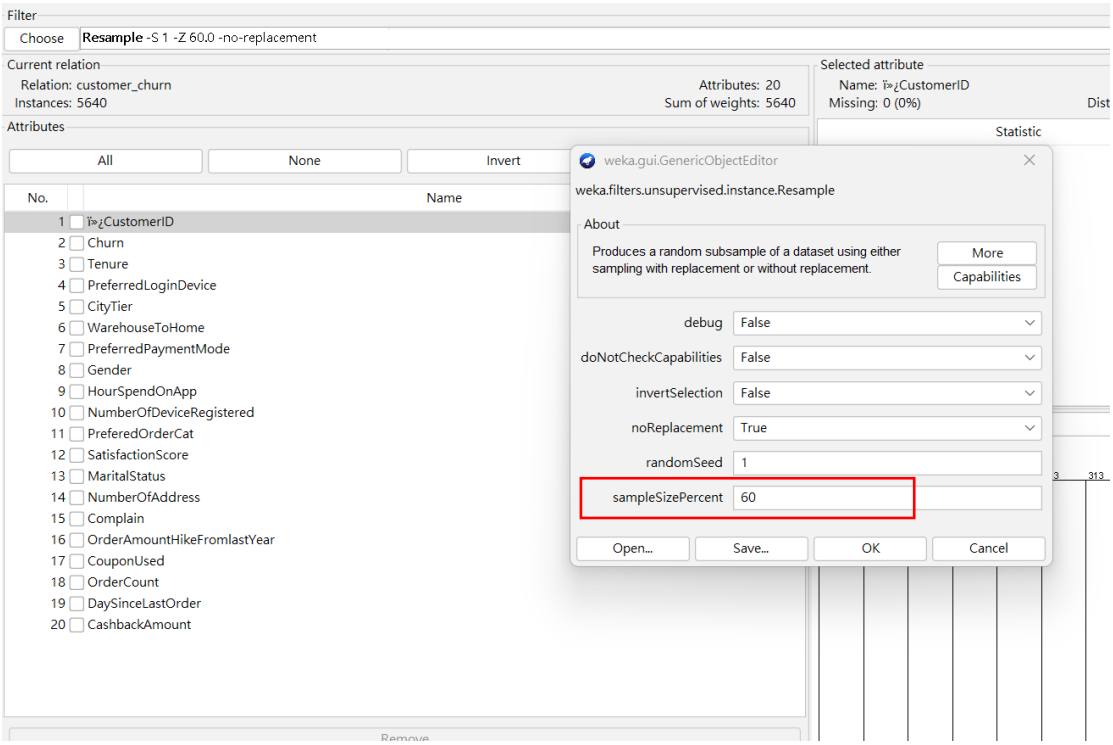
ECT_HW5_109403021_Weka 題

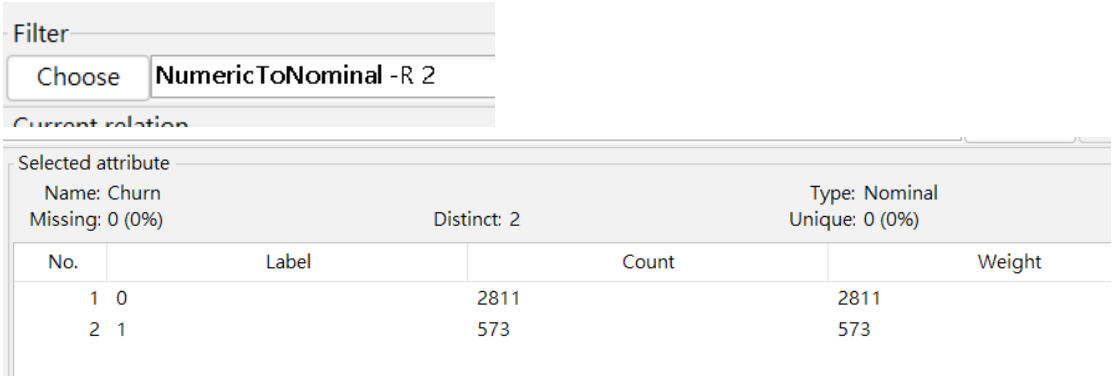
1. 取 60%資料

從 Filter 選取 Resample 並將 sampleSizePercent 調為 60 後 Apply



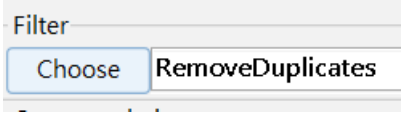
2. 列處 churn 資料個數

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

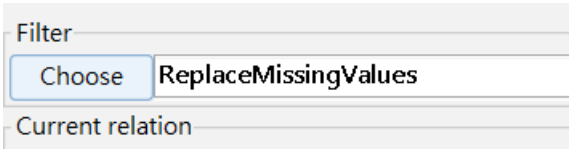


3. 資料前處理

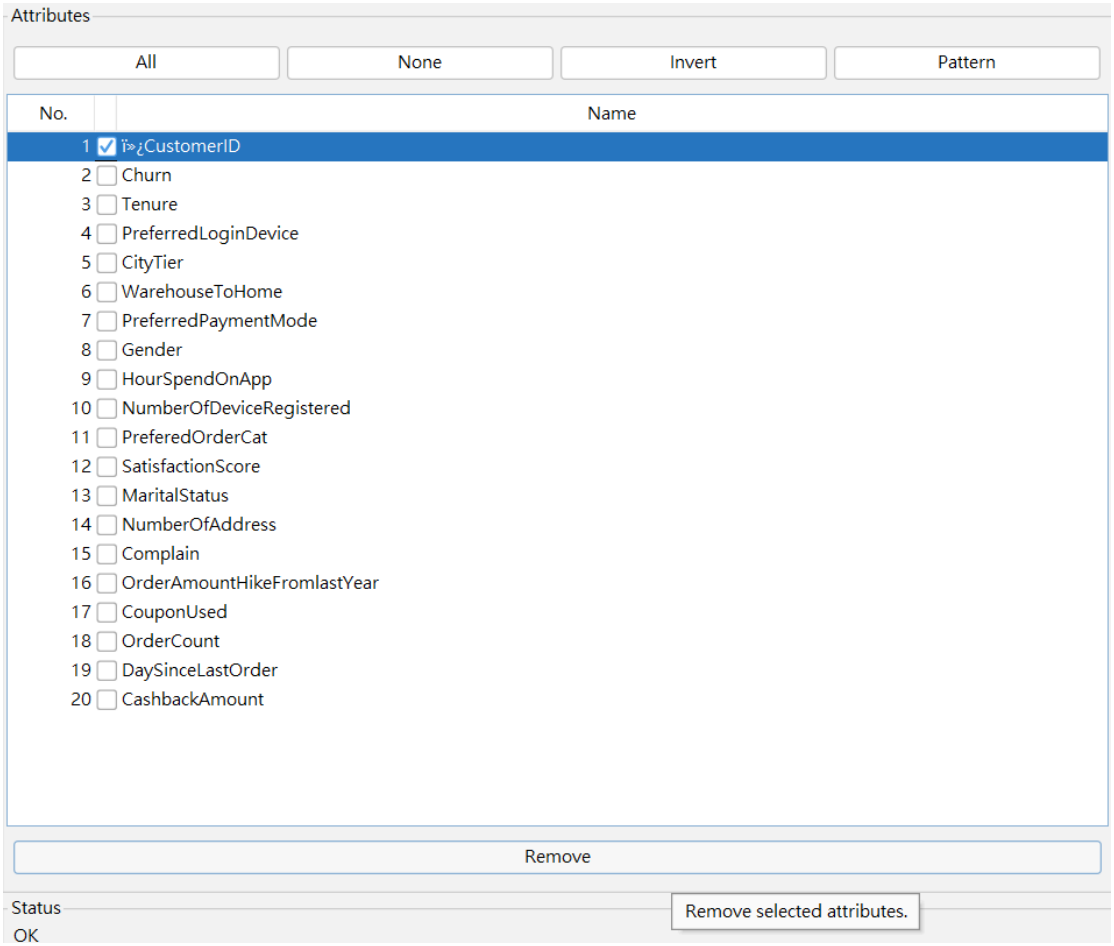
(1)先刪除重複多餘的資料(保留一筆)



(2)用平均數填補空值

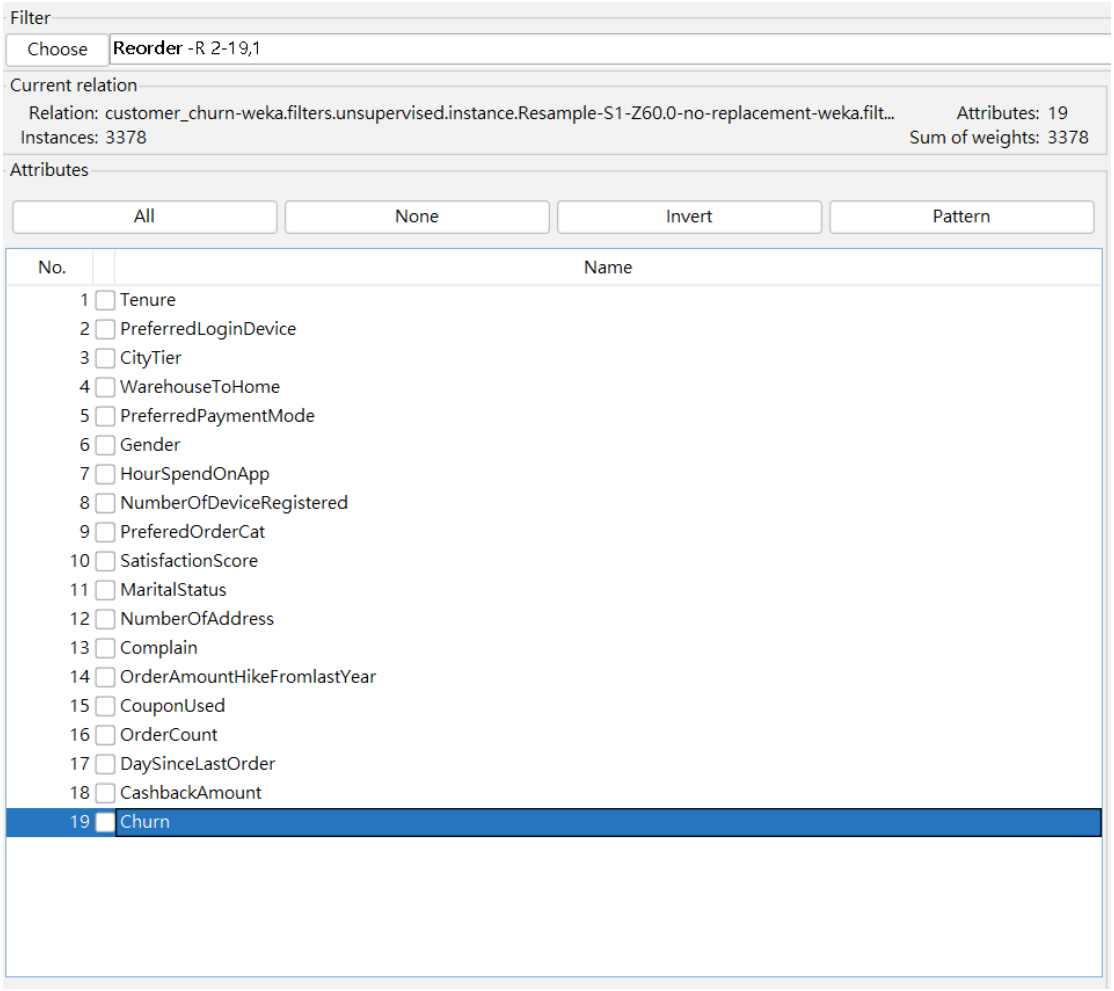


(3)刪除 CustomerID 欄位

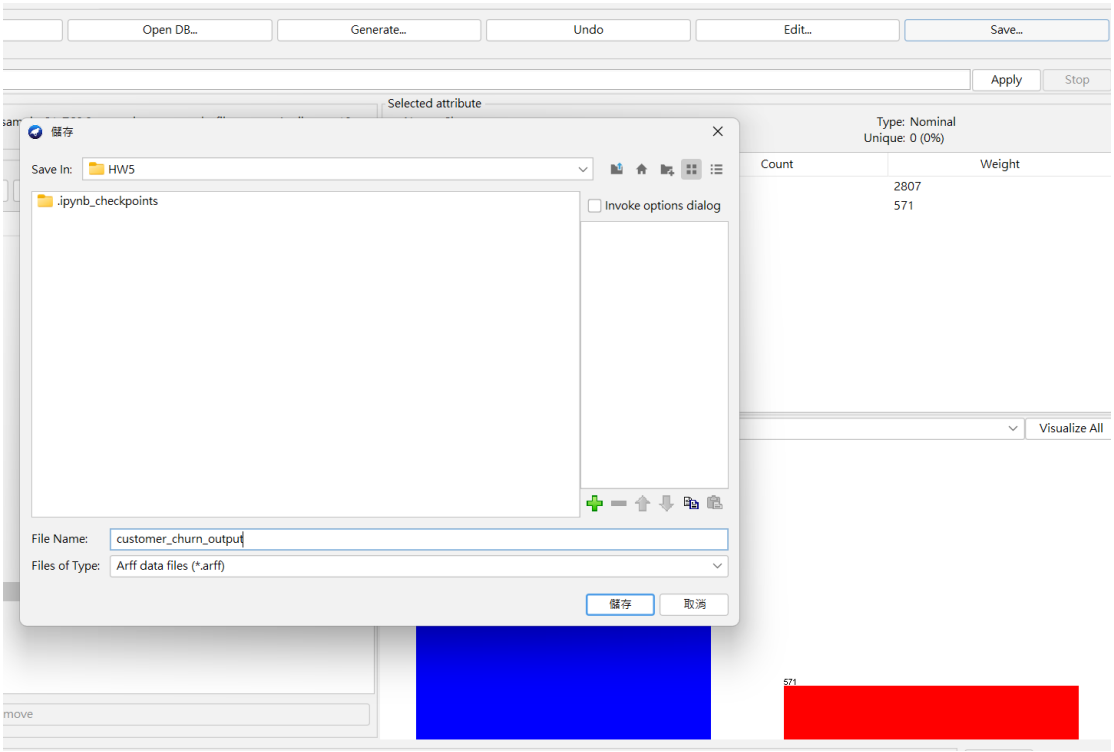


4. 用 Experimenter 比較兩模型

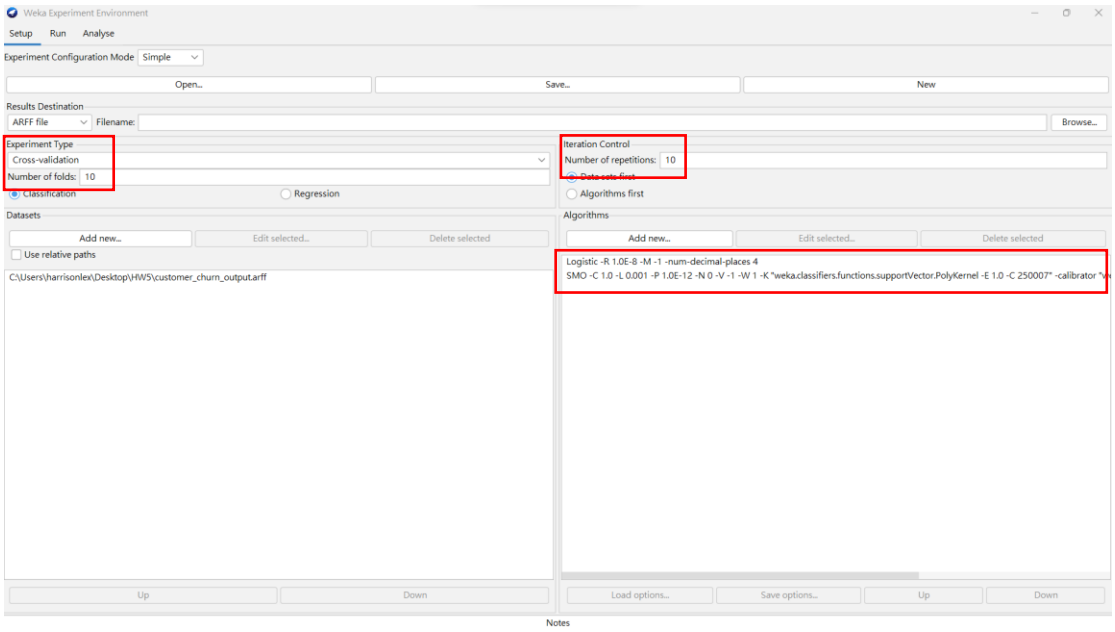
我先把 Churn 欄位調到最後讓等等 experimenter 直接識別其為目標變數



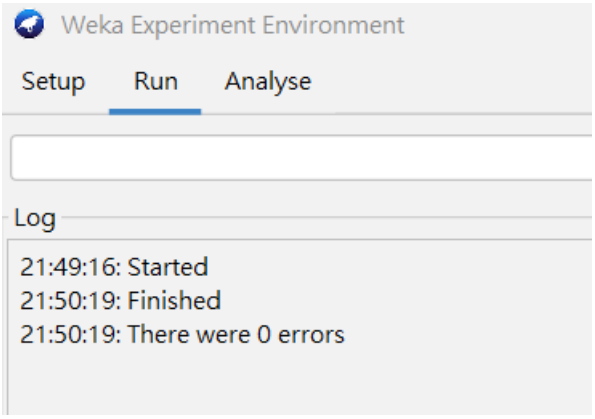
然後按 save 儲存檔案



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



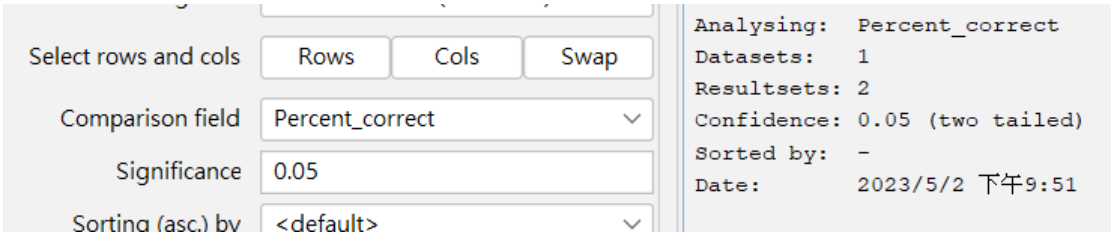
得到結果:

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。



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。