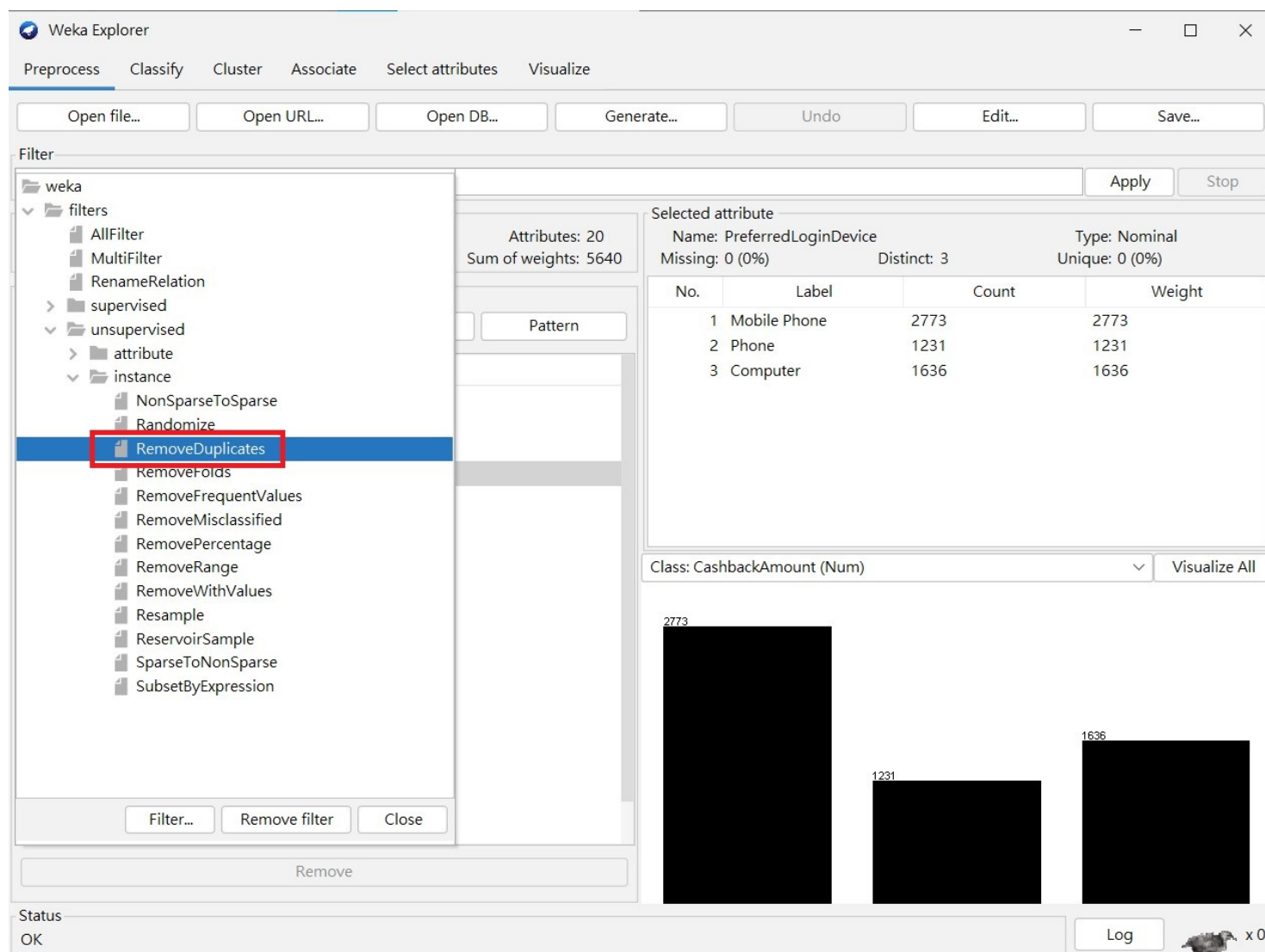


# ECT\_HW5\_108403523

## 1. 使用 **Stratified sampling** 從原本的資料集中取**60%**的資料

(因為 Weka 的 stratified resample 要先做前處理才能執行，所以我把前處理的步驟截圖放在第1題，第3題就不再做前處理，直接以處理後的檔案完成後續題目要求。)

### 1.1 刪除重複 **CustomerID** 的資料



The screenshot shows the Weka Explorer interface. The 'Filter' list on the left has 'RemoveDuplicates' selected. The 'Selected attribute' panel on the right shows 'PreferredLoginDevice' with 3 distinct values. A bar chart visualizes the counts for each value.

No.	Label	Count	Weight
1	Mobile Phone	2773	2773
2	Phone	1231	1231
3	Computer	1636	1636

Class: CashbackAmount (Num) Visualize All

### 1.2 刪除 **CustomerID**

Weka Explorer

Preprocess   Classify   Cluster   Associate   Select attributes   Visualize

Open file...   Open URL...   Open DB...   Generate...   Undo   Edit...   Save...

Filter: Choose **None**   Apply   Stop

Current relation  
Relation: customer\_churn   Attributes: 20   Sum of weights: 5640  
Instances: 5640

Attributes  
All   None   Invert   Pattern

No.	Name
1	<input checked="" type="checkbox"/> i>CustomerID
2	<input type="checkbox"/> Churn
3	<input type="checkbox"/> Tenure
4	<input type="checkbox"/> PreferredLoginDevice
5	<input type="checkbox"/> CityTier
6	<input type="checkbox"/> WarehouseToHome
7	<input type="checkbox"/> PreferredPaymentMode
8	<input type="checkbox"/> Gender
9	<input type="checkbox"/> HourSpendOnApp
10	<input type="checkbox"/> NumberOfDeviceRegistered
11	<input type="checkbox"/> PreferredOrderCat
12	<input type="checkbox"/> SatisfactionScore
13	<input type="checkbox"/> MaritalStatus
14	<input type="checkbox"/> NumberOfAddress
15	<input type="checkbox"/> Complain
16	<input type="checkbox"/> OrderAmountHikeFromlastYear
17	<input type="checkbox"/> CouponUsed
18	<input type="checkbox"/> OrderCount

Remove

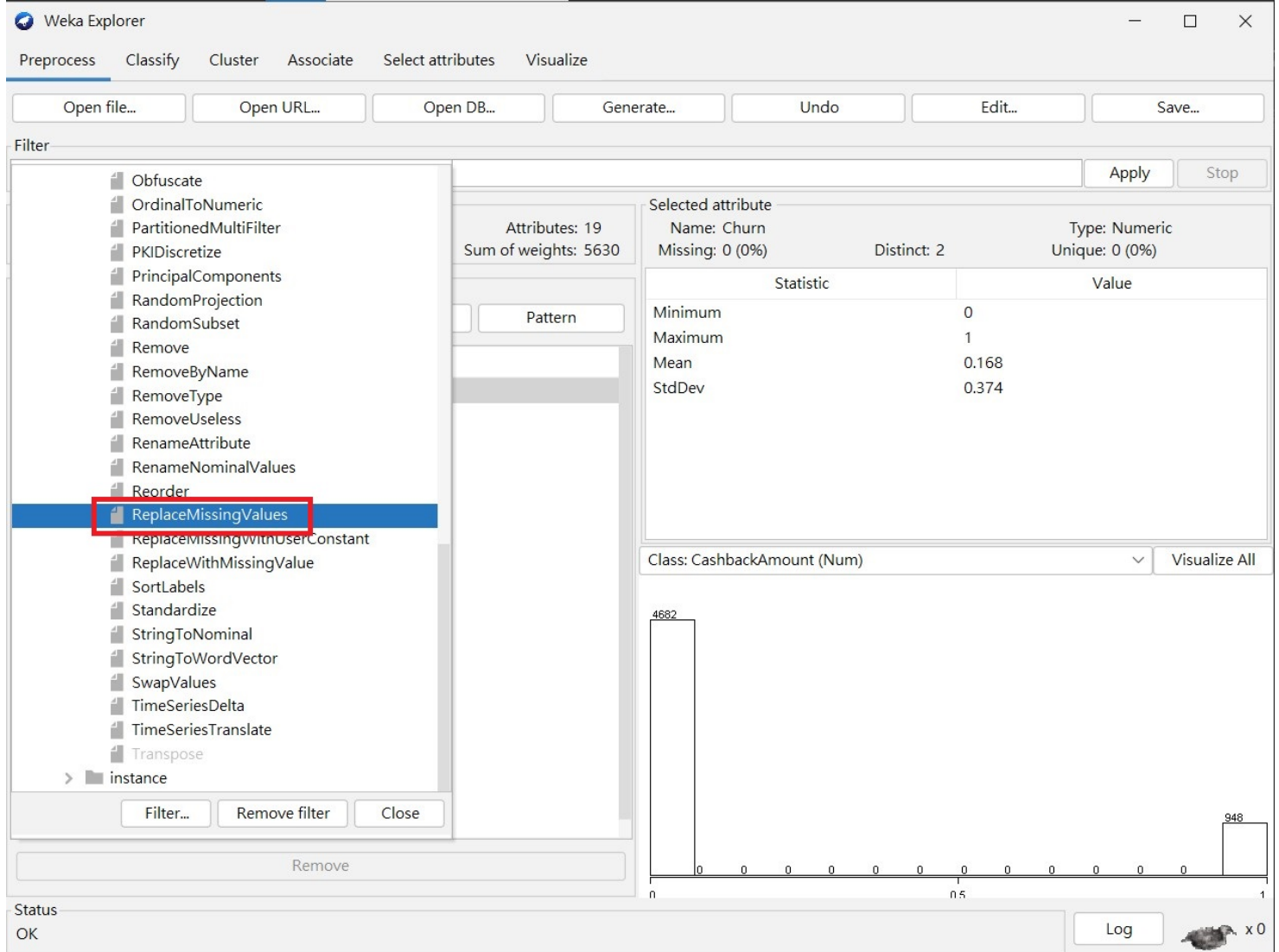
Selected attribute  
Name: i>CustomerID   Type: Numeric  
Missing: 0 (0%)   Distinct: 5630   Unique: 5624 (100%)

Statistic	Value
Minimum	50001
Maximum	55630
Mean	52812.458
StdDev	1627.394

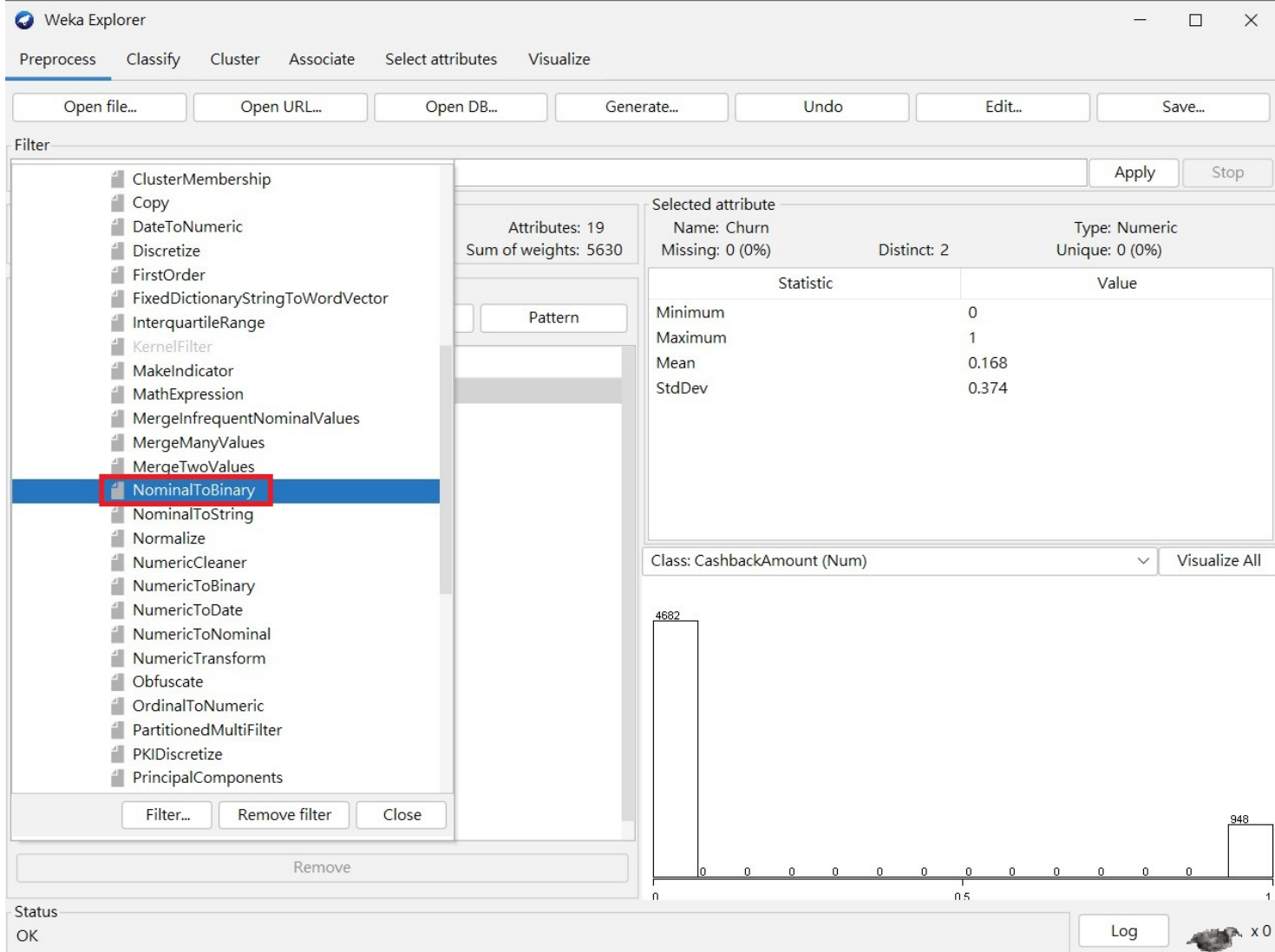
Class: CashbackAmount (Num)   Visualize All

Status: OK   Log   x 0

## 1.3 填补缺失值



## 1.4 將資料進行轉換，僅保留 Churn 為唯一的 "Nominal" 欄



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **NominalToBinary -R 3,6,7,10,12** Apply Stop

Current relation: Relation: customer\_churn-weka.filters.unsupervised.attribute.NominalToBinary Instances: 5630

Attributes: All None

No.	
1	<input type="checkbox"/> Churn
2	<input type="checkbox"/> Tenure
3	<input type="checkbox"/> PreferredLoginDevice
4	<input type="checkbox"/> CityTier
5	<input type="checkbox"/> WarehouseToHome
6	<input type="checkbox"/> PreferredPaymentMode
7	<input type="checkbox"/> Gender
8	<input type="checkbox"/> HourSpendOnApp
9	<input type="checkbox"/> NumberOfDeviceRegistered
10	<input type="checkbox"/> PreferredOrderCat
11	<input type="checkbox"/> SatisfactionScore
12	<input type="checkbox"/> MaritalStatus
13	<input type="checkbox"/> NumberOfAddress
14	<input type="checkbox"/> Complain
15	<input type="checkbox"/> OrderAmountHikeFromlastYear
16	<input type="checkbox"/> CouponUsed
17	<input type="checkbox"/> OrderCount
18	<input type="checkbox"/> DaySinceLastOrder

Remove

Status: OK

Log x 0

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.NominalToBinary

About: Converts all nominal attributes into binary numeric attributes. More Capabilities

attributeIndices: 3,6,7,10,12

binaryAttributesNominal: False

debug: False

doNotCheckCapabilities: False

invertSelection: False

spreadAttributeWeight: False

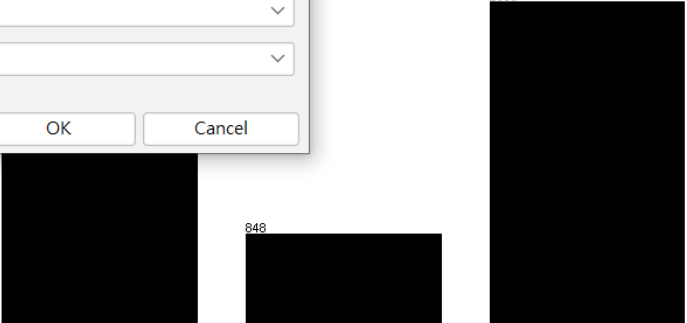
transformAllValues: False

Open... Save... OK Cancel

Type: Nominal Unique: 0 (0%)

Count	Weight
96	1796
8	848
86	2986

Visualize All



Apply Stop

- Filter... Remove filter Close

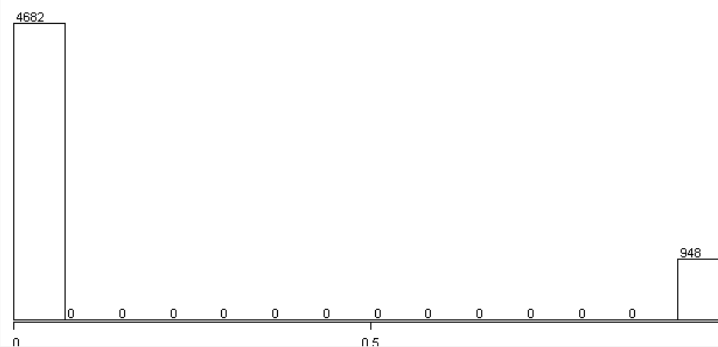
### Pattern

Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.168
StdDev	0.374

Class: Churn (Num)

Visualize All



Log

X



Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **NumericToNominal -R last** Apply Stop

Current relation: Relation: customer\_churn-weka.filters.unsupervised.insta... Attributes: 34 Instances: 5630 Sum of weights: 5630

Attributes: All None Invert Pattern

No. Name

- 17 ☐ PreferredOrderCat=Laptop & Accessory
- 18 ☐ NumberOfDeviceRegistered
- 19 ☐ HourSpendOnApp
- 20 ☐ Gender=Male
- 21 ☐ PreferredPaymentMode=Credit Card
- 22 ☐ PreferredPaymentMode=COD
- 23 ☐ PreferredPaymentMode=E wallet
- 24 ☐ PreferredPaymentMode=Cash on Delivery
- 25 ☐ PreferredPaymentMode=CC
- 26 ☐ PreferredPaymentMode=UPI
- 27 ☐ PreferredPaymentMode=Debit Card
- 28 ☐ WarehouseToHome
- 29 ☐ CityTier
- 30 ☐ PreferredLoginDevice=Computer
- 31 ☐ PreferredLoginDevice=Phone
- 32 ☐ PreferredLoginDevice=Mobile Phone
- 33 ☐ Tenure
- 34 ☒ Churn**

Remove

Selected attribute: **Name: Churn** Type: Nominal  
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	4682	4682
2	1	948	948

Class: Churn (Nom) Visualize All

Status: OK Log x 0

## 1.5 重新排列屬性(顛倒排列，Churn 排最後)

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

- Obfuscate
- OrdinalToNumeric
- PartitionedMultiFilter
- PKIDiscretize
- PrincipalComponents
- RandomProjection
- RandomSubset
- Remove
- RemoveByName
- RemoveType
- RemoveUseless
- RenameAttribute
- RenameNominalValues
- Reorder**
- ReplaceMissingValues
- ReplaceMissingWithUserConstant
- ReplaceWithMissingValue
- SortLabels
- Standardize
- StringToNominal
- StringToWordVector
- SwapValues
- TimeSeriesDelta
- TimeSeriesTranslate
- Transpose

> instance

Filter... Remove filter Close

Attributes: 34  
Sum of weights: 5630

Pattern

Selected attribute

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

Statistic	Value
Minimum	0
Maximum	1
Mean	0.168
StdDev	0.374

Class: CashbackAmount (Num) Visualize All

4682 948

0 0.5 1

Status  
OK

Log x 0



Weka Explorer

Preprocess   Classify   Cluster   Associate   Select attributes   Visualize

Open file...   Open URL...   Open DB...   Generate...   Undo   Edit...   Save...

Filter: Choose **Reorder -R last-first**   Apply   Stop

Current relation  
Relation: customer\_churn-weka.filters.unsupervised.insta...   Attributes: 34   Sum of weights: 5630  
Instances: 5630

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

Attributes

All   None

No.	
1	<input checked="" type="checkbox"/> Churn
2	<input type="checkbox"/> Tenure
3	<input type="checkbox"/> PreferredLoginDevice=Mobile Phone
4	<input type="checkbox"/> PreferredLoginDevice=Phone
5	<input type="checkbox"/> PreferredLoginDevice=Computer
6	<input type="checkbox"/> CityTier
7	<input type="checkbox"/> WarehouseToHome
8	<input type="checkbox"/> PreferredPaymentMode=Debit Card
9	<input type="checkbox"/> PreferredPaymentMode=UPI
10	<input type="checkbox"/> PreferredPaymentMode=CC
11	<input type="checkbox"/> PreferredPaymentMode=Cash on delivery
12	<input type="checkbox"/> PreferredPaymentMode=E wallet
13	<input type="checkbox"/> PreferredPaymentMode=COD
14	<input type="checkbox"/> PreferredPaymentMode=Credit Card
15	<input type="checkbox"/> Gender=Male
16	<input type="checkbox"/> HourSpendOnApp
17	<input type="checkbox"/> NumberOfDeviceRegistered
18	<input type="checkbox"/> PreferredOrderCat=Laptop & Accessory

Remove

Status: OK

Log   x 0

Statistic

Statistic	Value
	0
	1
	0.168
	0.374

Visualize All

948

0   0.5   1

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.Reorder

About

A filter that generates output with a new order of the attributes.

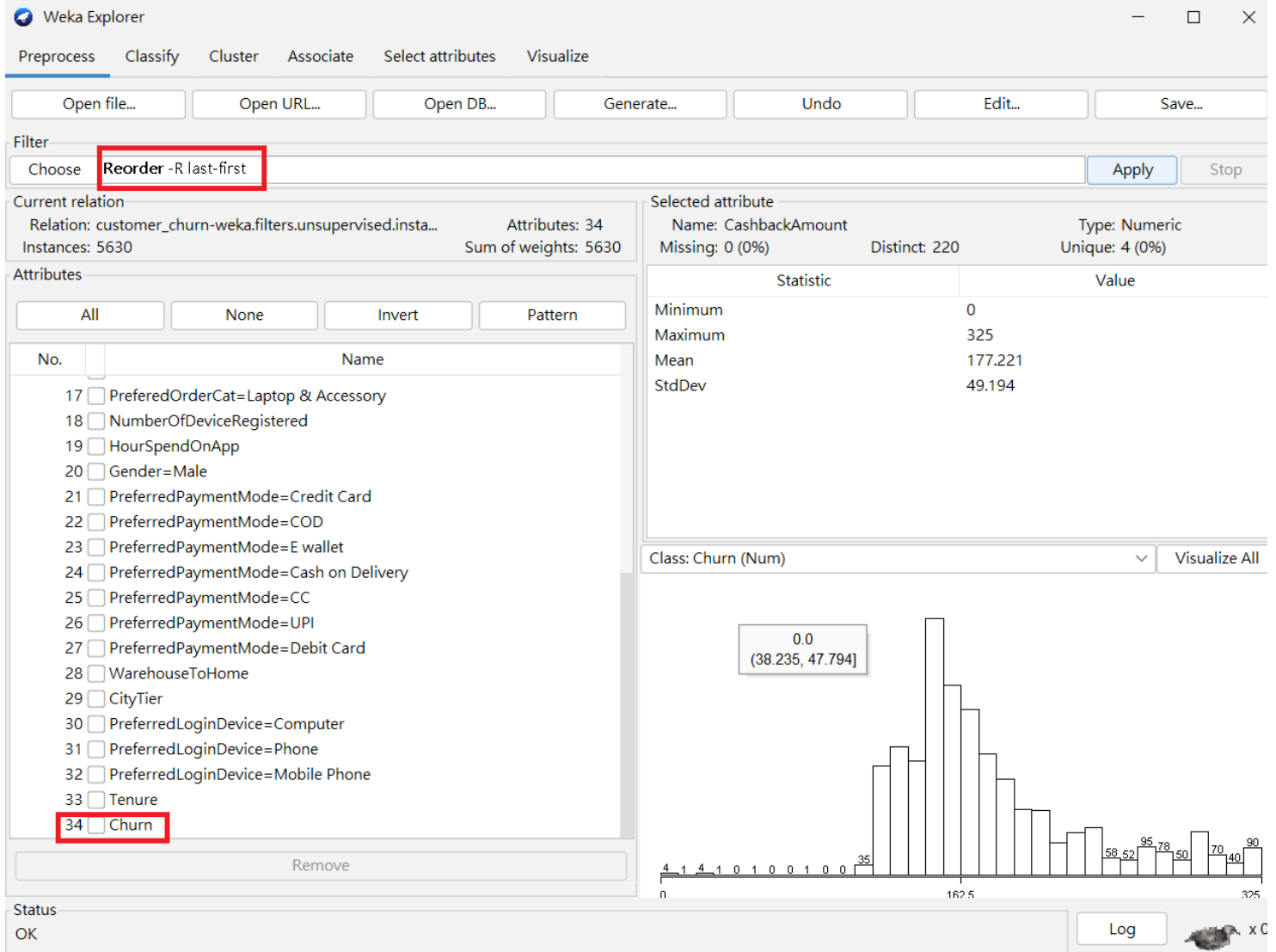
More   Capabilities

attributeIndices   last-first

debug   False

doNotCheckCapabilities   False

Open...   Save...   OK   Cancel



## 1.6 做 60% stratified resample

Weka Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Open file...

Open URL...

Open DB...

Generate...

Undo

Edit...

Save...

Filter

weka

filters

AllFilter

MultiFilter

RenameRelation

supervised

attribute

instance

ClassBalancer

Resample

SpreadSubsample

StratifiedRemoveFolds

unsupervised

Attributes: 34

Sum of weights: 5630

Pattern

Selected attribute

Name: Churn

Missing: 0 (0%)

Distinct: 2

Type: Nominal

Unique: 0 (0%)

No.	Label	Count	Weight
1	0	4682	4682
2	1	948	948

Class: Churn (Nom)

Visualize All

4682

948

Filter...

Remove filter


Close

Remove

Status

OK

Log

 x 0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **Resample -B 0.0 -S 1 -Z 60.0** Apply Stop

Current relation: Relation: customer\_churn-weka.filters.unsupervised.instance.Resample Instances: 5630

Attributes: All None

No. 17 PreferredOrderCat=Laptop & Access 18 NumberOfDeviceRegistered 19 HourSpendOnApp 20 Gender=Male 21 PreferredPaymentMode=Credit Card 22 PreferredPaymentMode=COD 23 PreferredPaymentMode=E wallet 24 PreferredPaymentMode=Cash on delivery 25 PreferredPaymentMode=CC 26 PreferredPaymentMode=UPI 27 PreferredPaymentMode=Debit Card 28 WarehouseToHome 29 CityTier 30 PreferredLoginDevice=Computer 31 PreferredLoginDevice=Phone 32 PreferredLoginDevice=Mobile Phone 33 Tenure 34 Churn

Remove

Status OK Log x 0

weka.gui.GenericObjectEditor  
weka.filters.supervised.instance.Resample

About  
Produces a random subsample of a dataset using either sampling with replacement or without replacement. More Capabilities

biasToUniformClass 0.0  
debug False  
doNotCheckCapabilities False  
invertSelection False  
noReplacement False  
randomSeed 1  
**sampleSizePercent 60.0**

Open... Save... OK Cancel

Type: Nominal  
Unique: 0 (0%)

Count	Weight
82	4682
8	948

Visualize All

948

## 2. 顯示取樣後各類別的資料數量

### 2.1 紅框內為分層抽樣後各類別資料數量

Weka Explorer

Preprocess   Classify   Cluster   Associate   Select attributes   Visualize

Open file...   Open URL...   Open DB...   Generate...   Undo   Edit...   Save...

Filter  
Choose **Resample -B 0.0 -S 1 -Z 60.0**   Apply   Stop

Current relation  
Relation: customer\_churn-weka.filters.unsupervised.insta...   Attributes: 34   Sum of weights: 3377  
Instances: 3377

Attributes  
All   None   Invert   Pattern

No.	Name
17	<input type="checkbox"/> PreferredOrderCat=Laptop & Accessory
18	<input type="checkbox"/> NumberOfDeviceRegistered
19	<input type="checkbox"/> HourSpendOnApp
20	<input type="checkbox"/> Gender=Male
21	<input type="checkbox"/> PreferredPaymentMode=Credit Card
22	<input type="checkbox"/> PreferredPaymentMode=COD
23	<input type="checkbox"/> PreferredPaymentMode=E wallet
24	<input type="checkbox"/> PreferredPaymentMode=Cash on Delivery
25	<input type="checkbox"/> PreferredPaymentMode=CC
26	<input type="checkbox"/> PreferredPaymentMode=UPI
27	<input type="checkbox"/> PreferredPaymentMode=Debit Card
28	<input type="checkbox"/> WarehouseToHome
29	<input type="checkbox"/> CityTier
30	<input type="checkbox"/> PreferredLoginDevice=Computer
31	<input type="checkbox"/> PreferredLoginDevice=Phone
32	<input type="checkbox"/> PreferredLoginDevice=Mobile Phone
33	<input type="checkbox"/> Tenure
34	<input checked="" type="checkbox"/> Churn

Remove

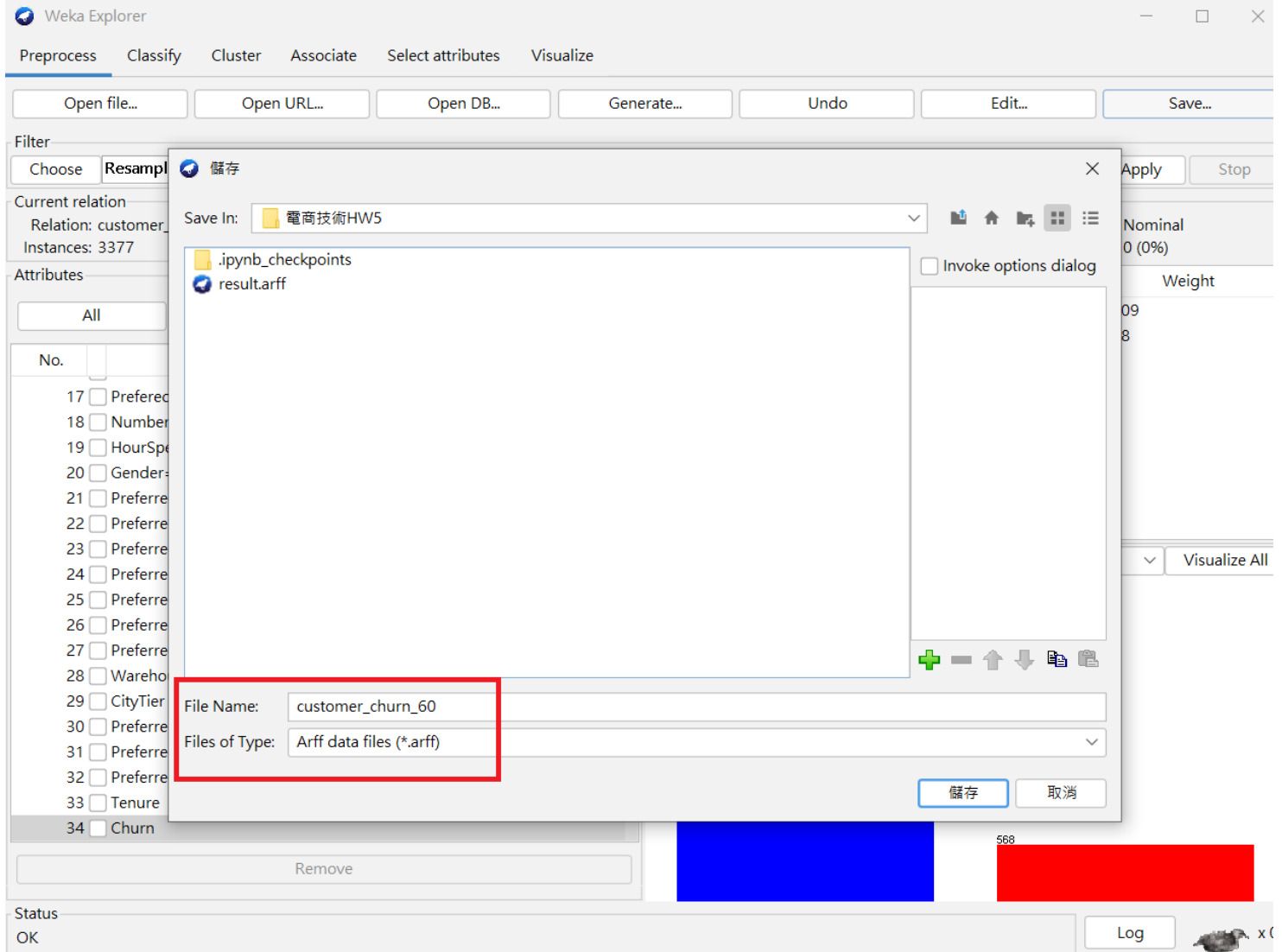
Selected attribute  
Name: **Churn**   Type: Nominal  
Missing: 0 (0%)   Distinct: 2   Unique: 0 (0%)

No.	Label	Count	Weight
1	0	2809	2809
2	1	568	568

Class: Churn (Nom)   Visualize All

Status  
OK   Log   x 0

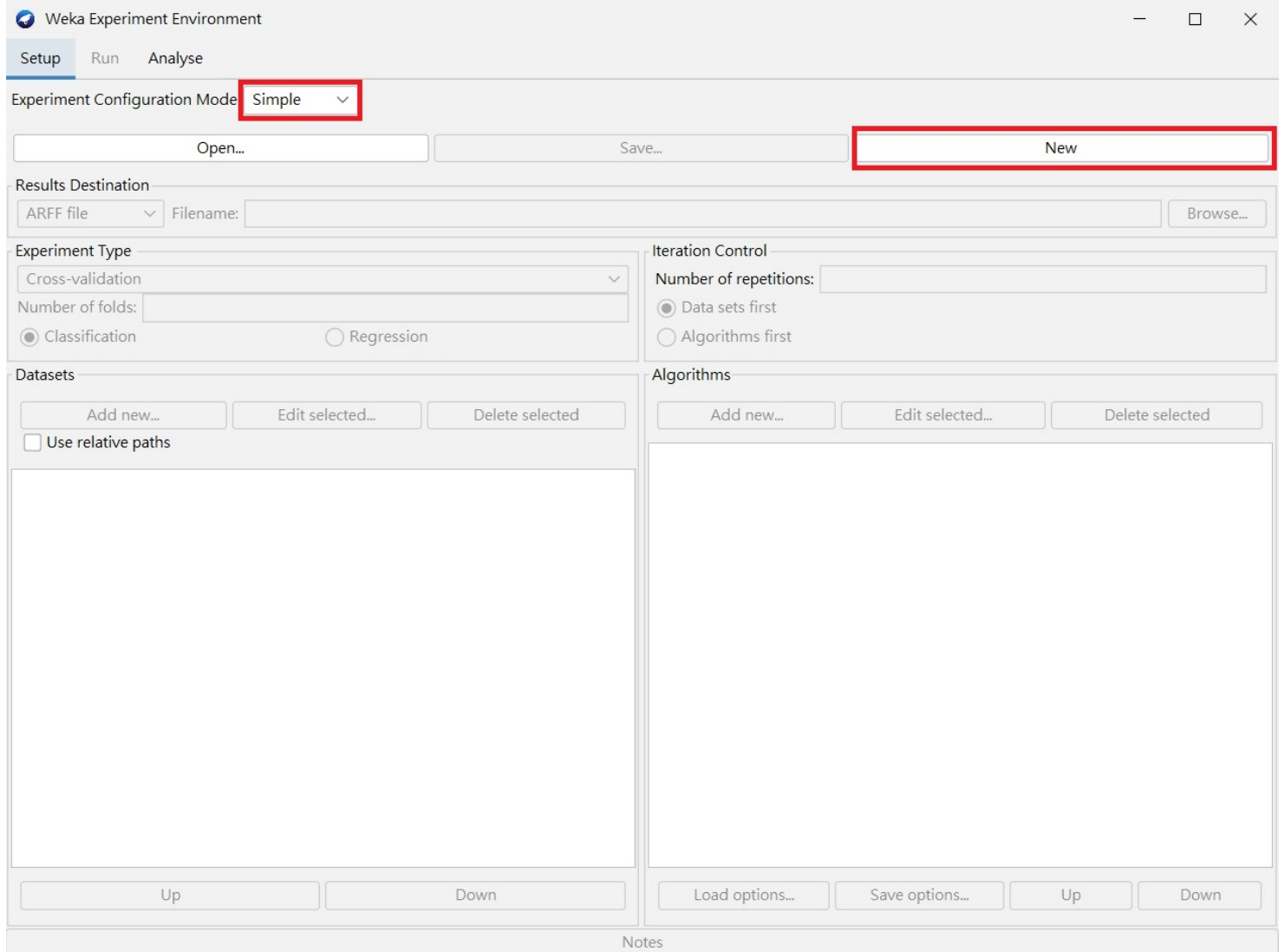
## 2.2 存成新的檔案做後續第3、4題



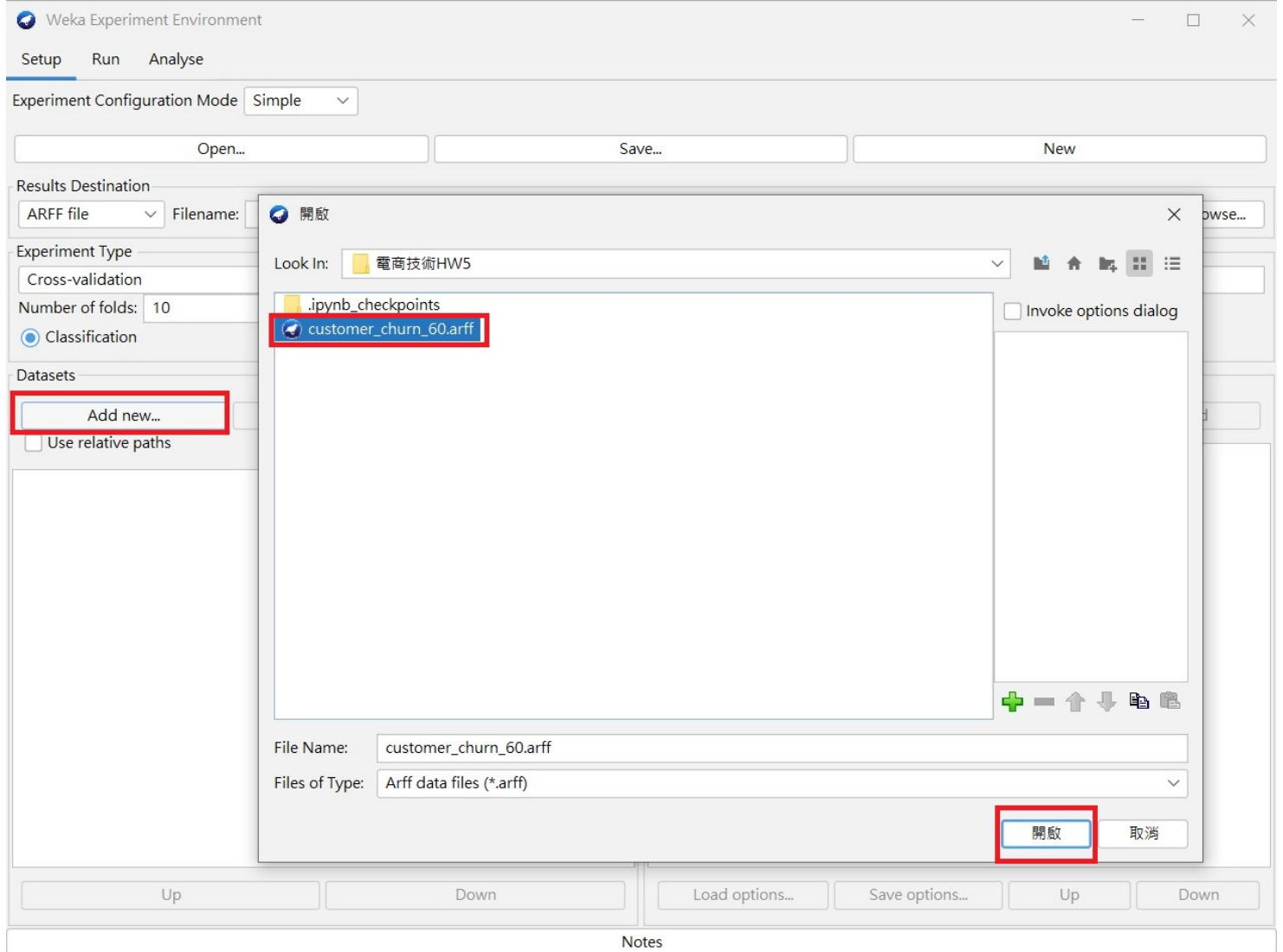
### 3. 資料前處理，並以 **repeated 10 folds cross-validation (重複 10 次) Paired t-test** 比較 **Logistic Regression** 及 **SVM** 模型

(在第1、2題已完成資料前處理，並將分層抽樣結果另存新檔)

#### 3.1 New Experiment

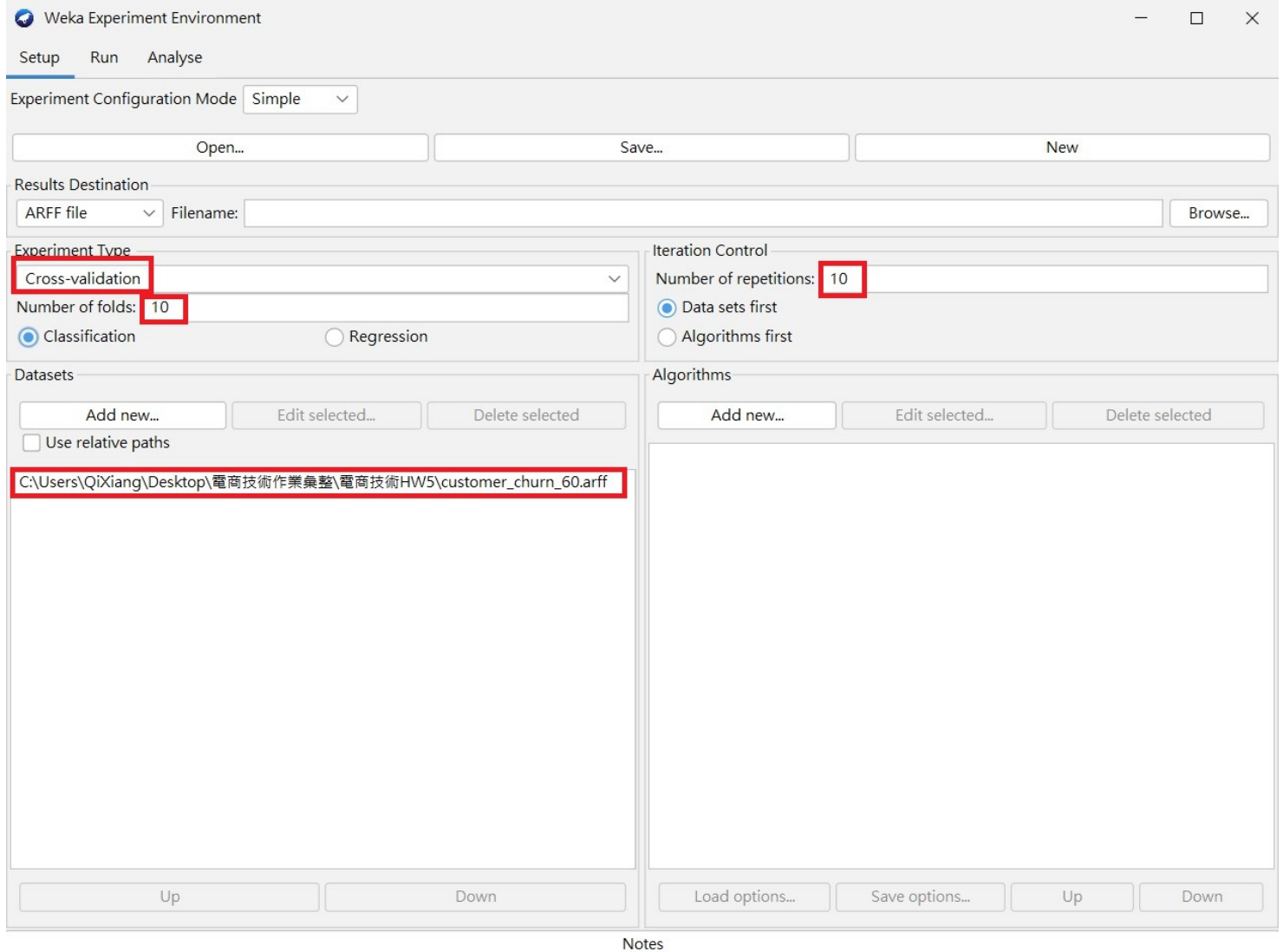


## 3.2 導入資料



### 3.3 設定 10 run 10 folds cross validation





### 3.4 設定要跑的兩個模型

(Logistic Regression)

Weka Experiment Environment

SetupRunAnalyse

Experiment Configuration ModeSimple

Open...Save...New

Results DestinationARFF fileFilename:Browse...

Experiment TypeCross-validation

Number of folds:10

Classification

DatasetsAdd new...Edit se

Use relative paths

C:\Users\QiXiang\Desktop\電商技術作業集

UpDownLoad options...Save options...UpDown

Notes

weka.gui.GenericObjectEditor

Chooseweka.classifiers.functions.Logistic

About

Class for building and using a multinomial logistic regression model with a ridge estimator.

MoreCapabilities

batchSize100

debugFalse

doNotCheckCapabilitiesFalse

doNotStandardizeAttributesFalse

maxIts-1

numDecimalPlaces4

ridge1.0E-8

useConjugateGradientDescentFalse

Open...Save...OKCancel

# (SVM)

Weka Experiment Environment

SetupRunAnalyse

Experiment Configuration ModeSimple

Open...

Results Destination  
ARFF file ▾Filename:

Experiment Type  
Cross-validation

Number of folds: 10  
☒ Classification

Datasets  
Add new... Edit s  
☐ Use relative paths  
C:\Users\QiXiang\Desktop\電商技術作業集

UpDownLoad options...Save options...UpDown

weka.gui.GenericObjectEditor

Chooseweka.classifiers.functions.SGD

About  
Implements stochastic gradient descent for learning various linear models (binary class SVM, binary class logistic regression, squared loss, Huber loss and epsilon-insensitive loss linear regression).

MoreCapabilities

batchSize100

debugFalse ▾

doNotCheckCapabilitiesFalse ▾

dontNormalizeFalse ▾

dontReplaceMissingFalse ▾

epochs500

epsilon0.001

lambda1.0E-4

learningRate0.01

lossFunctionHinge loss (SVM) ▾

numDecimalPlaces2

seed1

Open...Save...OKCancel

New

Browse...

Delete selected

Notes

(各跑10次)

Weka Experiment Environment

Setup Run Analyse

Experiment Configuration Mode Simple

Open... Save... New

Results Destination  
ARFF file Filename: Browse...

Experiment Type  
Cross-validation  
Number of folds: 10  
☒ Classification ☐ Regression

Datasets  
Add new... Edit selected... Delete selected  
☐ Use relative paths  
C:\Users\QiXiang\Desktop\電商技術作業彙整\電商技術HW5\customer\_churn\_60.arff

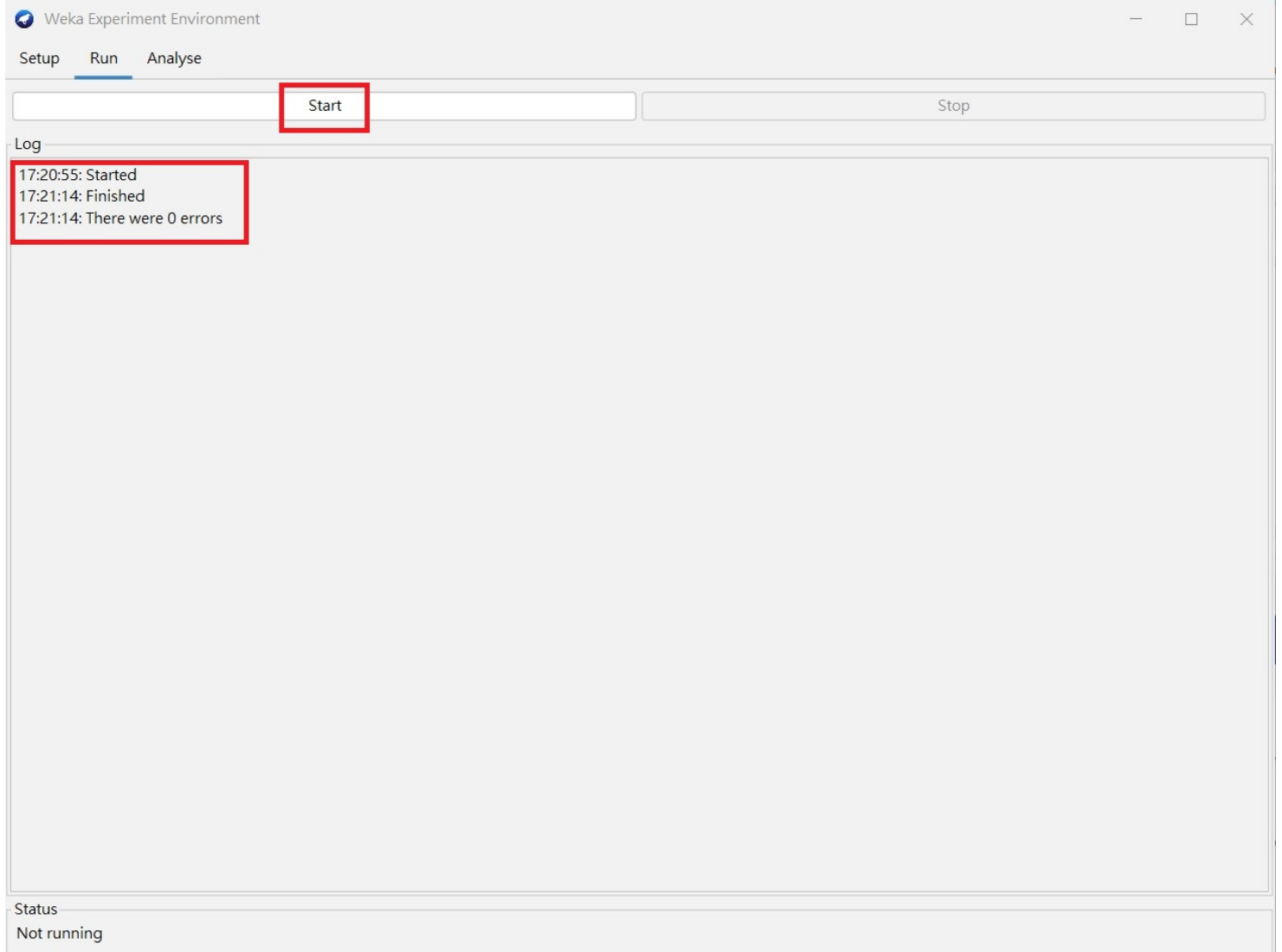
Iteration Control  
Number of repetitions: 10  
☐ Data sets first  
☒ Algorithms first

Algorithms  
Add new... Edit selected... Delete selected  
Logistic -R 1.0E-8 -M -1 -num-decimal-places 4  
SGD -F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1

Up Down Load options... Save options... Up Down

Notes

### 3.5 開始跑 Experiment



## 3.6 分析

# (點擊Experiment)

Weka Experiment Environment

Setup

Run

Analyse

Source

No source

File...Database...Experiment

Actions

Perform testSave outputOpen Explorer...

Configure test

Testing with

Paired T-Tester (corrected)

Select rows and cols

Rows

Cols

Swap

Comparison field

Significance

0.05

Sorting (asc.) by

Test base

Select

Displayed Columns

Select

Show std. deviations

Output Format

Select

Test output

Result list

## (做Paired T-test、參數調整)

The screenshot shows the Weka Experiment Environment window. The 'Analyse' tab is selected. The 'Source' section indicates 'Got 200 results'. The 'Actions' section has buttons for 'Perform test', 'Save output', and 'Open Explorer...'. The 'Configure test' section is on the left, and the 'Test output' section is on the right.

**Configure test**

- Testing with: **Paired T-Tester (corrected)**
- Select rows and cols: Rows, Cols, Swap
- Comparison field: **Percent\_correct**
- Significance: **0.05**
- Sorting (asc.) by: <default>
- Test base: Select
- Displayed Columns: Select
- Show std. deviations: ☐
- Output Format: Select

**Test output**

Available resultsets

```
(1) functions.Logistic '-R 1.0E-8 -M -1 -num-decimal-places 4' 3932117032546553727
(2) functions.SGD '-F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1' -3732968666673530290
```

**Result list**

17:22:23 - Available resultsets

## 3.7 輸出結果

Weka Experiment Environment

Setup Run **Analyse**

Source

Got 200 results File... Database... Experiment

Actions

**Perform test** Save output Open Explorer...

Configure test

Testing with: Paired T-Tester (corrected)

Select rows and cols: Rows Cols Swap

Comparison field: Percent\_correct

Significance: 0.05

Sorting (asc.) by: <default>

Test base: Select

Displayed Columns: Select

Show std. deviations: ☐

Output Format: Select

Result list

- 17:22:23 - Available resultsets
- 17:23:02 - Percent\_correct - functions.Logistic '-R 1.0E-8 -M -

Test output

Tester: weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-ma  
Analysing: Percent\_correct  
Datasets: 1  
Resultsets: 2  
Confidence: 0.05 (two tailed)  
Sorted by: -  
Date: 2022/5/10 下午5:23

Dataset	(1) function	(2) funct
'customer_churn-weka.filt(100)	88.36	88.15

(v/ /\*) | (0/1/0)

Key:

- (1) functions.Logistic '-R 1.0E-8 -M -1 -num-decimal-places 4' 3932117032546553727
- (2) functions.SGD '-F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1' -3732968666673530290

## 4. 根據weka 的輸出說明結論

### 4.1 最終結果



## Test output

```
Tester:      weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -result-ma
Analysing:   Percent_correct
Datasets:    1
Resultsets:  2
Confidence:  0.05 (two tailed)
Sorted by:   -
Date:        2022/5/10 下午5:23
```

Dataset	(1) function	(2) funct
'customer_churn-weka.filt(100)	88.36	88.15
	(v/ /*)	(0/1/0)

Key:

```
(1) functions.Logistic '-R 1.0E-8 -M -1 -num-decimal-places 4' 3932117032546553727
(2) functions.SGD '-F 0 -L 0.01 -R 1.0E-4 -E 500 -C 0.001 -S 1' -3732968666673530290
```

做完兩個模型的 Paired T-test 後，可以發現 SVM 和 Logistic Regression 在預測成功比例上的表現並沒有什麼差異，所以只要任選其一即可。