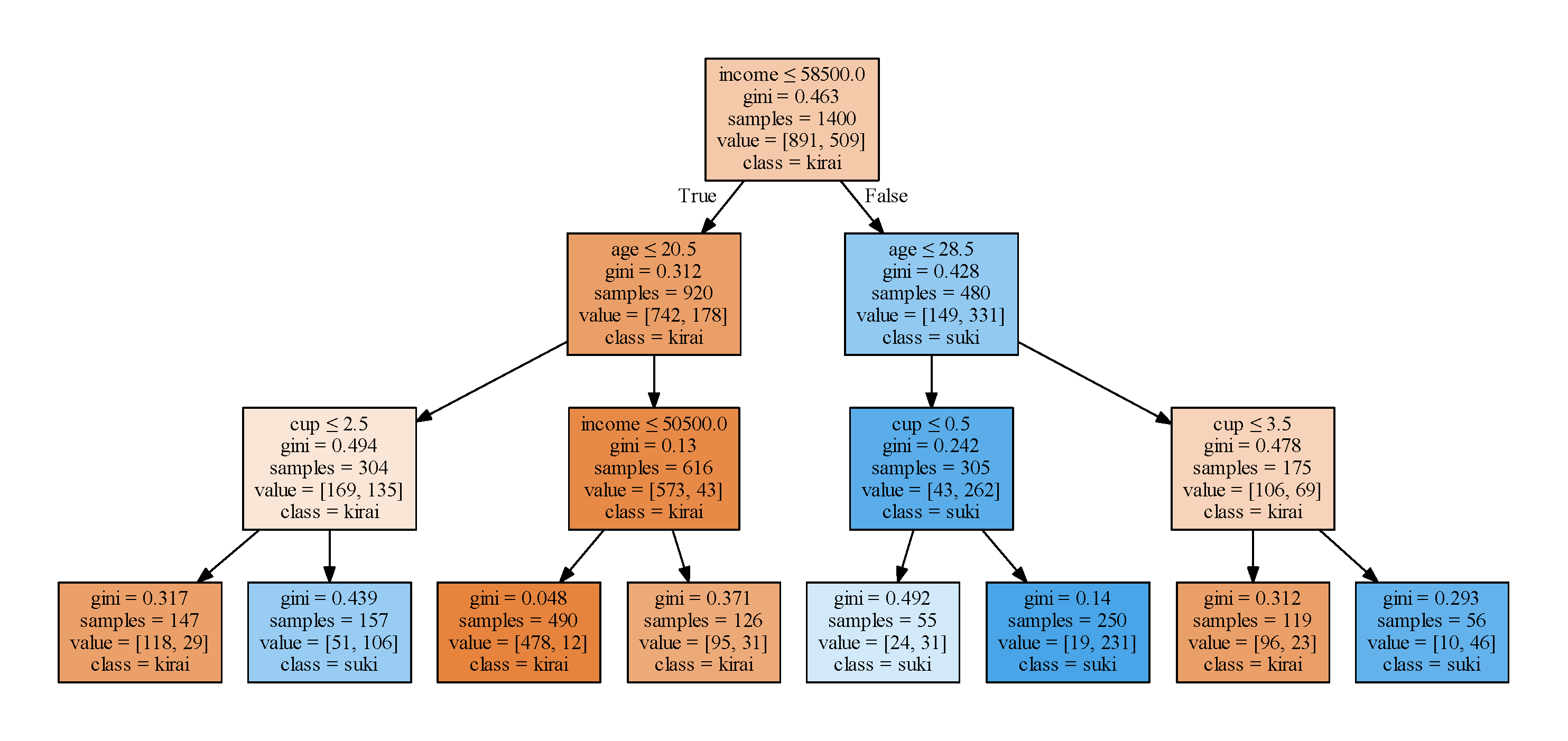


‧其中一個設定成最不重要的特徵height已經被捨棄了

‧準確率0.888

‧就算丟了一個特徵，準確率還是不低，可見這個特徵真的不算重要(0.903 -> 0.888)

* Depth = 3

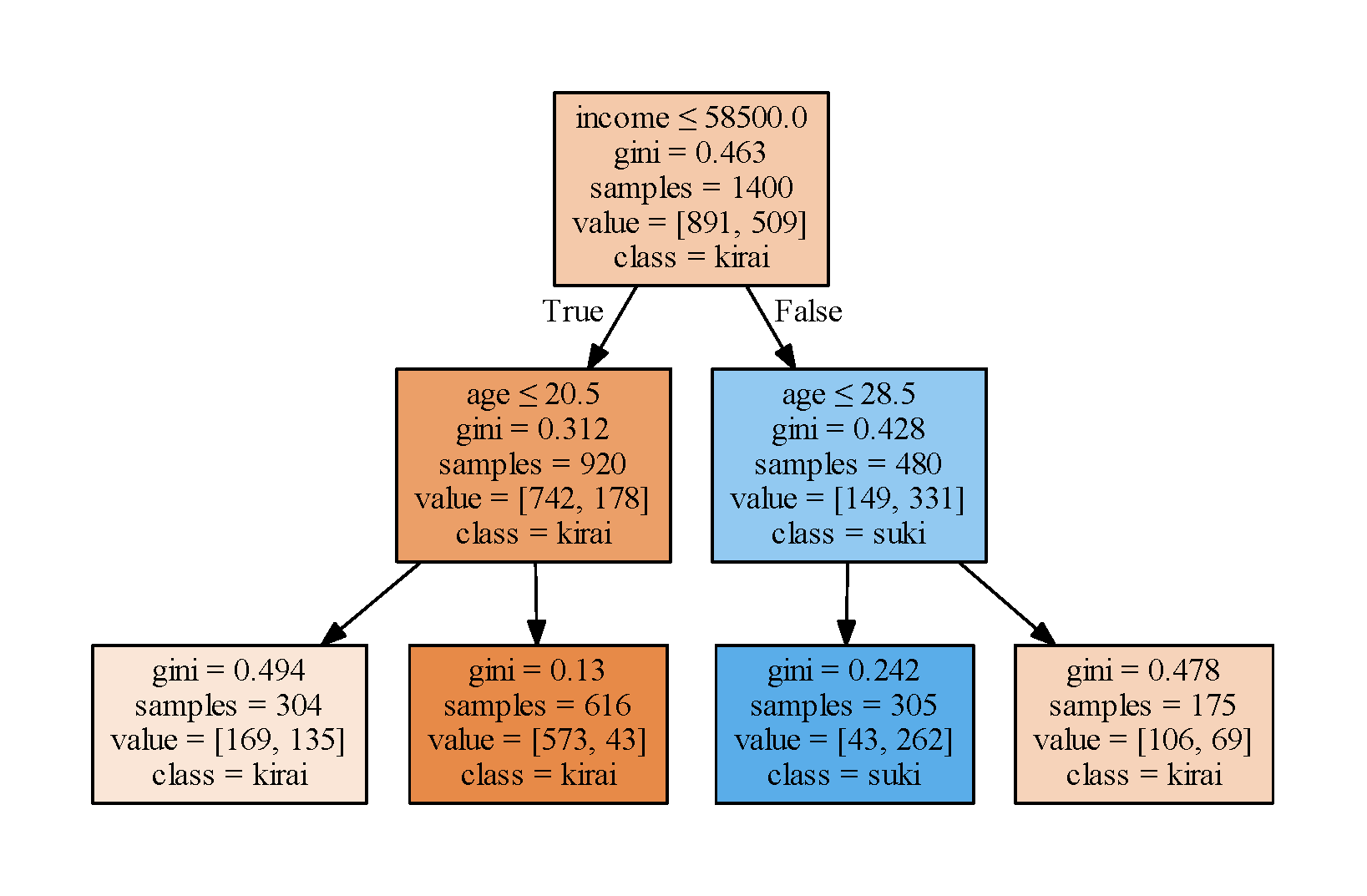


‧第二個設定成最不重要的特徵weight也被捨棄了

‧準確率0.848

‧雖然都不算重要，但還是拿掉了兩項特徵，準確率開始稍微明顯的下降(0.888 -> 0.848)

* Depth = 2

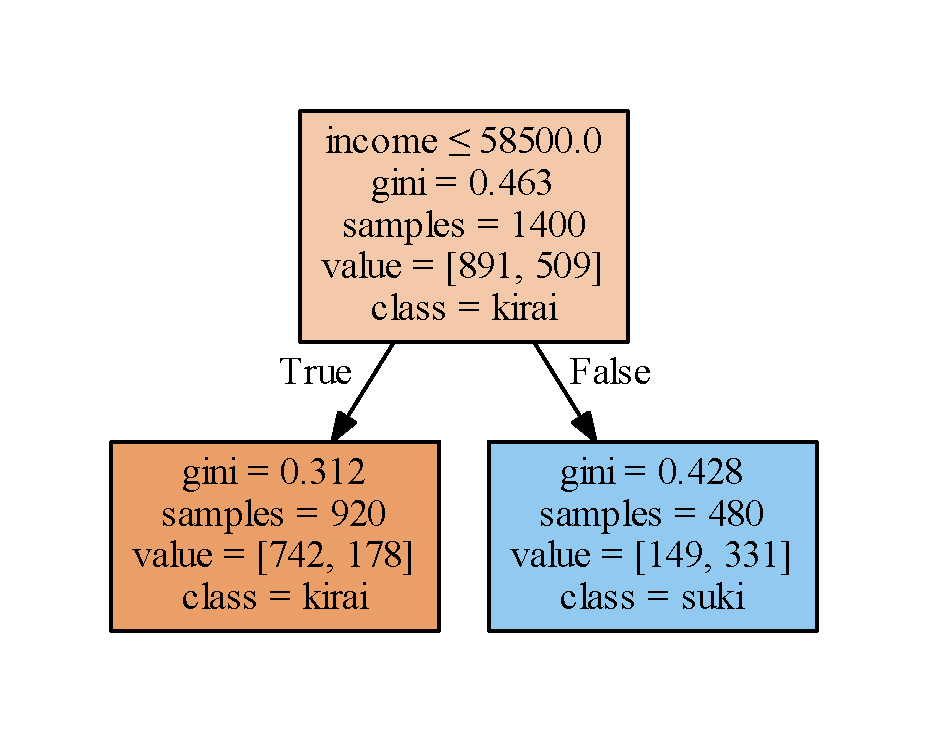


‧第二重要的特徵cup被放棄了

‧準確率0.788

‧準確率穩定下降(0.848 -> 0.788)

* Depth = 1



‧剩下最後的一個特徵income

‧準確率0.73

‧只用了一個特徵，但準確率比想像的還要高。可能是資料設計的有問題，導致有權重相同的特徵造成的影響程度不相同。

1. Random forest

讓很多decision tree來投票選出這個女生我喜不喜歡。這裡測試看看decision tree數量不同(n\_estimators參數的調整)，結果會有什麼變化。

* n\_estimators = 2(兩棵樹)

‧準確率0.888，跟一棵樹差不多準

* n\_estimators = 10(十棵樹)

‧數量增加到十棵樹，準確率有明顯的提升，來到了0.931

* n\_estimators = 100(一百棵樹)

‧一次把樹的數量增加到100，但結果準確率還是在0.94左右，看來已經是極限了，再增加也沒有用。

1. 結論
2. 特徵越多，準確率越高。但硬體負擔會越大。
3. 如果特徵夠重要，就可以有效分類。也就是選擇正確的特徵，可以更有效率的預測。
4. 只要月收入58500以上，就是我的菜!