Data Mining Project 1

資工碩一 P76084300 施逢怡

**Dataset**: Adult Data Set(<https://archive.ics.uci.edu/ml/datasets/Adult>)

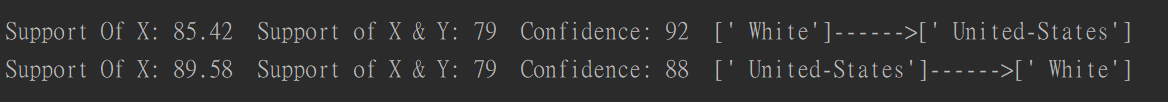
* Attributes: (age, workclass, education, marital-status, occupation, relationship,

race, sex, hrs-per-work, native-country, salary)

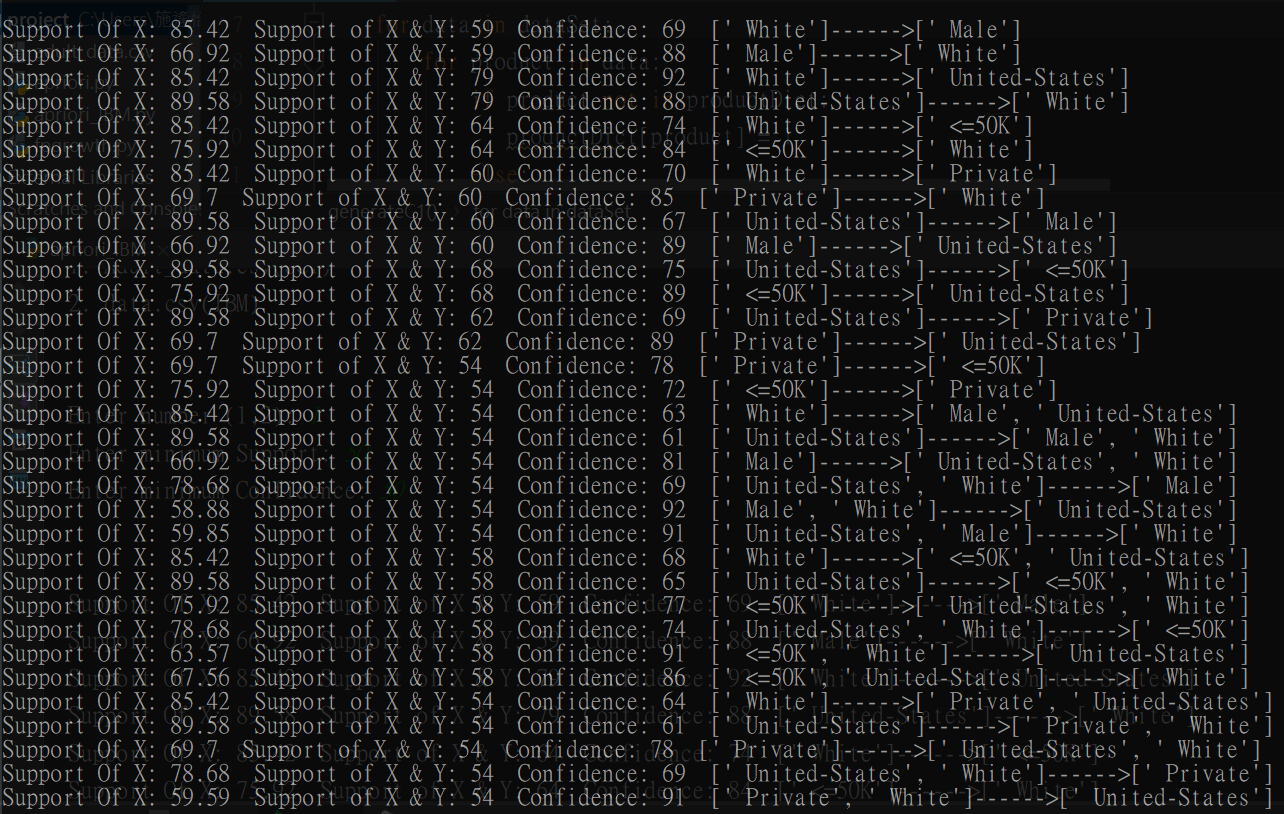
**Result**:

* **Apriori實作結果**

**(support, confidence) = (0.7,0.7)**

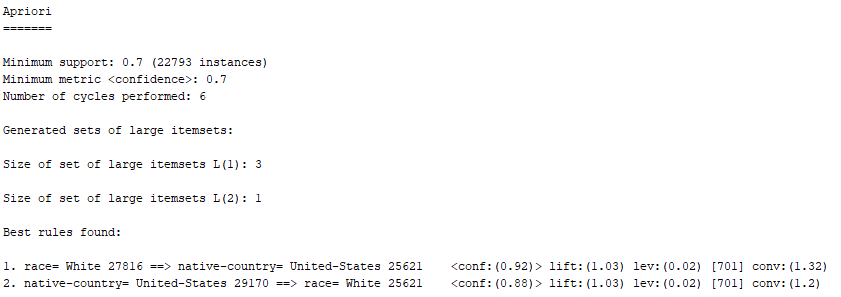
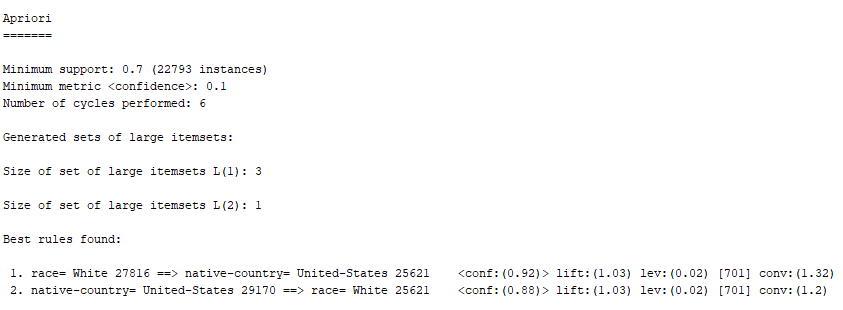


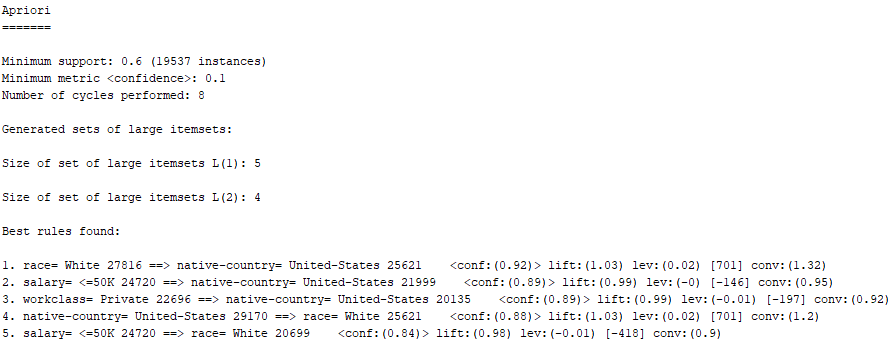
**(support, confidence) = (0.5,0.5)**



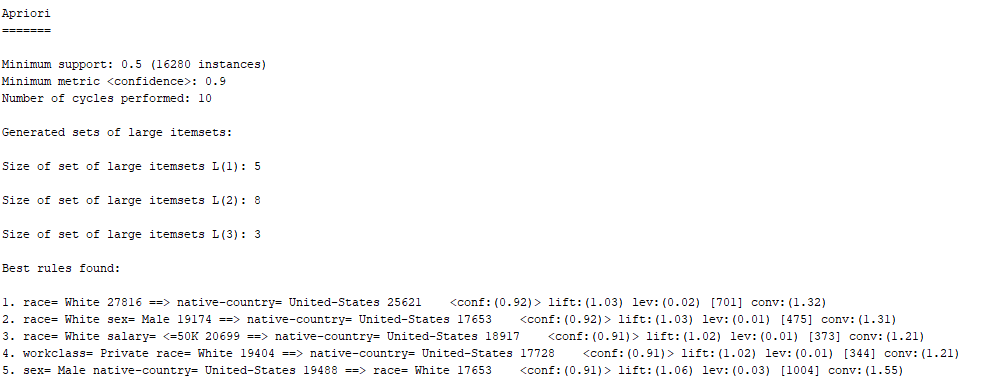
* **Apriori (by Weka)**

**//每次最多顯示5筆**

1. high support, high confidence (0.7, 0.7) 
2. high support, low confidence(0.7, 0.1) 
3. Low support, low confidence(0.1, 0.1)



(4) Low support, high confidence(0.1, 0.9)



* **Apriori比較與心得:**

這次使用了美國人關於職業、教育、性別、種族等的資料集，希望能透過關聯法則找出各個資料是否有關連。

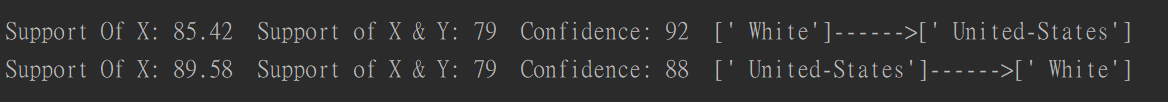
透過實作和Weka，我從(support, confidence)=(0.9,0.9)開始往下測試，發現support值太高，使得程式或是Weka都找不出關聯法則，直到support值調到0.7之後，找到2個關聯:United Stated->white和white->United Stated，代表的意義是在這個資料集裡面，國籍是美國的人通常是白人，且白人通常是住在美國。

當繼續調低support值之後，因為在每個階段中，進入下一個階段的關聯組合太多，除了因為計算量太大而使的程式執行速度太久，得出的關聯法則量也大量成長。

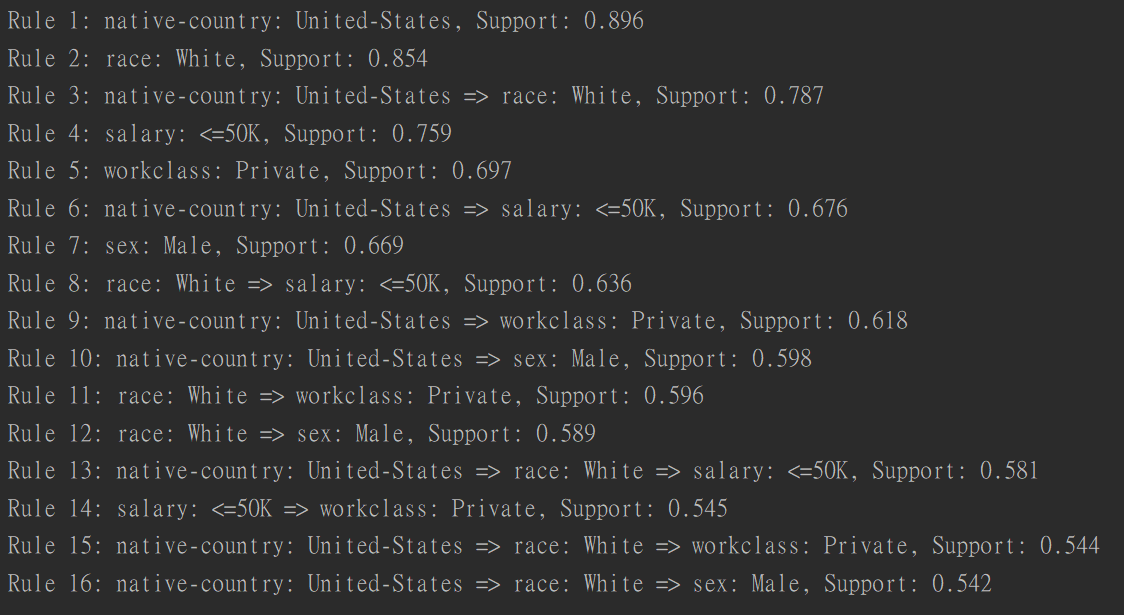
觀察其他的關聯之後，發現出來的法則往往都互相關聯(A->B且B->A)，而且感覺都環繞著幾個特徵(美國人，男人，白人，在私人企業工作)。

* **FPGrowth實作結果**

Support:0.8



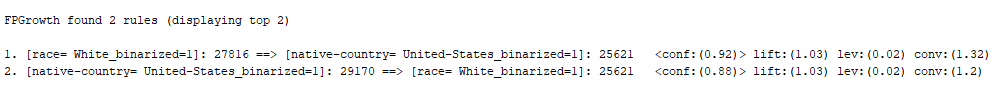
Support:0.5

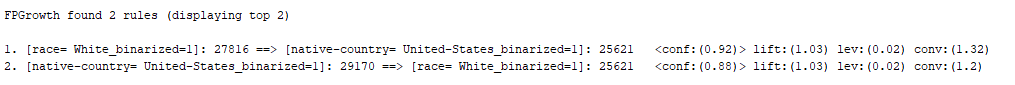
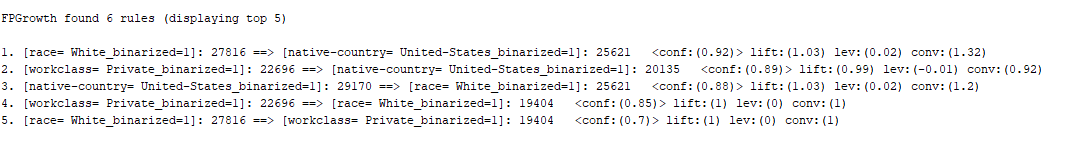
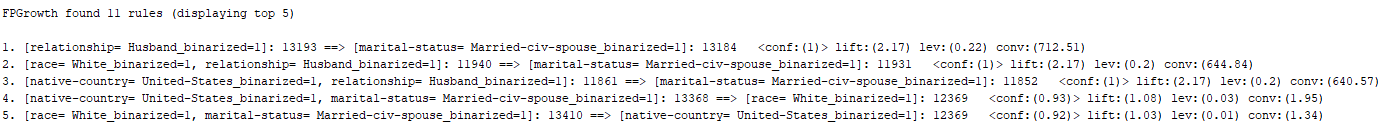


**FPGrowth**

**//每次最多顯示5筆**

1. high support, high confidence(0.7, 0.7)



1. high support, low confidence(0.7, 0.2)
2. Low support, low confidence(0.1, 0.1)
3. Low support, high confidence(0.2, 0.9)

* **FPGrowth比較與心得:**

執行FPGrowth時，明顯發現速度快了不少，FPGrowth算是個遞迴演算法，期間需要反覆遍歷樹和建構fp-tree，但是在直接計算單路徑產生所有組合的時候就會便捷很多。

執行結果在support=0.8時產生了2個關聯法則，和Apriori的結果一樣，都是United Stated->white和white->United Stated。

觀察support往下調後產生更多的關聯法則，發現結果和Apriori幾乎一樣，都是發現

(美國人，男人，白人，在私人企業工作)這些特徵關聯特別強。

反思Apriori與FPgrowth的結果，原本預期以為能發現工作和教育程度的關聯，或是年紀與婚姻之間的關係，結果在support較高的條件下幾乎沒有出現。我了解到資料集對於關聯法則有著重要的影響，使得某些潛在的關聯可能會被資料出現機率被過濾掉，如果想要對找出特定關聯的話，可能對資料集的選擇十分重要，亦或是資料集要經過前處理，對於影響演算法裡面權重太多的資料做sample或是正規化。

* **利用程式實作和Weka的結果來探討最後的問題: (high,low)與(support,confidence)各項組合產生的關聯法則**

1. High support, High confidence:

美國國籍，白人

1. High support, low confidence

白人，美國國籍

1. Low support, low confidence

白人，私人企業工作

1. Low support, high confidence

丈夫，已經結婚