電商技術 HW3 第三部分

1.載入 customer_churn.csv，列出資料筆數、屬性數量以及每個欄位的空值個數

customer_churn.csv 原始資料共3083筆、屬性數20:

Current relation

Relation: customer_churn	Attributes: 20
Instances: 3083	Sum of weights: 3083

以下為每個欄位的空值個數：

1. CustomerID 無空值

Selected attribute

Name: CustomerID		Type: Numeric
Missing: 0 (0%)	Distinct: 3078	Unique: 3075 (100%)

Statistic	Value
Minimum	50001
Maximum	55628
Mean	52807.294
StdDev	1617.13

2. Churn 無空值

Selected attribute

Name: Churn		Type: Numeric
Missing: 0 (0%)	Distinct: 2	Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.308
StdDev	0.462

3. Tenure 有153筆空值

Selected attribute

Name: Tenure | Type: Numeric

Missing: 153 (5%) | Distinct: 35 | Unique: 3 (0%)

Statistic	Value
Minimum	0
Maximum	61
Mean	9.109
StdDev	8.549

4. PreferredLoginDevice 無空值

Selected attribute

Name: PreferredLoginDevice | Type: Nominal

Missing: 0 (0%) | Distinct: 3 | Unique: 0 (0%)

No.	Label	Count	Weight
1	Computer	928	928
2	Mobile Phone	1458	1458
3	Phone	697	697

5. CityTier 無空值

Selected attribute

Name: CityTier | Type: Numeric

Missing: 0 (0%) | Distinct: 3 | Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	3
Mean	1.682
StdDev	0.925

6. WarehouseToHome 154筆空值

Selected attribute

Name: WarehouseToHome　Type: Numeric

Missing: 154 (5%)　Distinct: 33　Unique: 1 (0%)

Statistic	Value
Minimum	5
Maximum	126
Mean	15.771
StdDev	8.557

7. PreferredPaymentMode 無空值

Selected attribute

Name: PreferredPaymentMode　Type: Nominal

Missing: 0 (0%)　Distinct: 7　Unique: 0 (0%)

No.	Label	Count	Weight
1	Debit Card	1255	1255
2	Credit Card	789	789
3	E wallet	368	368
4	CC	147	147
5	COD	225	225
6	UPI	222	222
7	Cash on Delivery	77	77

8. Gender 無空值

Selected attribute

Name: Gender　Type: Nominal

Missing: 0 (0%)　Distinct: 2　Unique: 0 (0%)

No.	Label	Count	Weight
1	Male	1854	1854
2	Female	1229	1229

9. HourSpendOnApp 150筆空值

Selected attribute

Name: HourSpendOnApp | Type: Numeric
Missing: 150 (5%) | Distinct: 6 | Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	5
Mean	2.946
StdDev	0.719

10. NumberOfDeviceRegistered 無空值

Selected attribute

Name: NumberOfDeviceRegistered | Type: Numeric
Missing: 0 (0%) | Distinct: 6 | Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	6
Mean	3.75
StdDev	1.005

11. PreferredOrderCat 無空值

Selected attribute

Name: PreferedOrderCat | Type: Nominal
Missing: 0 (0%) | Distinct: 6 | Unique: 0 (0%)

No.	Label	Count	Weight
1	Mobile Phone	772	772
2	Grocery	198	198
3	Laptop & Accessory	1066	1066
4	Mobile	463	463
5	Fashion	449	449
6	Others	135	135

12. SatisfactionScore 無空值

Selected attribute

Name: SatisfactionScore | Type: Numeric

Missing: 0 (0%) | Distinct: 5 | Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	5
Mean	3.134
StdDev	1.382

13. MaritalStatus 無空值

Selected attribute

Name: MaritalStatus | Type: Nominal

Missing: 0 (0%) | Distinct: 3 | Unique: 0 (0%)

No.	Label	Count	Weight
1	Single	1102	1102
2	Married	1547	1547
3	Divorced	434	434

14. NumberOfAddress 無空值

Selected attribute

Name: NumberOfAddress | Type: Numeric

Missing: 0 (0%) | Distinct: 14 | Unique: 3 (0%)

Statistic	Value
Minimum	1
Maximum	21
Mean	4.335
StdDev	2.663

15. Complain 無空值

Selected attribute

Name: Complain　　Type: Numeric

Missing: 0 (0%)　　Distinct: 2　　Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.328
StdDev	0.47

16. OrderAmountHikeFromlastYear 131筆空值

Selected attribute

Name: OrderAmountHikeFromlastYear　　Type: Numeric

Missing: 131 (4%)　　Distinct: 16　　Unique: 0 (0%)

Statistic	Value
Minimum	11
Maximum	26
Mean	15.714
StdDev	3.765

17. CouponUsed 126筆空值

Selected attribute

Name: CouponUsed　　Type: Numeric

Missing: 126 (4%)　　Distinct: 17　　Unique: 3 (0%)

Statistic	Value
Minimum	0
Maximum	16
Mean	1.753
StdDev	1.886

18. OrderCount 128筆空值

Selected attribute

Name: OrderCount　　Type: Numeric

Missing: 128 (4%)　　Distinct: 16　　Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	16
Mean	2.994
StdDev	2.948

19. DaySinceLastOrder 166筆空值

Selected attribute

Name: DaySinceLastOrder　　Type: Numeric

Missing: 166 (5%)　　Distinct: 20　　Unique: 1 (0%)

Statistic	Value
Minimum	0
Maximum	46
Mean	4.349
StdDev	3.609

20. CashbackAmount 無空值

Selected attribute

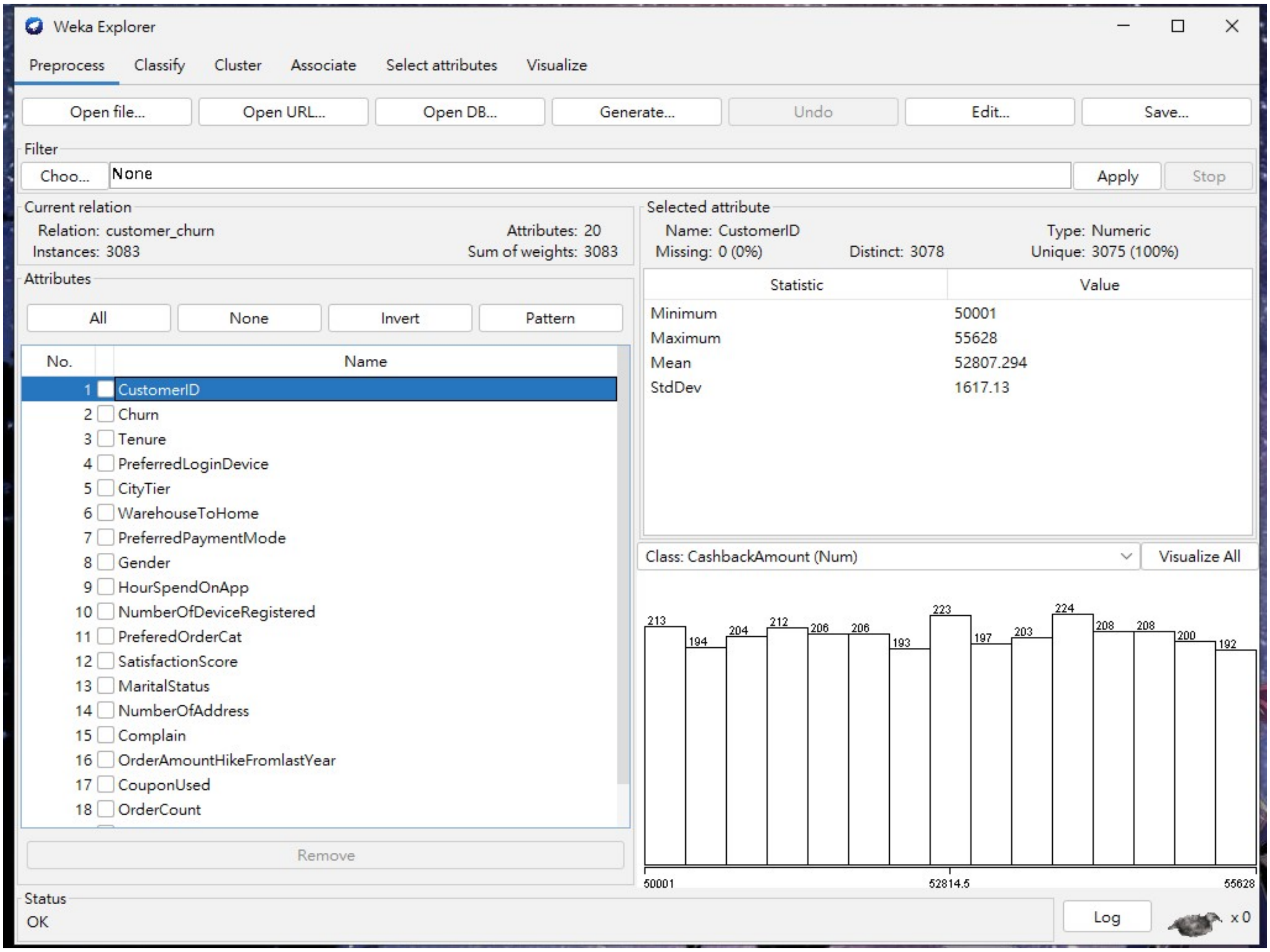
Name: CashbackAmount　　Type: Numeric

Missing: 0 (0%)　　Distinct: 212　　Unique: 20 (1%)

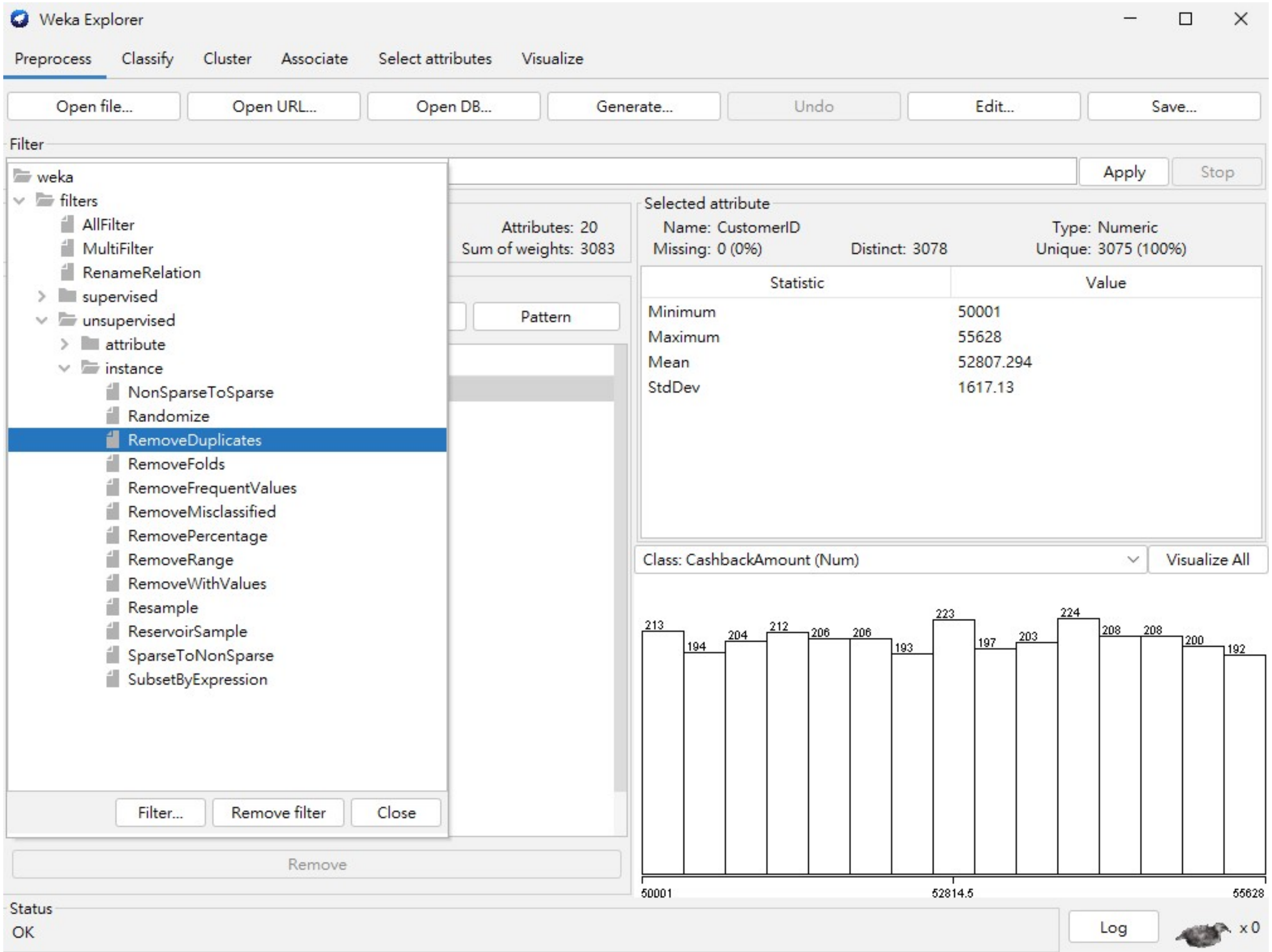
Statistic	Value
Minimum	0
Maximum	324
Mean	174.904
StdDev	47.829

2.請刪除重覆多餘的資料 (僅保留一筆)，並列出剩餘的資料筆數

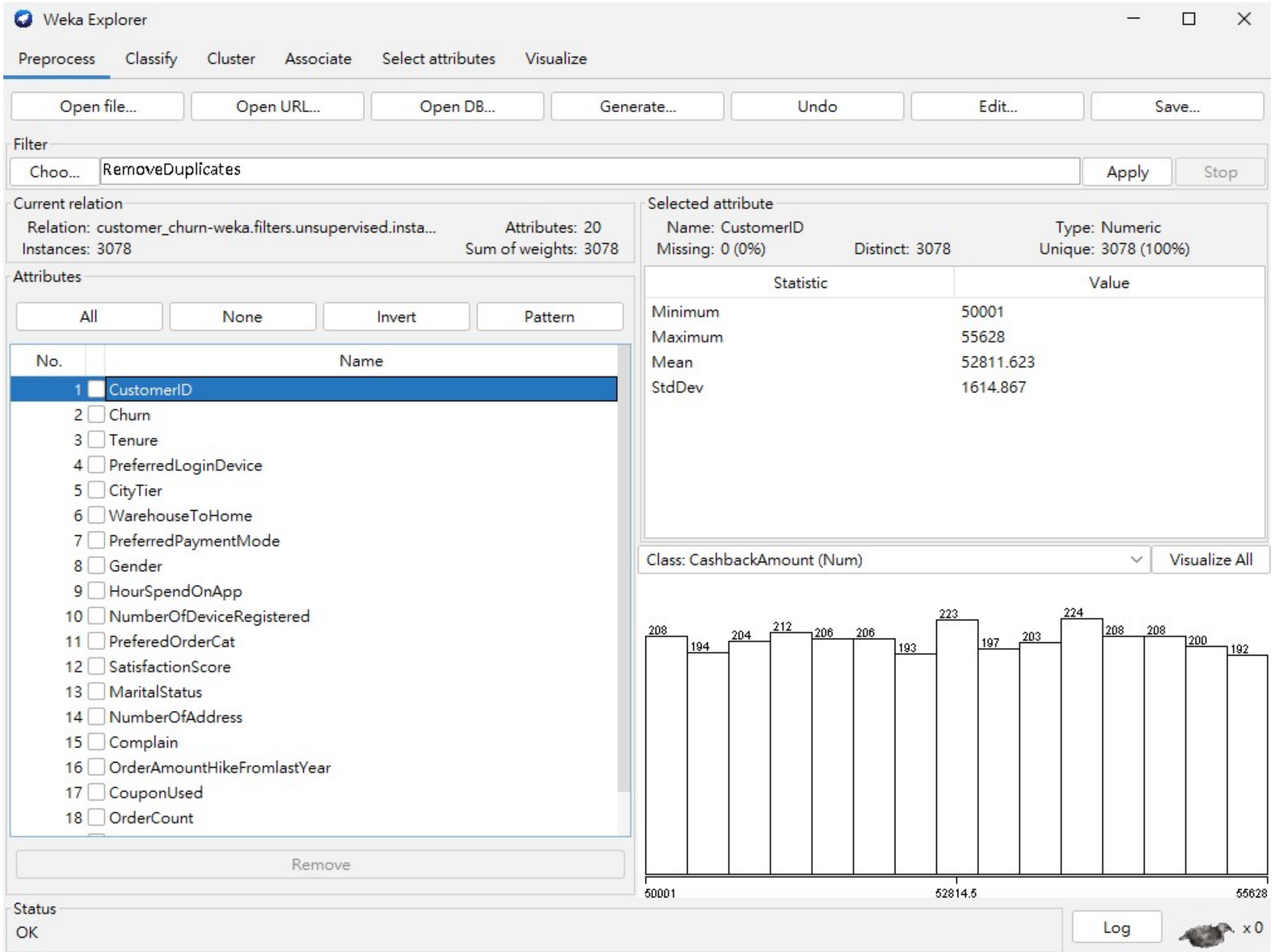
有重複的 CustomerID 表示有重複的紀錄，必須將其刪除：



透過 fiter 找到 RemoveDuplicates 刪去重複資料：



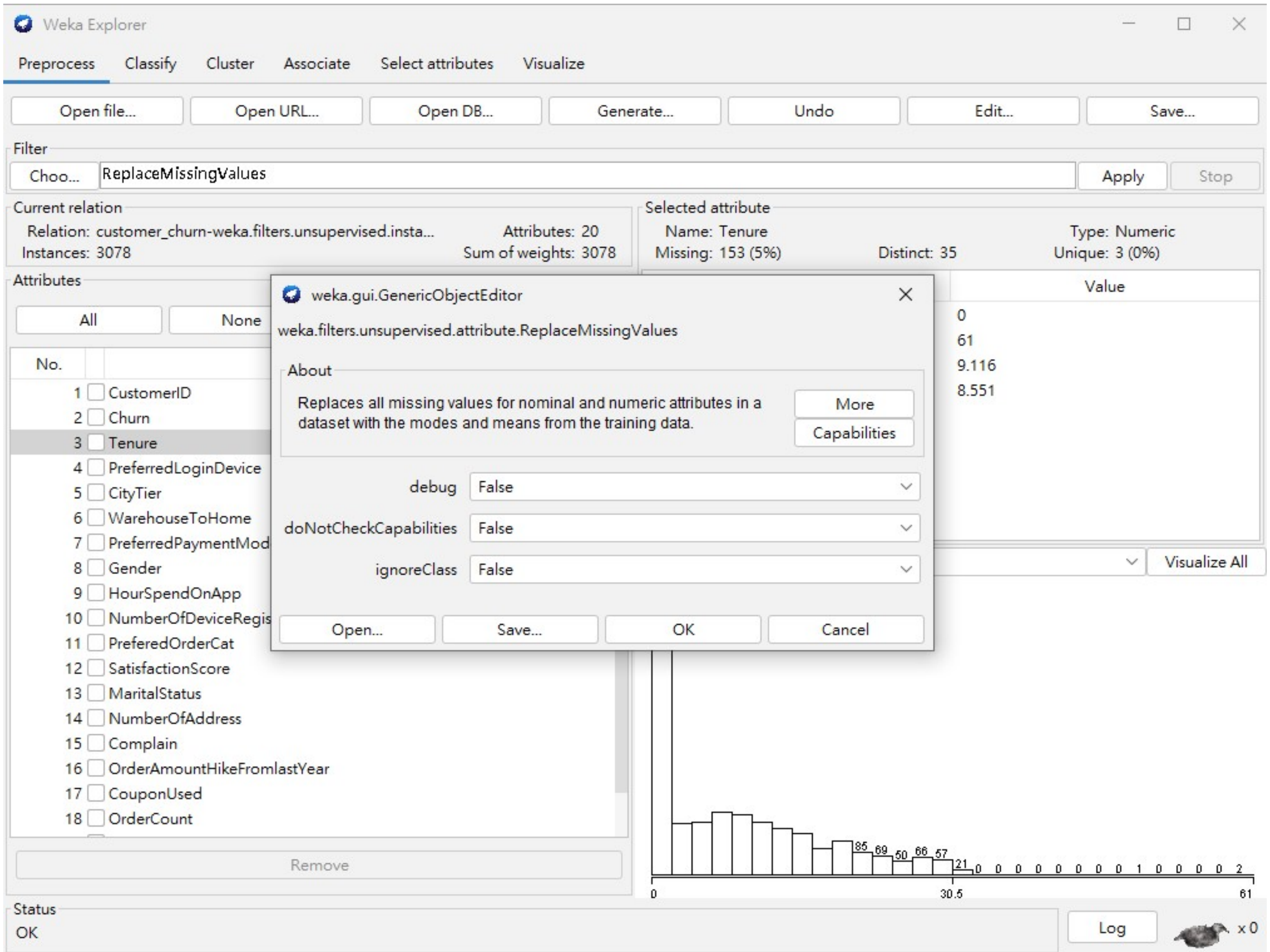
Apply 之後，可以觀察到已經沒有重複的 CustomerID，亦即沒有重複的資料：



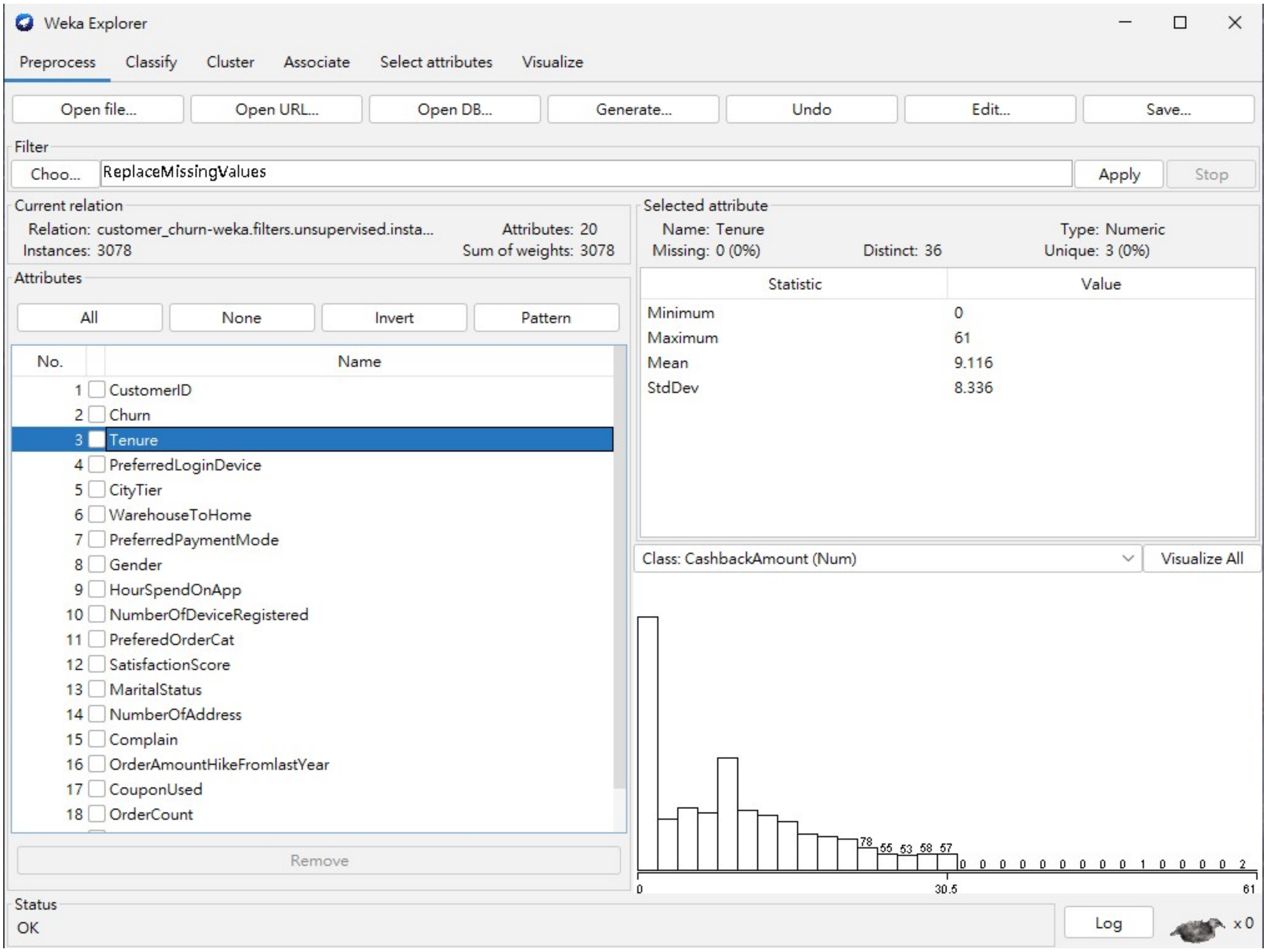
3.資料前處理

填補 Missing Value

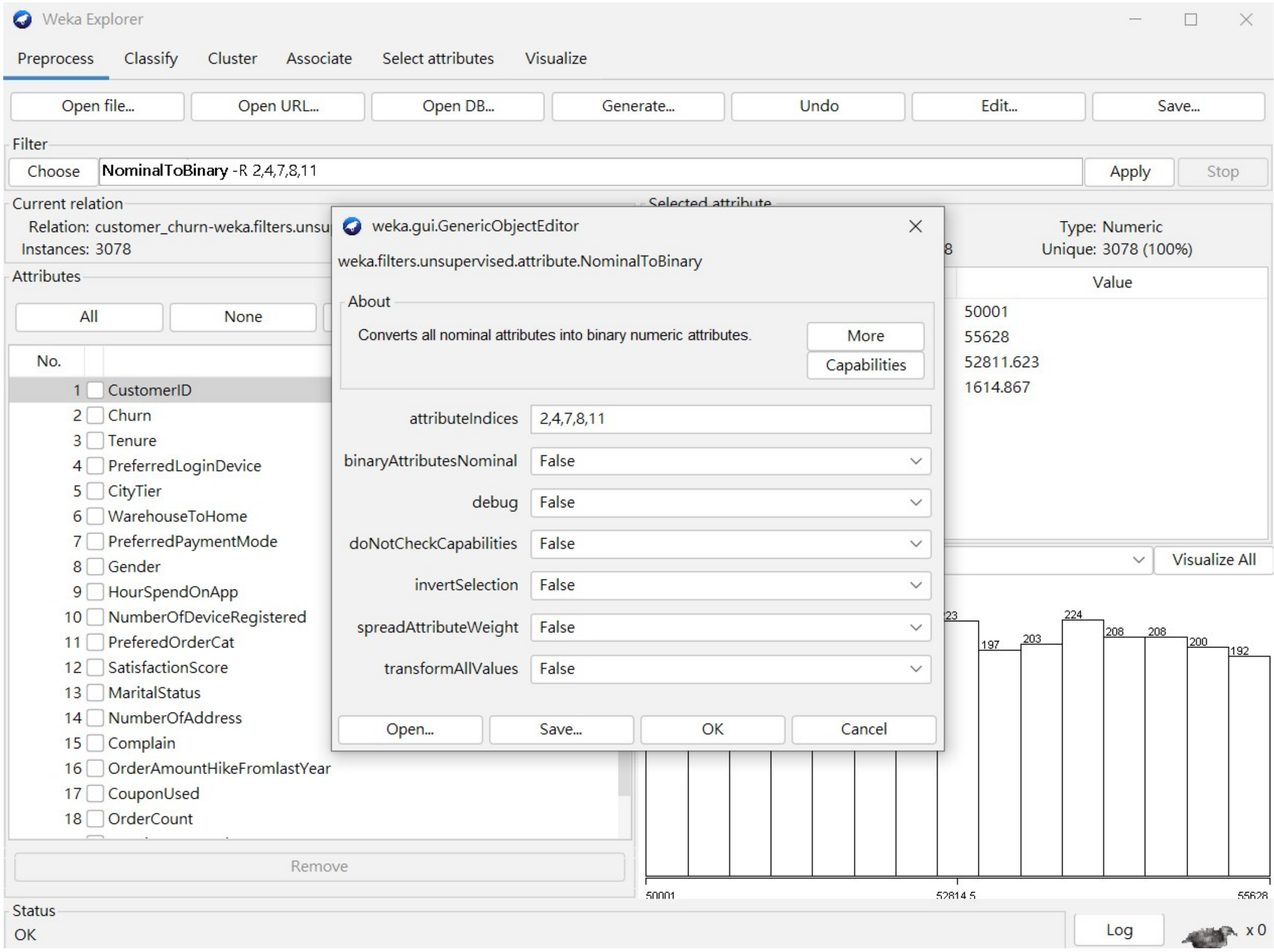
nominal 以 mode 填補 missing value
numeric 以 mean 填補 missing value

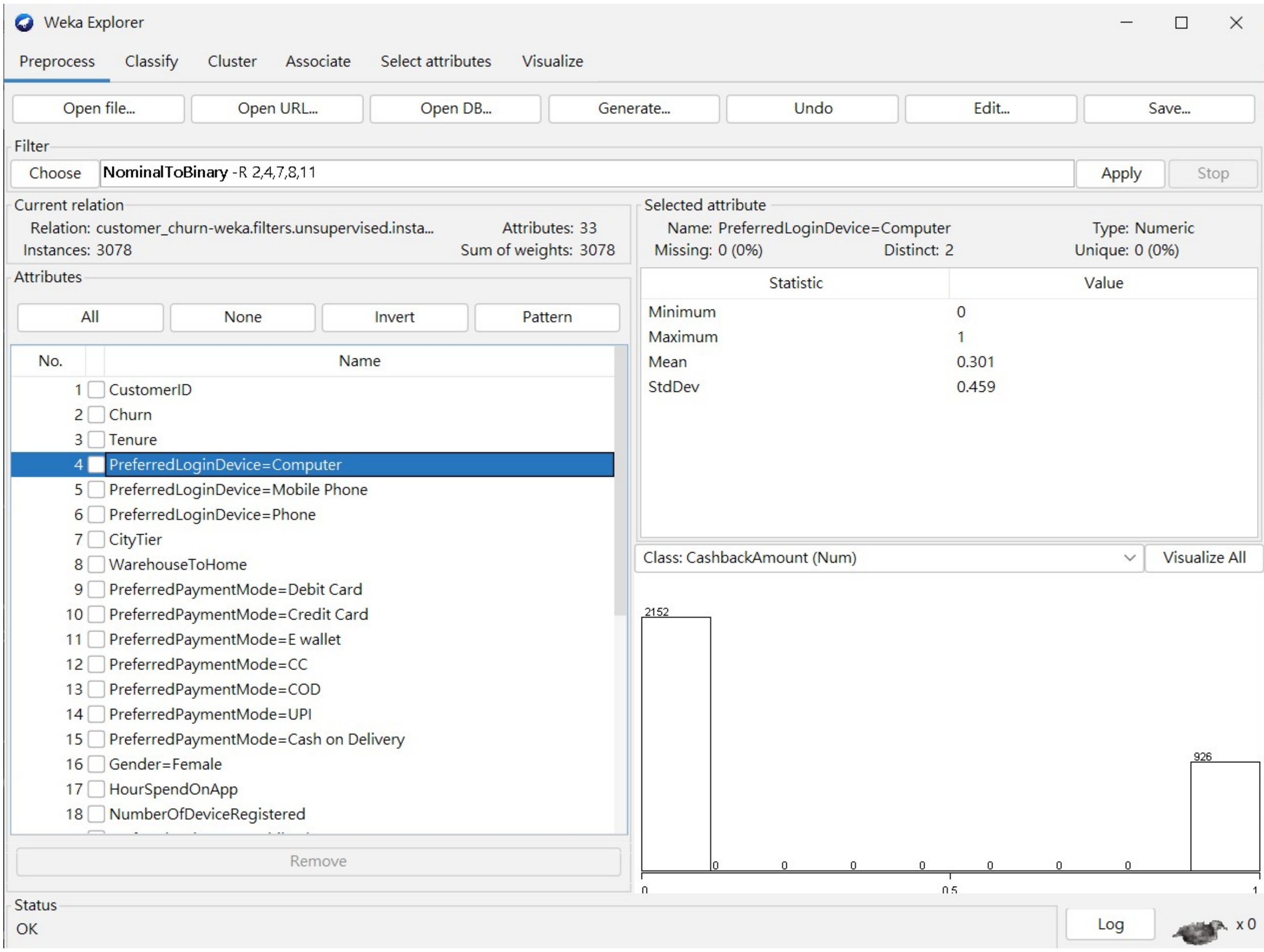


以 Tenure 為例，missing value 都不見了，其餘屬性也是如此

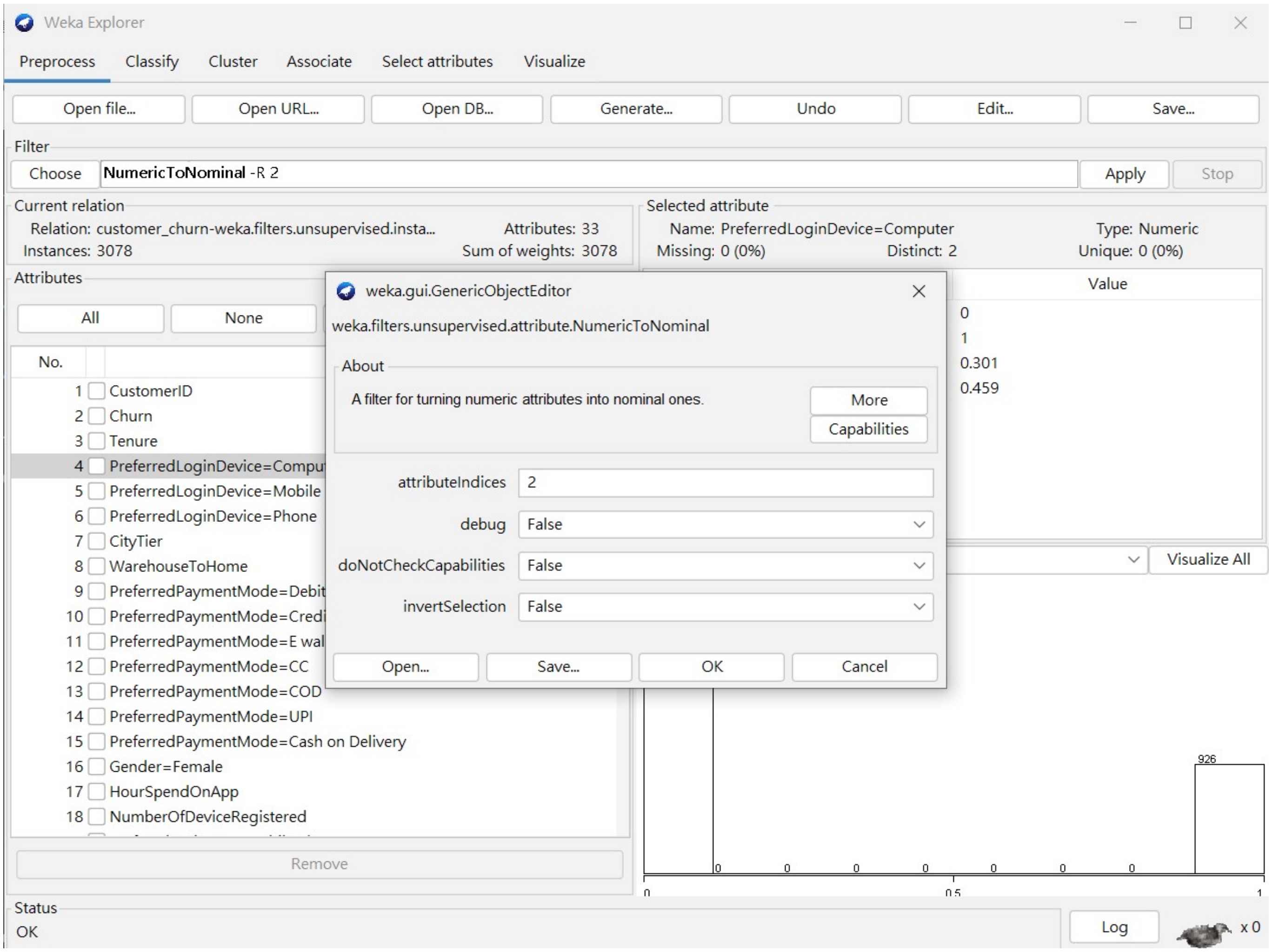


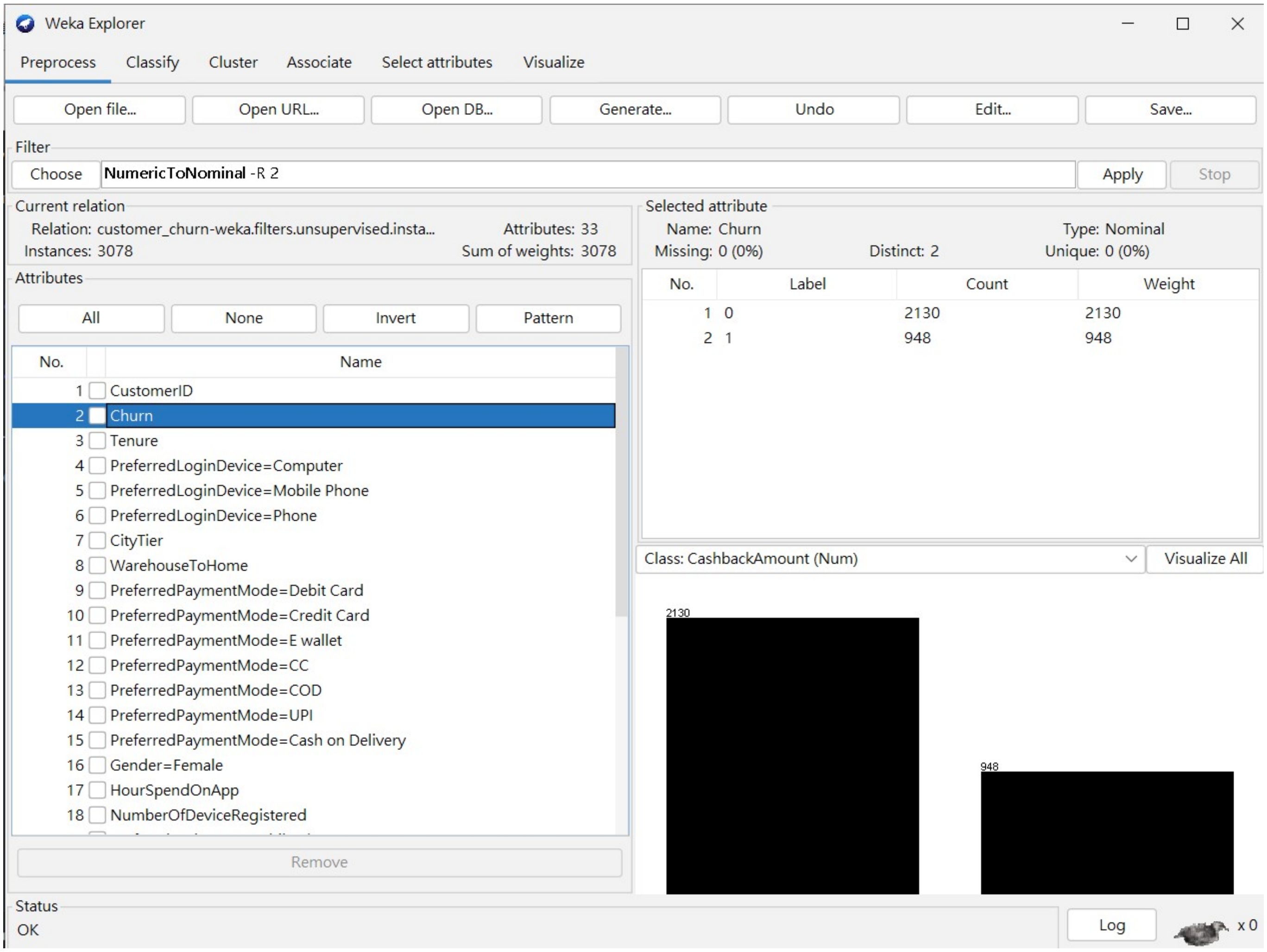
接下來將 categorical 屬性值都轉成 numeric value，我選擇的是 filter 中的 NominalToBinary，在屬性值不是二元可分的情況，其實就跟 Python 的 pandas.get_dummies 很像，把屬性變得更多了。


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
NominalToBinary -R 2,4,7,8,11
Apply
Stop
Current relation
Relation: customer_churn-weka.filters.unsu
Instances: 3078
Attributes
All
None
No.
1 CustomerID
2 Churn
3 Tenure
4 PreferredLoginDevice
5 CityTier
6 WarehouseToHome
7 PreferredPaymentMode
8 Gender
9 HourSpendOnApp
10 NumberOfDeviceRegistered
11 PreferedOrderCat
12 SatisfactionScore
13 MaritalStatus
14 NumberOfAddress
15 Complain
16 OrderAmountHikeFromlastYear
17 CouponUsed
18 OrderCount
Remove
Type: Numeric
Unique: 3078 (100%)
Value
50001
55628
52811.623
1614.867
Visualize All
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.NominalToBinary
About
Converts all nominal attributes into binary numeric attributes.
More
Capabilities
attributeIndices 2,4,7,8,11
binaryAttributesNominal False
debug False
doNotCheckCapabilities False
invertSelection False
spreadAttributeWeight False
transformAllValues False
Open...
Save...
OK
Cancel
Status
OK
Log
x 0



最後，把 churn 用 filter 中的 NumericToNominal 就可以套用 SVM、Logistic Regression、Decision Tree 來預測 churn 了。


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
NumericToNominal -R 2
Apply
Stop
Current relation
Relation: customer_churn-weka.filters.unsupervised.insta...
Attributes: 33
Instances: 3078
Sum of weights: 3078
Selected attribute
Name: PreferredLoginDevice=Computer
Type: Numeric
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)
Attributes
All
None
No.
1 CustomerID
2 Churn
3 Tenure
4 PreferredLoginDevice=Compu
5 PreferredLoginDevice=Mobile
6 PreferredLoginDevice=Phone
7 CityTier
8 WarehouseToHome
9 PreferredPaymentMode=Debit
10 PreferredPaymentMode=Credi
11 PreferredPaymentMode=E wal
12 PreferredPaymentMode=CC
13 PreferredPaymentMode=COD
14 PreferredPaymentMode=UPI
15 PreferredPaymentMode=Cash on Delivery
16 Gender=Female
17 HourSpendOnApp
18 NumberOfDeviceRegistered
Remove
Value
0
1
0.301
0.459
Visualize All
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.NumericToNominal
About
A filter for turning numeric attributes into nominal ones.
More
Capabilities
attributeIndices 2
debug False
doNotCheckCapabilities False
invertSelection False
Open...
Save...
OK
Cancel
926
Status
OK
Log
x 0

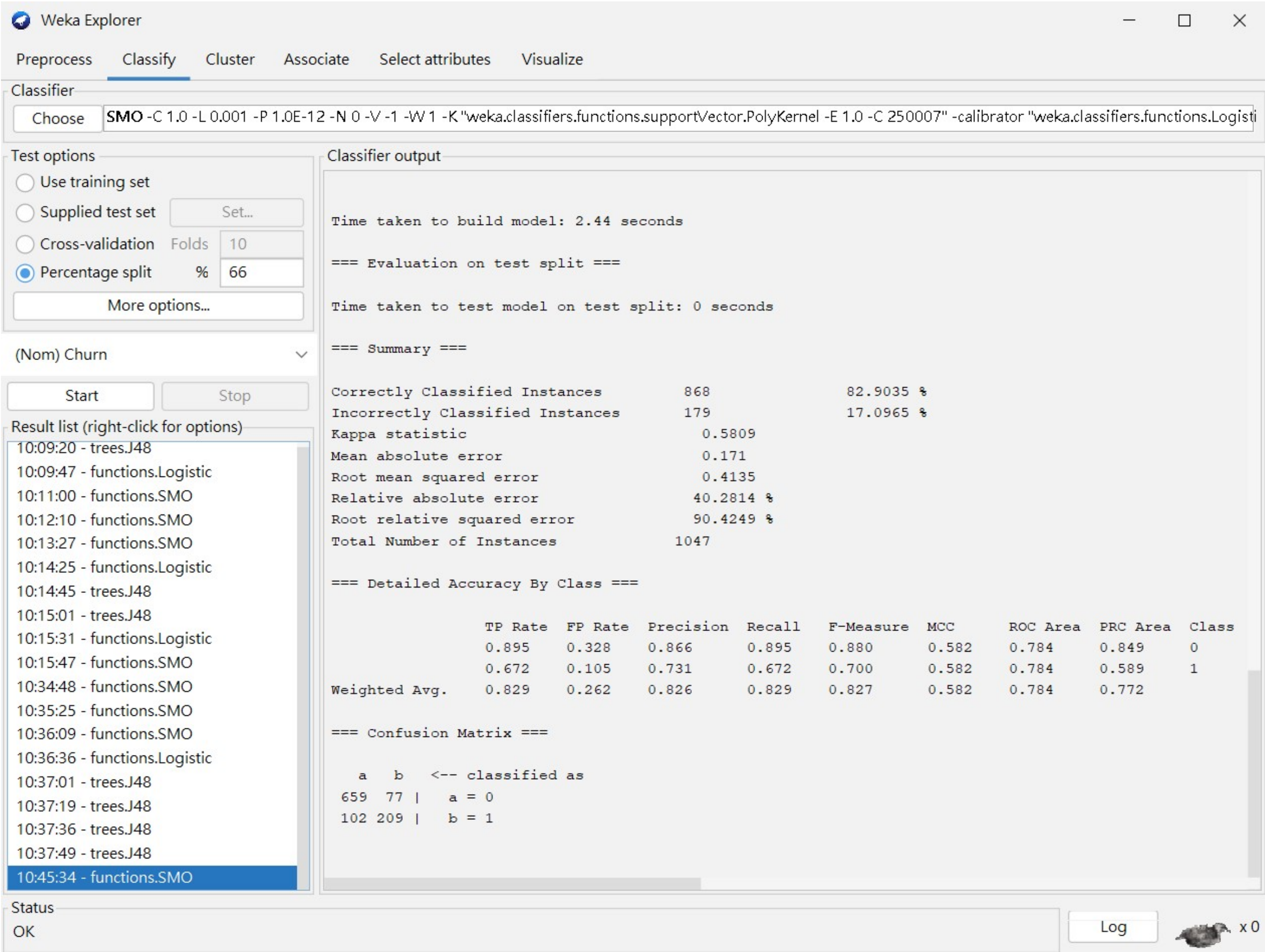


4.訓練、測試 SVM、Logistic Regression、Decision Tree 模型，請以 Accuracy 評估 模型表現

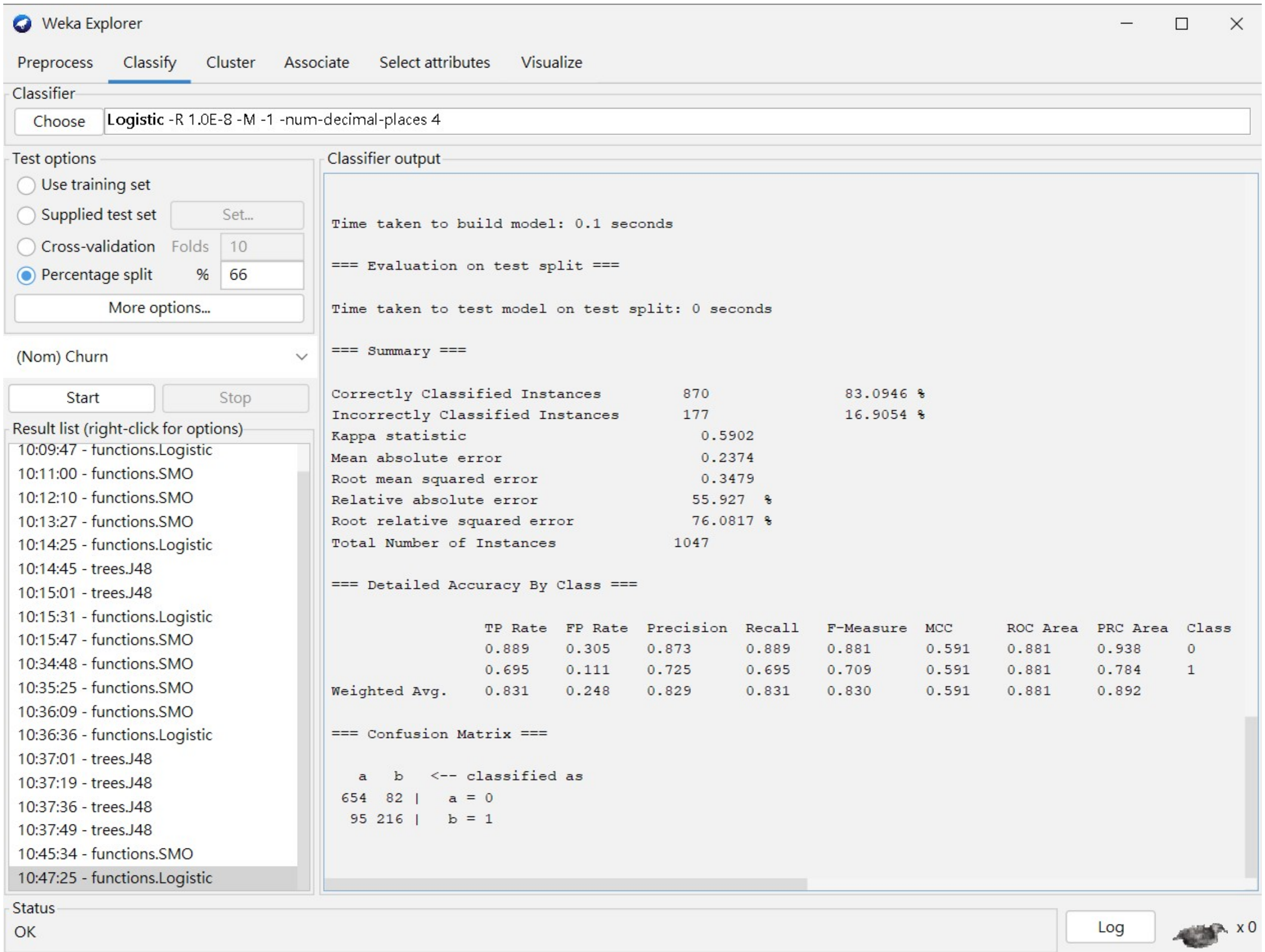
(模型皆為預設未調整，Training Data: 66%)

未用 NominalToBinary 前各模型準確率：

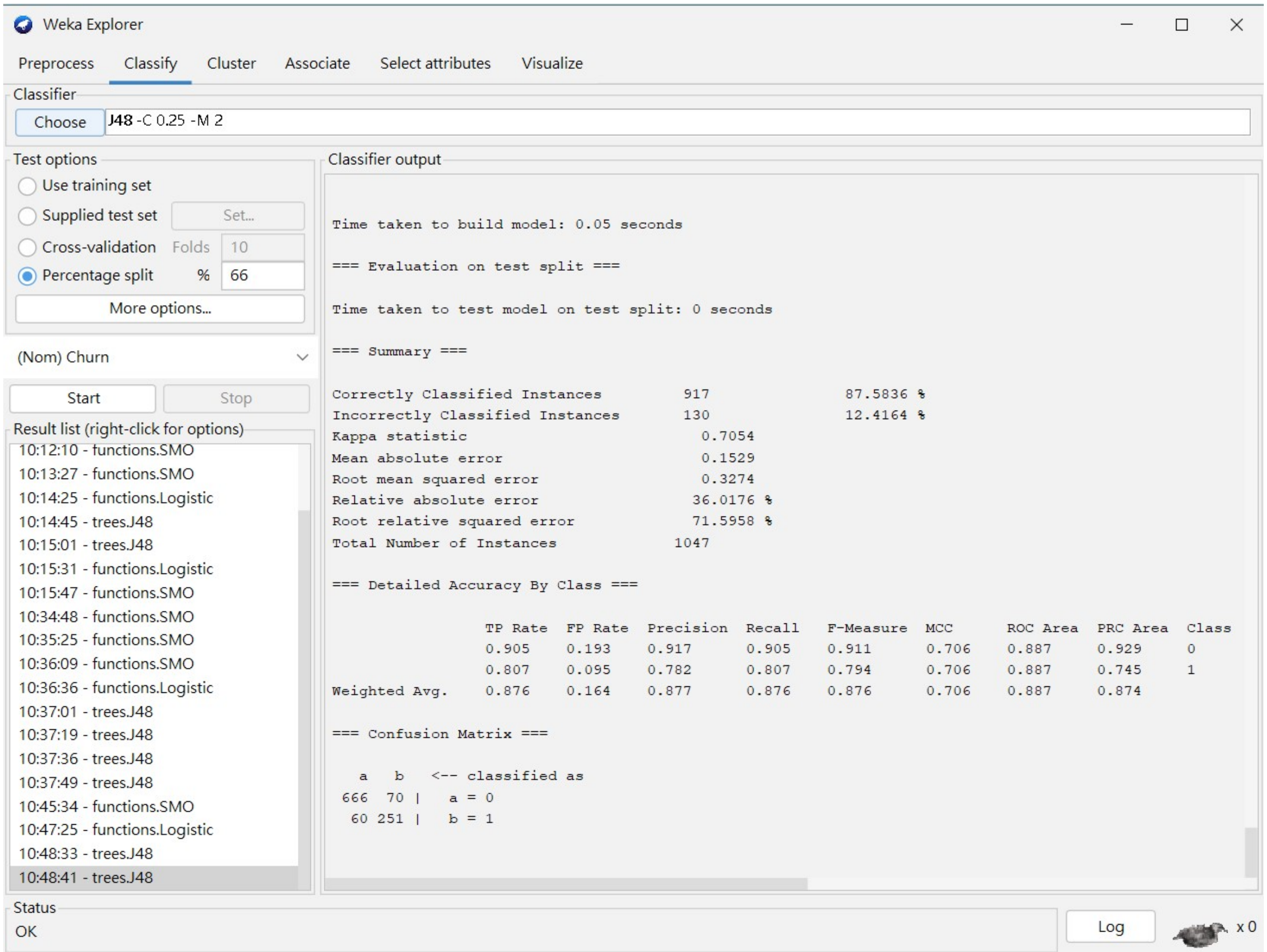
SVM：判斷正確為 82.9035%



Logistic Regression : 判斷正確為 83.0946%

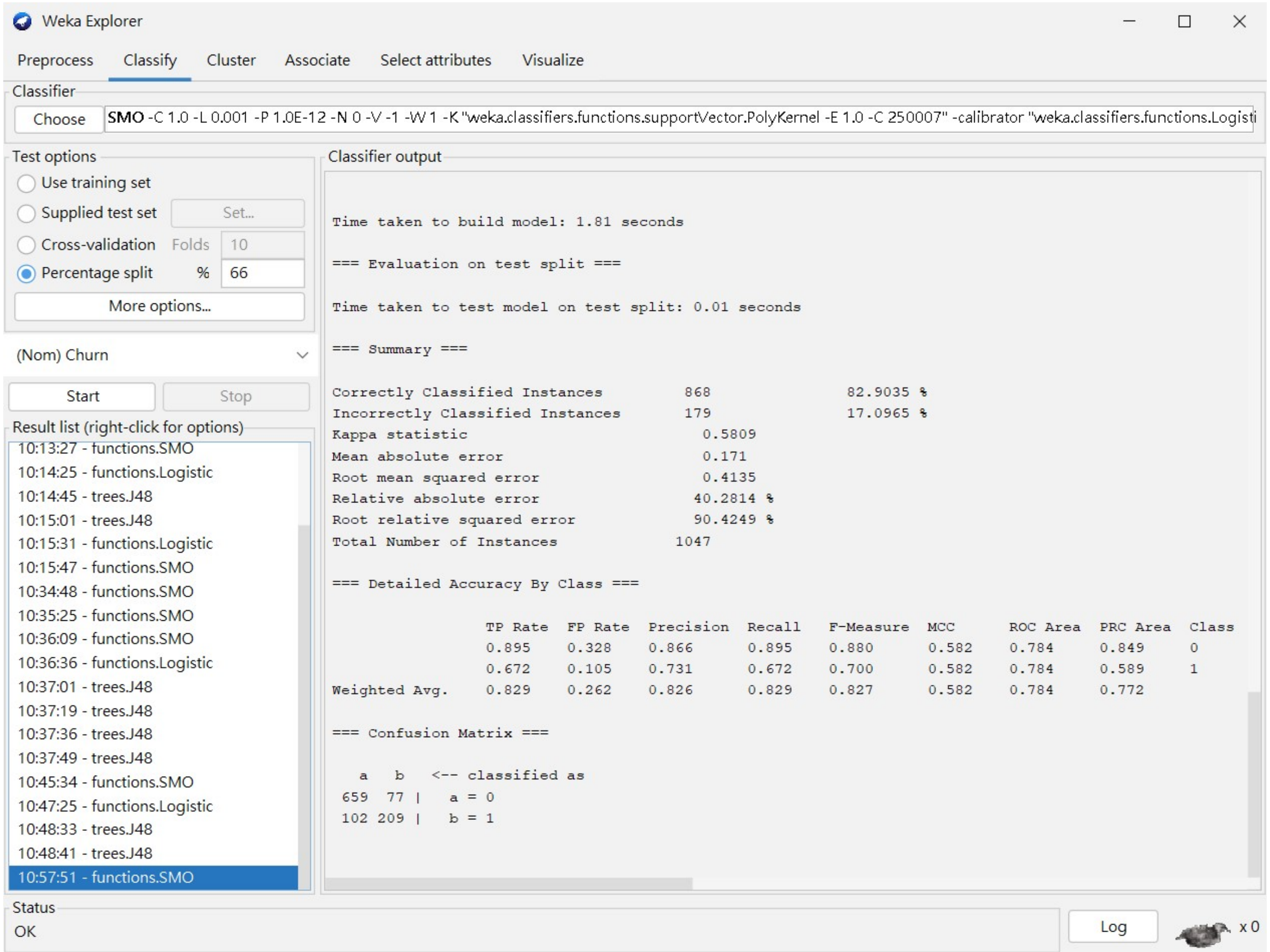


Decision Tree : 判斷正確為 87.5836%

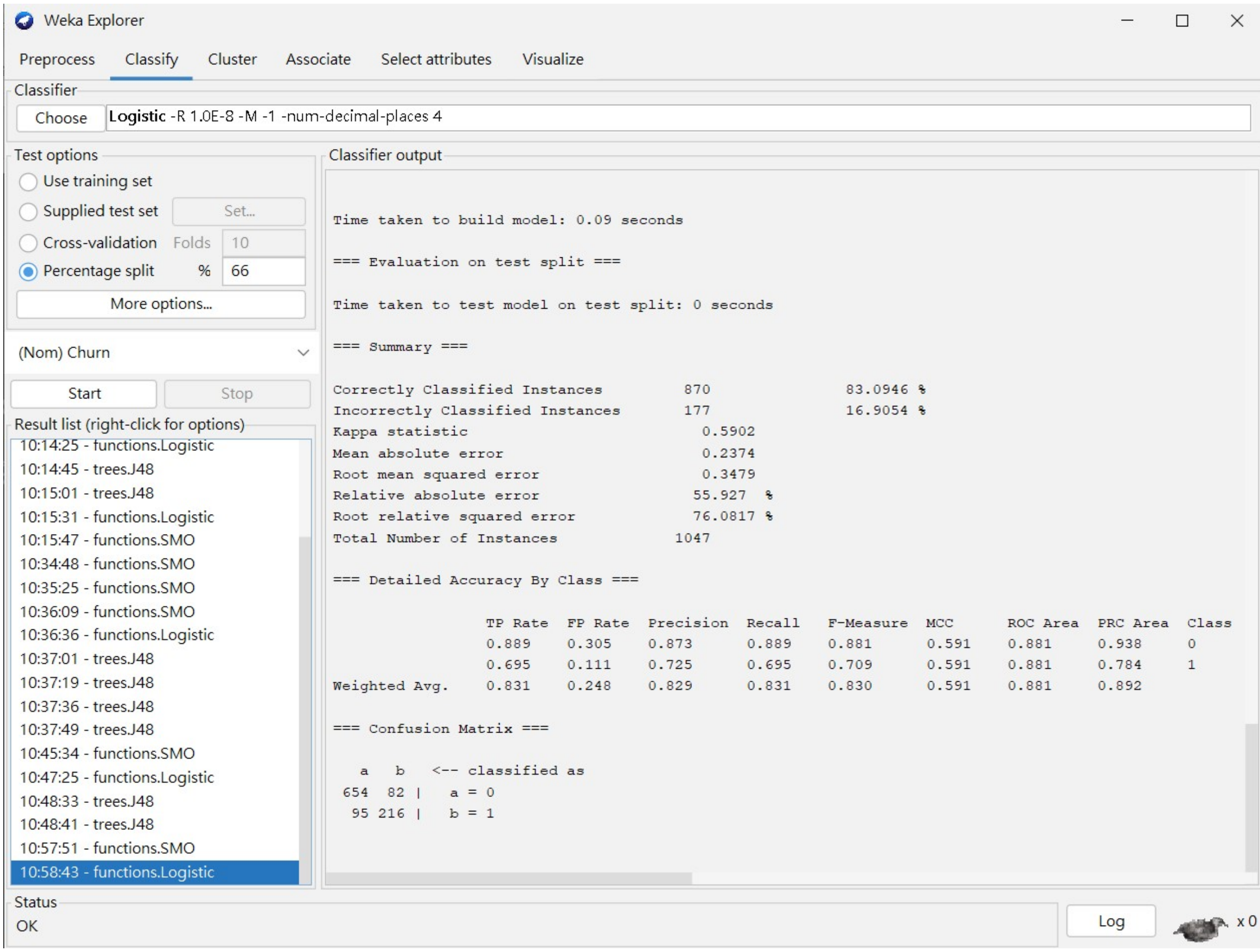


用 NominalToBinary 後各模型準確率：

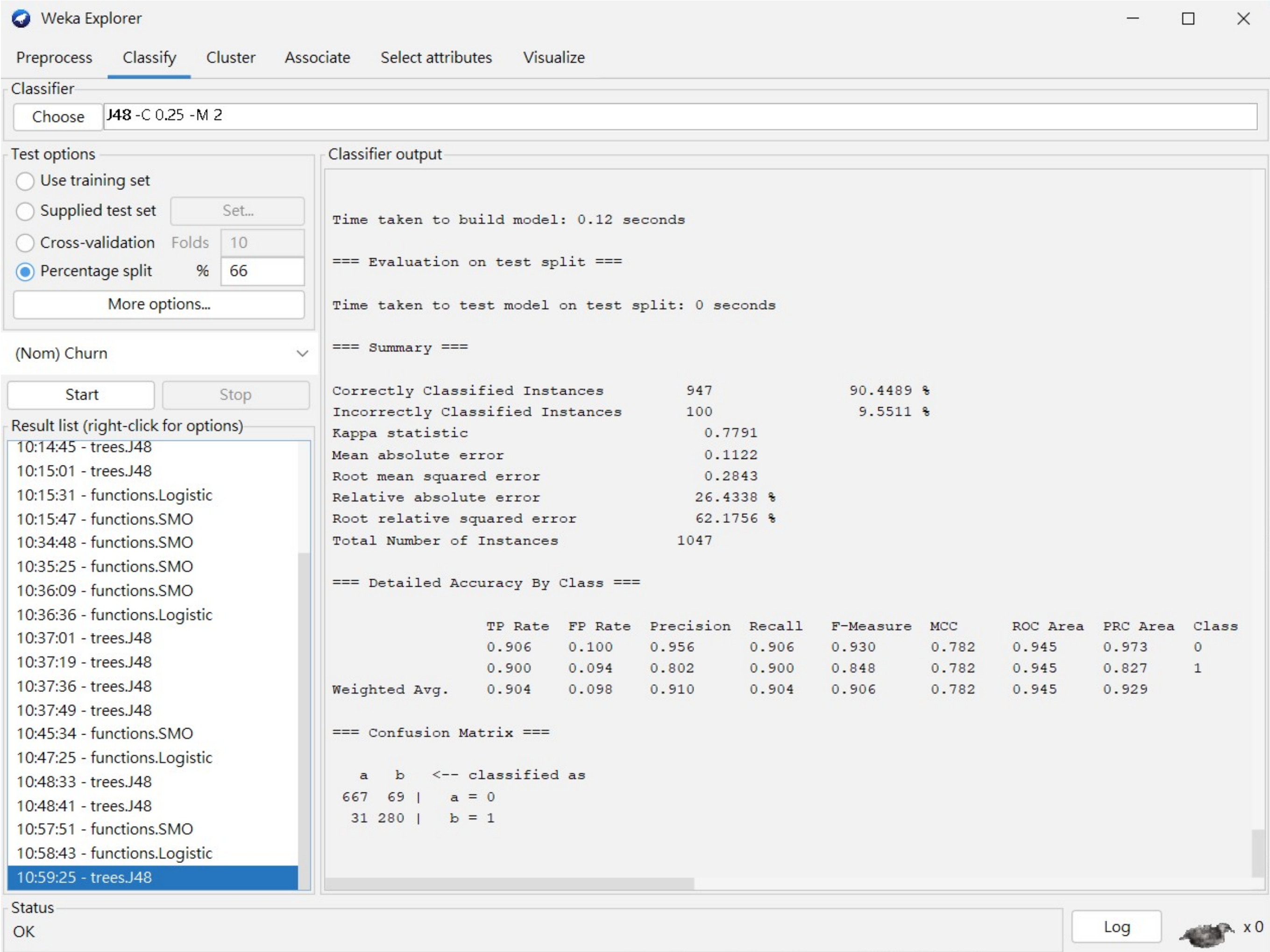
SVM：判斷正確為 82.9035%



Logistic Regression：判斷正確為 83.0946%



Decision Tree : 判斷正確為 90.4489%



結論：用 NominalToBinary 的前後，SVM 與 Logistic Regression 的準確率沒有影響，但 Decision Tree 的準確率有微幅上升。