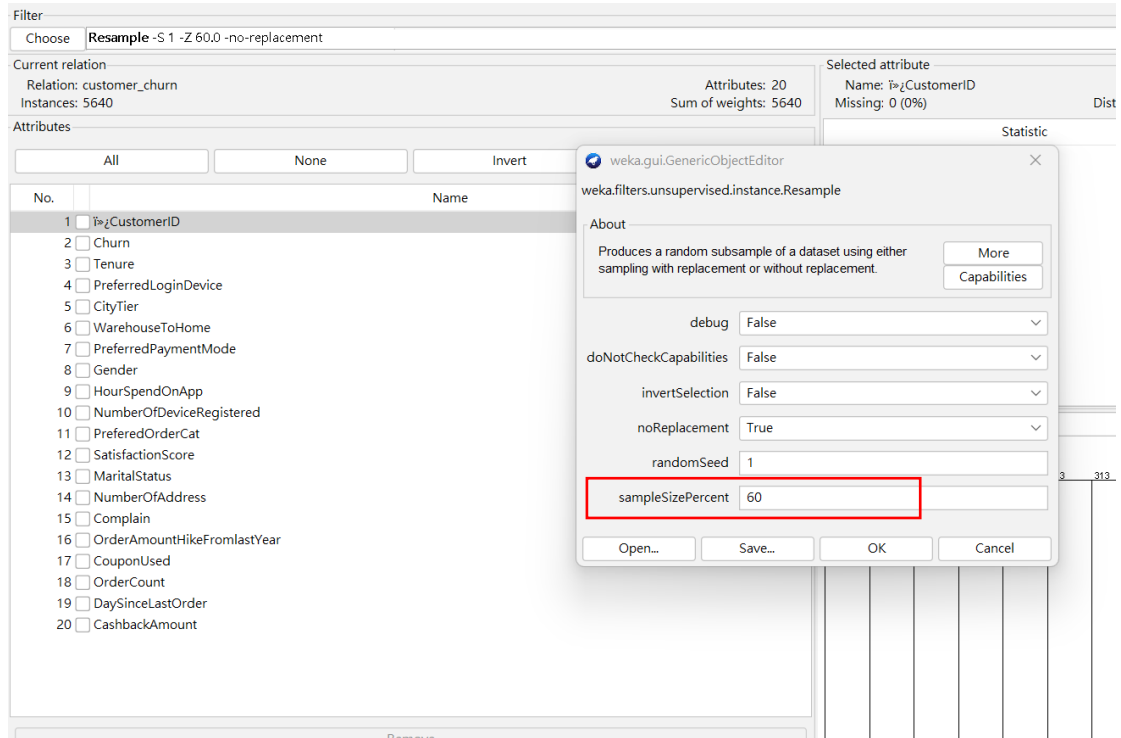


ECT_HW5_109403021_Weka 題

1. 取 60%資料

從 Filter 選取 Resample 並將 sampleSizePercent 調為 60 後 Apply



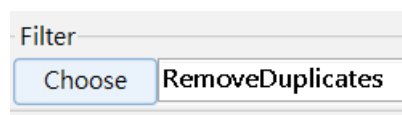
2. 列處 churn 資料個數

用 NumericToNominal 將 Churn 欄為轉為 nominal 後可看到各類別資料個數

| Filter | | | | |
|-------------------------------------|-------|----------------|--------|--|
| Choose NumericToNominal -R 2 | | | | |
| Current relation | | | | |
| Selected attribute | | | | |
| Name: Churn | | Type: Nominal | | |
| Missing: 0 (0%) | | Unique: 0 (0%) | | |
| | | Distinct: 2 | | |
| No. | Label | Count | Weight | |
| 1 | 0 | 2811 | 2811 | |
| 2 | 1 | 573 | 573 | |

3. 資料前處理

(1)先刪除重複多餘的資料(保留一筆)



(2)用平均數填補空值

Filter

Choose

ReplaceMissingValues

Current relation

(3)刪除 CustomerID 欄位

Attributes

All

None

Invert

Pattern

| No. | Name |
|-----|--|
| 1 | <input checked="" type="checkbox"/> 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 |
| 19 | <input type="checkbox"/> DaySinceLastOrder |
| 20 | <input type="checkbox"/> CashbackAmount |

Remove

Status
OK

Remove selected attributes.

4. 用 Experimenter 比較兩模型

我先把 Churn 欄位調到最後讓等等 experimenter 直接識別其為目標變數

Filter

Choose

Reorder - R 2-19;1

Current relation

Relation: customer_churn-weka.filters.unsupervised.instance.Resample-S1-Z60.0-no-replacement-weka.filt...

Instances: 3378

Attributes: 19

Sum of weights: 3378

Attributes

All

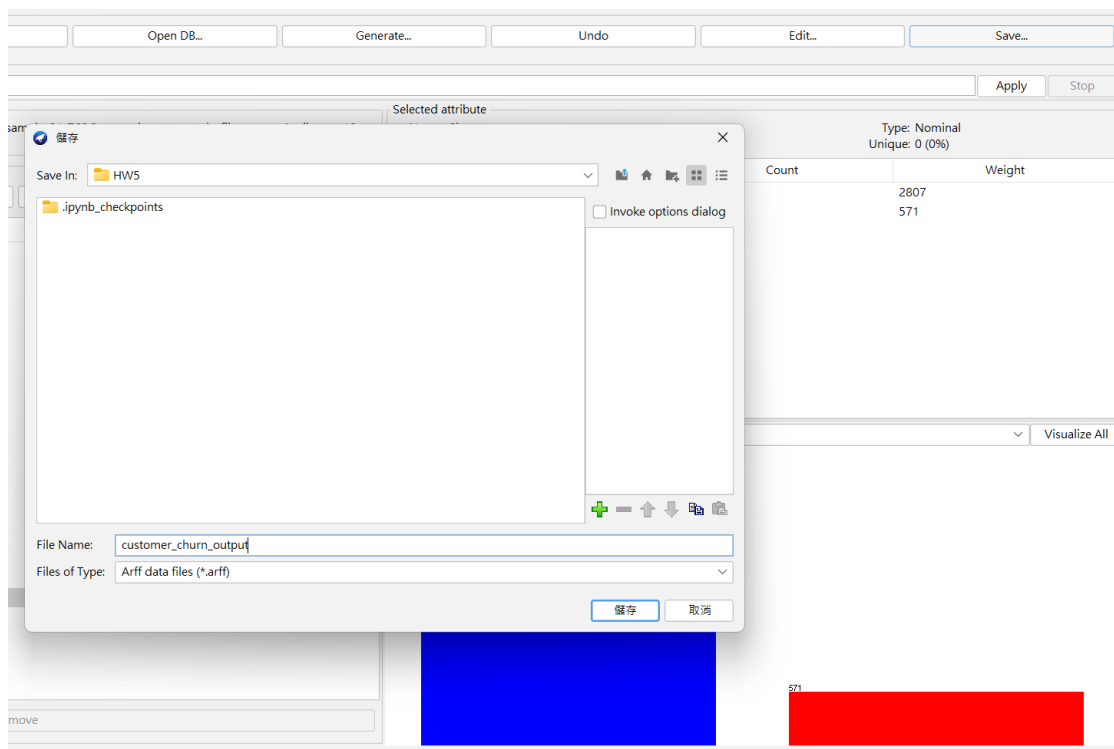
None

Invert

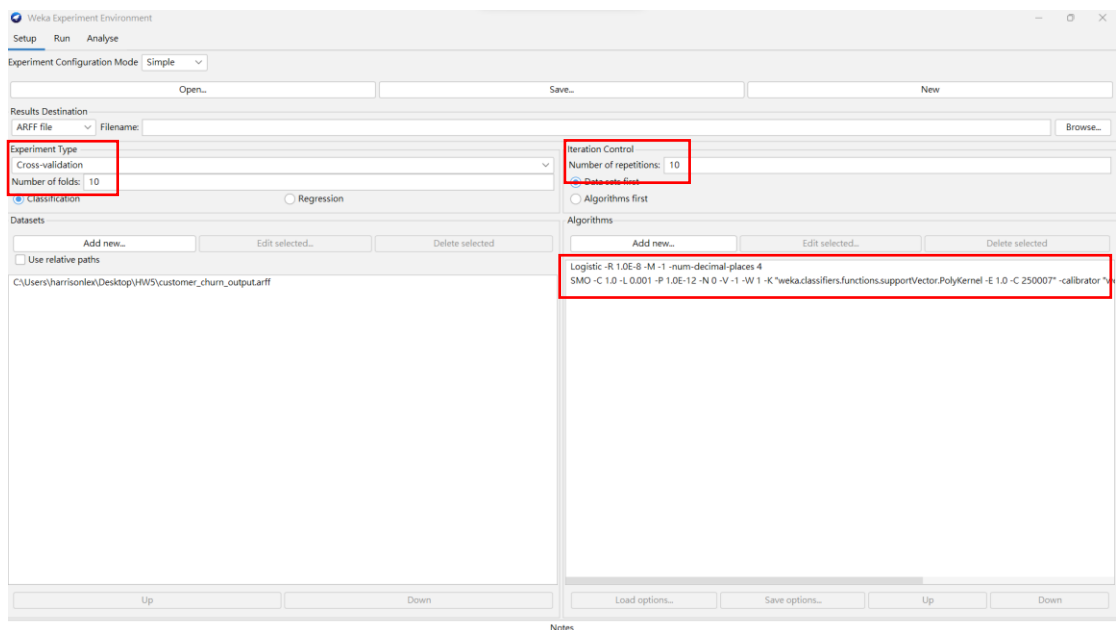
Pattern

| No. | | Name |
|-----|-------------------------------------|-----------------------------|
| 1 | <input type="checkbox"/> | Tenure |
| 2 | <input type="checkbox"/> | PreferredLoginDevice |
| 3 | <input type="checkbox"/> | CityTier |
| 4 | <input type="checkbox"/> | WarehouseToHome |
| 5 | <input type="checkbox"/> | PreferredPaymentMode |
| 6 | <input type="checkbox"/> | Gender |
| 7 | <input type="checkbox"/> | HourSpendOnApp |
| 8 | <input type="checkbox"/> | NumberOfDeviceRegistered |
| 9 | <input type="checkbox"/> | PreferedOrderCat |
| 10 | <input type="checkbox"/> | SatisfactionScore |
| 11 | <input type="checkbox"/> | MaritalStatus |
| 12 | <input type="checkbox"/> | NumberOfAddress |
| 13 | <input type="checkbox"/> | Complain |
| 14 | <input type="checkbox"/> | OrderAmountHikeFromlastYear |
| 15 | <input type="checkbox"/> | CouponUsed |
| 16 | <input type="checkbox"/> | OrderCount |
| 17 | <input type="checkbox"/> | DaySinceLastOrder |
| 18 | <input type="checkbox"/> | CashbackAmount |
| 19 | <input checked="" type="checkbox"/> | Churn |

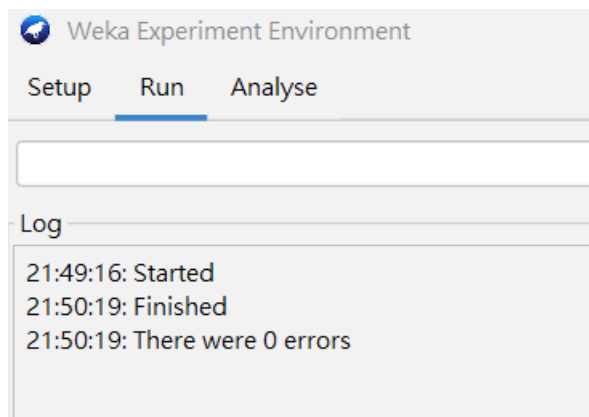
然後按 save 儲存檔案



來到 Experimenter。Datasets 選擇剛輸出的檔案，Algorithms 選到 Logistic 和 SMO(兩模型參數我都用 default 的)，Iteration Control 的 Number of repetitions 設為 10，Experiment Type 選 Cross-validation 並 Number of folds 設為 10 且選 Classification



跑完



得到結果：



5. 結論

根據第 4 題最後得出的 t-test 結果可以看到：

| Dataset | (1) functions.Logi | (2) functions.S |
|--------------------------------|--------------------|-----------------|
| 'customer_churn-weka.filt(100) | 88.85(1.31) | 88.84(1.22) |
| | (v/ /*) | (0/1/0) |

兩個模型的平均 Accuracy 分別為 88.85 和 88.84，標準差分別為 1.31 和 1.22。而且兩模型並沒有被標註 v 或* (表示某模型顯著較好/較差)，表示兩模型的平均 Accuracy 並無顯著差異。

| | | | |
|----------------------|-----------------|------|------|
| Select rows and cols | Rows | Cols | Swap |
| Comparison field | Percent_correct | | |
| Significance | 0.05 | | |
| Sortina (asc.) by | <default> | | |

Analysing: Percent_correct
Datasets: 1
Resultsets: 2
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 2023/5/2 下午9:51

由於顯著水準設置為 0.05，可得出的結論就是：

在信心水準 95%下，此次實驗 SVM 模型和 Logistic Regression 表現上並無顯著差異。