1. **對 CreditCardPromotion 進行 Association Rule，並使用 Apriori 演算法，設定 confidence = 0.9、minimum support = 0.2，並回答以下問題：**

使用 Weka 軟體：

(a) 請嘗試著修改 CreditCardPromotion.arff 的欄位與上圖相同,使其可以執 行 Association Rule，請說明使用的方法以及解釋原來的檔案不能執行的原因? (10%)

因為 Sex 屬性的型態應該是 Nominal，而不是 Numeric，而原本屬於 Sex 屬性的資料皆為 0 跟 1，屬於 Numeric 而不是 Nominal，所以無法執行。

(b) 請將 numRule 設成 5 和 10,其各別執行後的 Minimum support 為何,請比較兩 者並說明造成其差異的原因。(15%)

當 numRule 設成 10，Minimum Support 為 0.25

```
Apriori
=======

Minimum support: 0.25 (3 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 18

Size of set of large itemsets L(3): 10

Size of set of large itemsets L(4): 1
```

當 numRule 設成 5，Minimum Support 為 0.35

```
Apriori
=======

Minimum support: 0.35 (3 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 8

Size of set of large itemsets L(3): 2
```

結果不同的原因：當要求較多的 Rule 時，資料之間的關聯相對會較不緊密，因此 Minimum support 會降低。

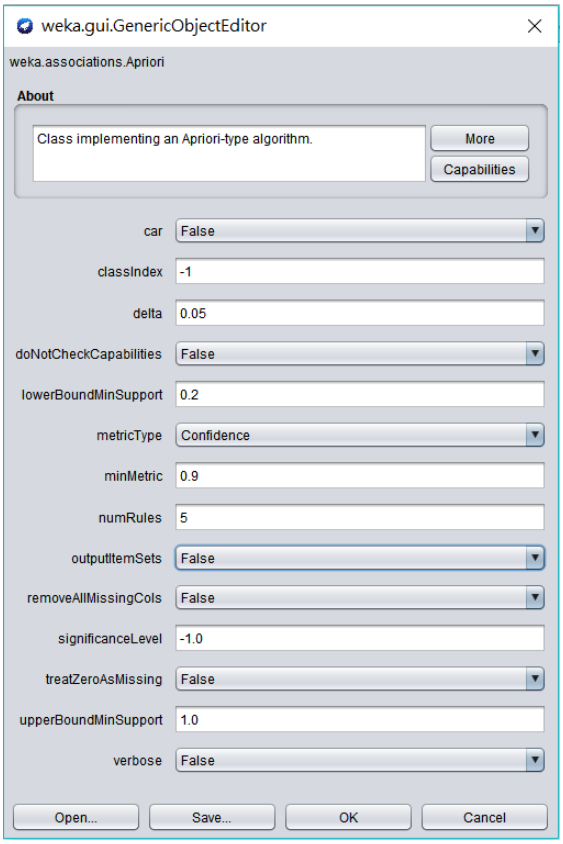
(c) 將 numRule 設成 10, 列出前 5 條 rule(15%)

```
Best rules found:

 1. Income Range=30-40000 4 ==> Magazine Promotion=Yes 4    <conf:(1)> lift:(1.43) lev:(0.12) [1] conv:(1.2)
 2. Sex=Female 4 ==> Credit Card Insurance=No 4    <conf:(1)> lift:(1.25) lev:(0.08) [0] conv:(0.8)
 3. Magazine Promotion=No 3 ==> Credit Card Insurance=No 3    <conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)
 4. Income Range=30-40000 Credit Card Insurance=No 3 ==> Magazine Promotion=Yes 3    <conf:(1)> lift:(1.43) lev:(0.09) [0] conv:(0.9)
 5. Magazine Promotion=Yes Sex=Female 3 ==> Credit Card Insurance=No 3    <conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)
```

(d) 如何在 Associator output 產生 Itemset, 請截圖說明並附上 Itemset 結果。(15%)

在參數設定的地方，將「outputItemSets」設為「True」，按下「OK」後「Start」



以下為 Itemset 的結果。

```
Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Large Itemsets L(1):
Income Range=20-30000 3
Income Range=30-40000 4
Magazine Promotion=Yes 7
Magazine Promotion=No 3
Credit Card Insurance=No 8
Sex=Male 6
Sex=Female 4

Size of set of large itemsets L(2): 8

Large Itemsets L(2):
Income Range=30-40000 Magazine Promotion=Yes 4
Income Range=30-40000 Credit Card Insurance=No 3
Magazine Promotion=Yes Credit Card Insurance=No 5
Magazine Promotion=Yes Sex=Male 4
Magazine Promotion=Yes Sex=Female 3
Magazine Promotion=No Credit Card Insurance=No 3
Credit Card Insurance=No Sex=Male 4
Credit Card Insurance=No Sex=Female 4

Size of set of large itemsets L(3): 2

Large Itemsets L(3):
Income Range=30-40000 Magazine Promotion=Yes Credit Card Insurance=No 3
Magazine Promotion=Yes Credit Card Insurance=No Sex=Female 3
```

在 Large Itemsets L(1)的部分：Weka 找出各屬性值「單一」存在的個數。例如：Income Range = 20-3000 的資料有 3 筆。

在 Large Itemsets L(2)的部分：由 Large Itemsets L(1)找出「任兩者」之間依賴關係的個數。例如：Income Range = 30-4000 且 Magazine Promotion= Yes 的資料有 4 筆。

在 Large Itemsets L(3)的部分：由 Large Itemsets L(1)找出「任三者」之間依賴關係的個數。例如：Income Range = 30-4000 且 Magazine Promotion= Yes 且 Credit Card Insurance = No 的資料有 3 筆。

使用 Python ：

(e) 將已修改過的 CreditCardPromotion.arff 轉成 csv 檔，使用 Apriori 演算法進行分析,設定 confidence = 0.9、minimum support = 0.2，過程中對所有重要程式步驟進行截圖並加以說明，越詳盡越好。(15%)

i. 首先，透過指令「pip install apyori」安裝與 import 引入所需套件，並讀取

CreditCardPromotion.csv 檔，將屬性變為 list。

安裝指令

pip install apyori

```
# import所需套件
# !pip install apyori
import pandas as pd
from apyori import apriori
```

讀取CSV檔

https://docs.python.org/3/library/csv.html

```
#hint:csv.reader
data = pd.read_csv('CreditCardPromotion.csv')
#將屬性合併
#變成list
feature=list(zip(data['Income Range'],data['Magizine Promotion'],data['Credit Card Insurance'],data['Sex']))
feature
```

```
[('40-50000', 'Yes', 'No', 'Male'),
 ('30-40000', 'Yes', 'No', 'Female'),
 ('40-50000', 'No', 'No', 'Male'),
 ('20-30000', 'Yes', 'Yes', 'Male'),
 ('50-60000', 'Yes', 'No', 'Female'),
 ('20-30000', 'No', 'No', 'Female'),
 ('30-40000', 'Yes', 'Yes', 'Male'),
 ('20-30000', 'No', 'No', 'Male'),
 ('30-40000', 'Yes', 'No', 'Male'),
 ('30-40000', 'Yes', 'No', 'Female')]
```

ii. 使用 aporiori()，將 min_support 設為 0.25，min_confidence 設為 0.9，再將結果轉為 list 並直接印出。

關聯式分析

https://pypi.org/project/efficient-apriori/

```
#hint:apriori()
association_rules  = apriori(feature, min_support=0.25,  min_confidence=0.9)
association_results = list(association_rules)

#print
for item in association_results:
    # first index of the inner list
    # Contains base item and add item
    pair = item[0]
    items = [x for x in pair]

    if len(items)<3 :
        print("Rule: " + items[0] + " -> " + items[1])
    else:
        print("Rule: " + items[0] + " -> " + items[1] + " -> " + items[2])

    #second index of the inner list
    print("Support: " + str(item[1]))

    #third index of the list located at 0th
    #of the third index of the inner list
    print("Confidence: " + str(item[2][0][2]))
    print("Lift: " + str(item[2][0][3]))
    print("====================================")
```

```
Rule: 30-40000 -> Yes
Support: 0.4
Confidence: 1.0
Lift: 1.4285714285714286
====================================
Rule: No -> Female
Support: 0.4
Confidence: 1.0
Lift: 1.25
====================================
Rule: 30-40000 -> No -> Yes
Support: 0.3
Confidence: 1.0
Lift: 1.4285714285714286
====================================
Rule: No -> Yes -> Female
Support: 0.3
Confidence: 1.0
Lift: 1.25
====================================
```

iii. 排序：將結果轉為 list 後再轉成 Dataframe 並使用，sort_values() 以 support 值進行排序(ascending = False 可讓排序由大至小)，再將結果轉為 list，並印出。

排序

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sort_values.html

```
#hint:darafrme.sort_values()
df = pd.DataFrame(association_results)
ordered = df.sort_values(by=['support'], ascending=False)
ordered_list = df.values.tolist()

for item in ordered_list:
    # first index of the inner list
    # Contains base item and add item
    pair = item[0]
    items = [x for x in pair]

    if len(items)<3:
        print("Rule: " + items[0] + " -> " + items[1] )
    else:
        print("Rule: " + items[0] + " -> " + items[1] +  " -> " + items[2] )

    #second index of the inner list
    print("Support: " + str(item[1]))

    #third index of the list located at 0th
    #of the third index of the inner list

    print("Confidence: " + str(item[2][0][2]))
    print("Lift: " + str(item[2][0][3]))
    print("====================================")
```

```
Rule: 30-40000 -> Yes
Support: 0.4
Confidence: 1.0
Lift: 1.4285714285714286
====================================
Rule: Female -> No
Support: 0.4
Confidence: 1.0
Lift: 1.25
====================================
Rule: 30-40000 -> Yes -> No
Support: 0.3
Confidence: 1.0
Lift: 1.4285714285714286
====================================
Rule: Yes -> Female -> No
Support: 0.3
Confidence: 1.0
Lift: 1.25
====================================
```

(f) 調整 apriori( )內的參數，產生與(c) 小題一樣的結果，截圖並加以說明(15%)

執行結果如下：

```
Rule: 30-40000 -> Yes
Support: 0.4
Confidence: 1.0
Lift: 1.4285714285714286
====================================
Rule: Female -> No
Support: 0.4
Confidence: 1.0
Lift: 1.25
====================================
Rule: 30-40000 -> Yes -> No
Support: 0.3
Confidence: 1.0
Lift: 1.4285714285714286
====================================
Rule: Yes -> Female -> No
Support: 0.3
Confidence: 1.0
Lift: 1.25
====================================
```

比較 Weka 執行結果（ numRule = 10 , confidence = 0.9, minimum support = 0.2 ）

```
Best rules found:

 1. Income Range=30-40000 4 ==> Magazine Promotion=Yes 4    <conf:(1)> lift:(1.43) lev:(0.12) [1] conv:(1.2)
 2. Sex=Female 4 ==> Credit Card Insurance=No 4    <conf:(1)> lift:(1.25) lev:(0.08) [0] conv:(0.8)
 3. Magazine Promotion=No 3 ==> Credit Card Insurance=No 3    <conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)
 4. Income Range=30-40000 Credit Card Insurance=No 3 ==> Magazine Promotion=Yes 3    <conf:(1)> lift:(1.43) lev:(0.09) [0] conv:(0.9)
 5. Magazine Promotion=Yes Sex=Female 3 ==> Credit Card Insurance=No 3    <conf:(1)> lift:(1.25) lev:(0.06) [0] conv:(0.6)
```

說明：相較於 Weka，Python 的執行結果只有四筆，而 Weka 的第四筆在 Python 中是不存在的，其餘的 Rule 在兩者之間都是一樣的。

(g) 請自己計算 (記錄在 Word 上或手算拍照附圖皆可),並與 (d) 小題結果做驗證。(15%)

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

i. 首先找出 Large Itemsets L(1)

Income Range = 20-3000，3 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Income Range = 30-4000，4 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Income Range = 40-5000，2 筆，數值太小不列入

Magazine Promotion = Yes，7 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Magazine Promotion = No，3 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Credit Card Insurance = No，8 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Credit Card Insurance = Yes，2 筆，數值太小不列入

Sex = Male，6 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Sex = Female，4 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

ii. 由 i. 找出列出所有可能的 Large Itemsets L(2)

Income Range = 20-3000，Magazine Promotion = Yes，1 筆，不列入

Income Range = 20-3000，Magazine Promotion = No，2 筆，不列入

Income Range = 20-3000，Credit Card Insurance = No，2 筆，不列入

Income Range = 20-3000，Sex = Male，2 筆，不列入

Income Range = 20-3000，Sex = Female，1 筆，不列入

Income Range = 30-4000，Magazine Promotion = Yes，4 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Income Range = 30-4000，Magazine Promotion = No，0 筆，不列入

Income Range = 30-4000，Credit Card Insurance = No，3 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Income Range = 30-4000，Sex = Male，2 筆，不列入

Income Range = 30-4000，Sex = Female，2 筆，不列入

Magazine Promotion = Yes，Credit Card Insurance = No，5 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Magazine Promotion = Yes，Sex = Male，4 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Magazine Promotion = Yes，Sex = Female，3 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Magazine Promotion = No，Credit Card Insurance = No，3 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Magazine Promotion = No，Sex = Male，2 筆，不列入

Magazine Promotion = No，Sex = Female，1 筆，不列入

Credit Card Insurance = No，Sex = Male，4 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Credit Card Insurance = No，Sex = Female，4 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

iii. 由 ii. 找出列出所有可能的 Large Itemsets L(3)

Income Range = 30-4000，Magazine Promotion = Yes，Credit Card Insurance = No，3 筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Income Range = 30-4000，Magazine Promotion = Yes，Sex = Male，2 筆，不列入

Income Range = 30-4000，Magazine Promotion = Yes，Sex = Female，2 筆，不列入

Income Range = 30-4000，Credit Card Insurance = No，Sex = Male，1 筆，不列入

Income Range = 30-4000，Credit Card Insurance = No，Sex = Female，2筆，不列入

Magazine Promotion = Yes，Credit Card Insurance = No，Sex = Male，2筆，不列入

Magazine Promotion = Yes，Credit Card Insurance = No，Sex = Female，3筆

Income Range	Magizine Promotion	Credit Card Insurance	Sex
40-50000	Yes	No	Male
30-40000	Yes	No	Female
40-50000	No	No	Male
20-30000	Yes	Yes	Male
50-60000	Yes	No	Female
20-30000	No	No	Female
30-40000	Yes	Yes	Male
20-30000	No	No	Male
30-40000	Yes	No	Male
30-40000	Yes	No	Female

Magazine Promotion = No，Credit Card Insurance = No，Sex = Male，2筆，不列入

Magazine Promotion = No，Credit Card Insurance = No，Sex = Female，1筆，不列入