2019 ECT 作業五

1. 請用 python 依照步驟對 voice.csv 進行 SVM 分析，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。

首先，引用所需套件，並讀取 csv 檔，

```
import pandas as pd
import numpy as np

#讀取CSV檔案
data = pd.read_csv('voice.csv')
```

使用 Python：(60%)

(a) 請檢查資料集是否有空值，如有空值即去掉該筆資料

運用 dropna()函數去掉資料的空值

```
data.dropna()
```

(b) 將最後一個屬性值”label”切分為 Target，其餘屬型切分為 Feature

```
Target = data['label']
#將屬性合併
#變成list
feature=list(zip(data['meanfreq'],data['sd'],data['median'],data['Q25'],data['Q75'],data['IQR'],
                 data['skew'],data['kurt'],data['sp.ent'],data['sfm'],data['mode'],data['centroid'],
                 data['meanfun'],data['minfun'],data['maxfun'],data['meandom'],data['mindom'],
                 data['maxdom'],data['dfrange'],data['modindx']))
#轉成array
features=np.asarray(feature)
```

(c) 將 Target 進行 encoded，用 LabelEncoder 將 male 轉為 1，female 轉為 0

```
#轉換屬性型態
#將屬性轉為數字label
le = preprocessing.LabelEncoder()
#將 label 轉為數字label
Target=le.fit_transform(Target)
```

(d) 將 Feature 用 sklearn.preprocessing 的 StandardScaler 進行標準化

```
from sklearn.preprocessing import StandardScaler
scaler  = preprocessing.StandardScaler().fit(features)
feature = scaler.transform(features)
```

(e) 切分資料集與測試集，設 test_size=0.33，random_state=1

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(feature, Target, test_size=0.33, random_state=1)
```

(f) 最後，使用 sklearn.svm 裡的 SVC 進行分析，kernel 設為' linear '，並印出模型最終的準確度

```
from sklearn.svm import SVC
clf = svm.SVC(kernel='linear')
clf.fit(X_train, y_train)
clf.fit(X_test, y_test)
print ('測試集準確度： %0.4f' % clf.score(X_test, y_test))
print ('訓練集準確度： %0.4f' % clf.score(X_train, y_train))
```

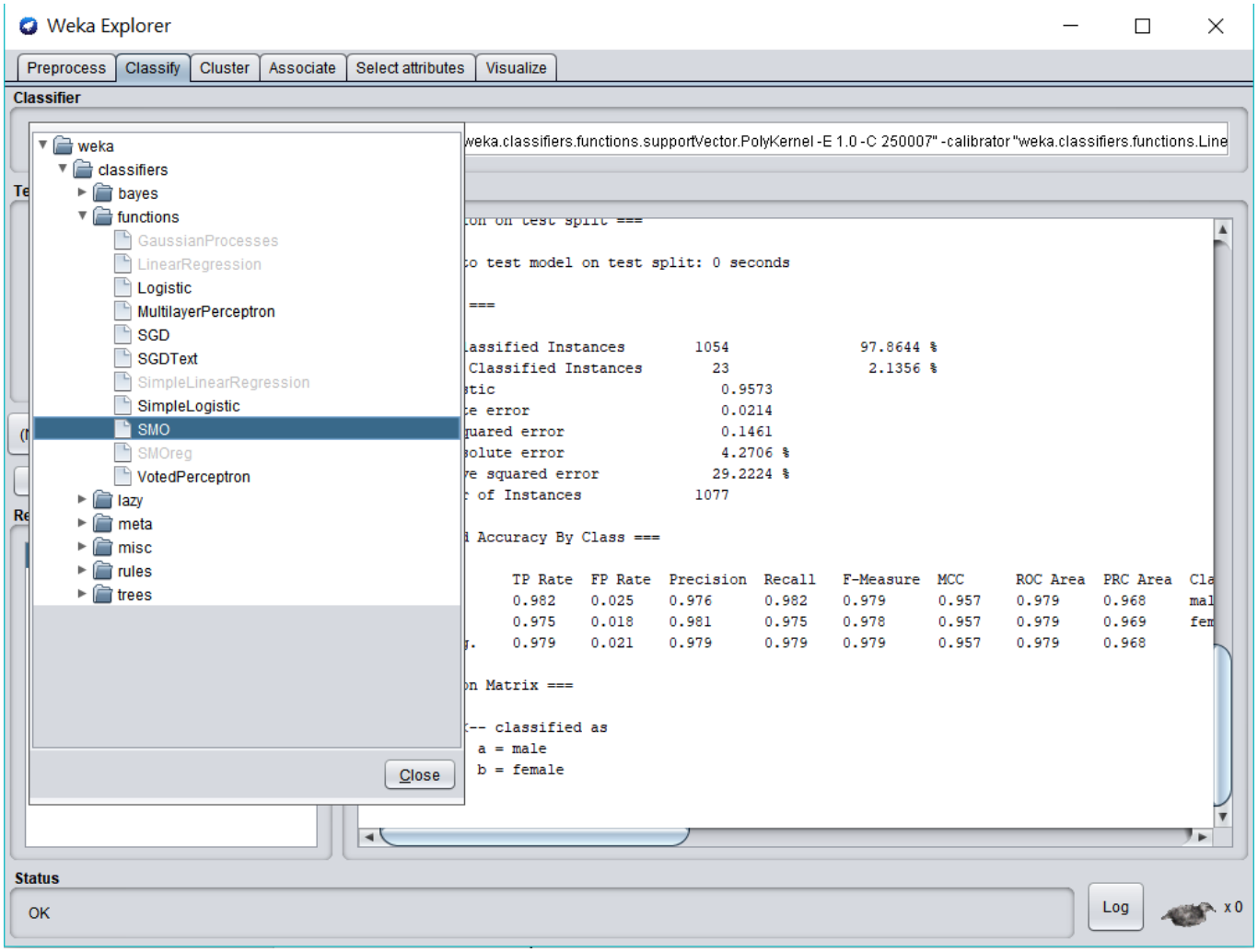
```
測試集準確度： 0.9771
訓練集準確度： 0.9689
```

使用 Weka 軟體：(40%)

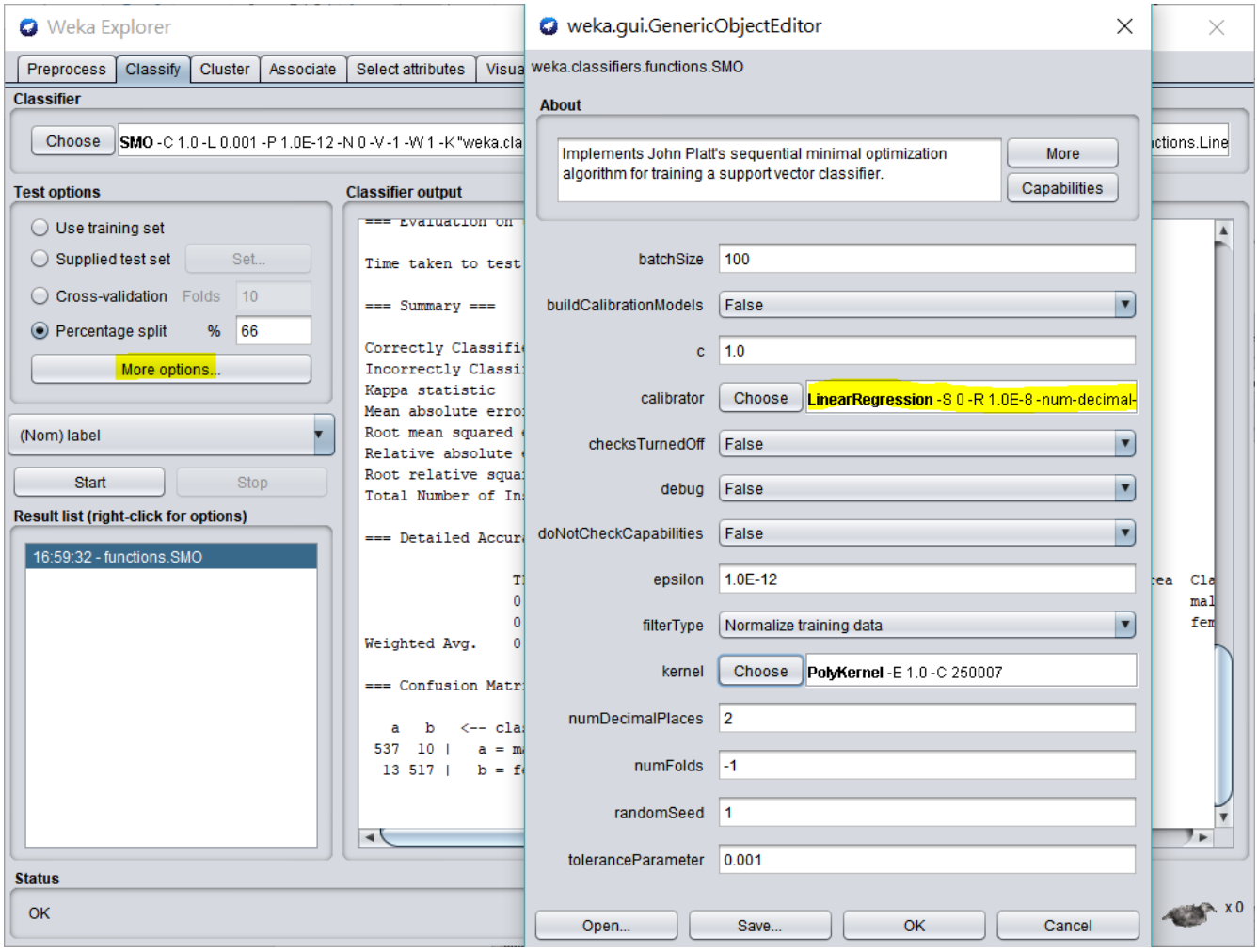
(a) 使用 weka 裡的 SMO function 對 voice.csv 進行 SVM 分析，kernel 設為' linear '，Percentage spilt 設為 66%，截圖並附上過程及準確率

i. 首先在 Weka 中開啟 voice.csv。

ii. 在 Classify 面板中，在 Classifier 選擇「weka / classifiers / functions / SMO」。

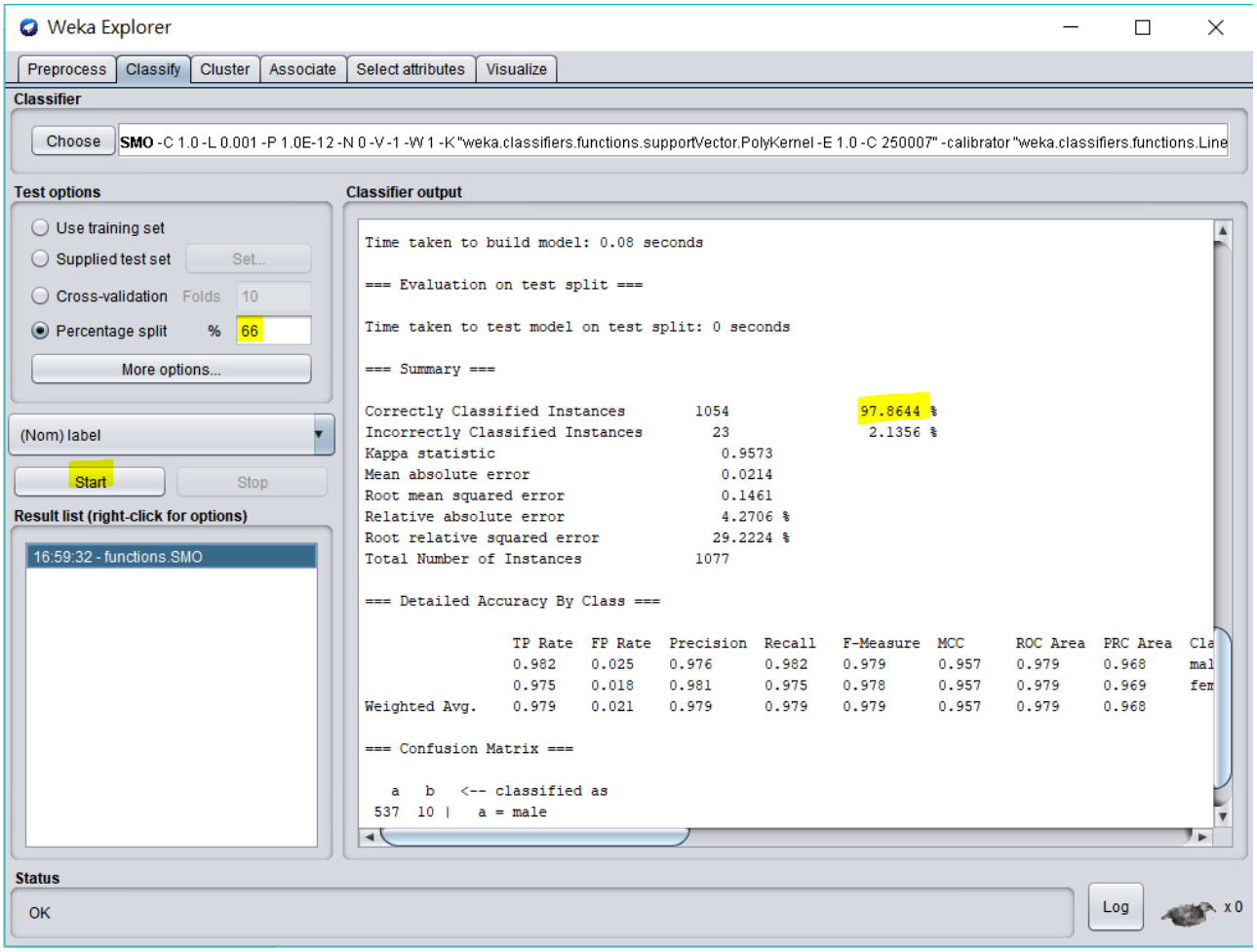


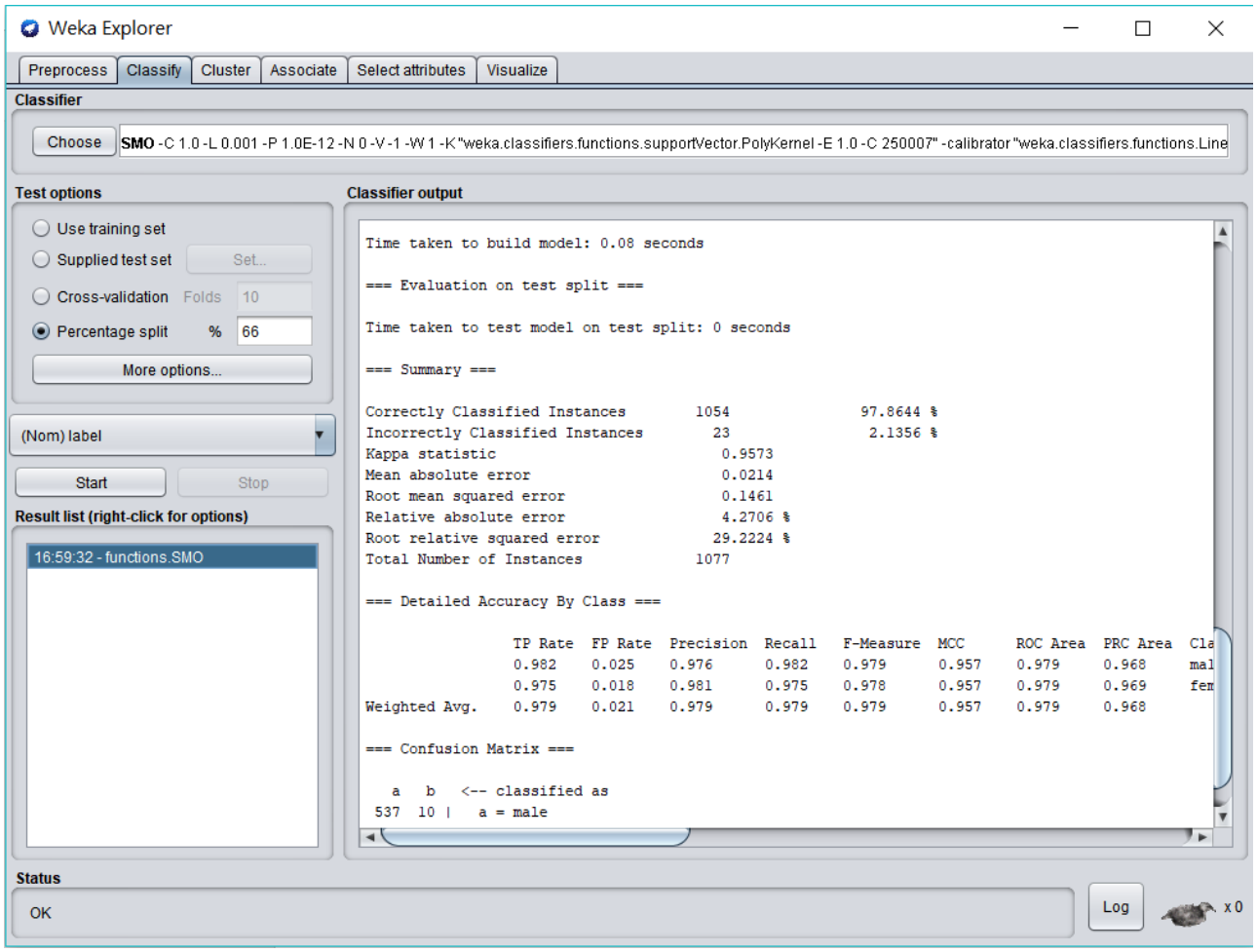
iii. 在「More options」中的「calibrator」選擇「LinearRegression」。



iv. 在「Test options」中選取「Percentage split」，並設定為 66%；選擇預測「(Nom)label」，並點選「Start」。

可得準確度為 97.8664%


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Line
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier output
Time taken to build model: 0.08 seconds
=== Evaluation on test split ===
Time taken to test model on test split: 0 seconds
=== Summary ===
Correctly Classified Instances 1054 97.8644 %
Incorrectly Classified Instances 23 2.1356 %
Kappa statistic 0.9573
Mean absolute error 0.0214
Root mean squared error 0.1461
Relative absolute error 4.2706 %
Root relative squared error 29.2224 %
Total Number of Instances 1077
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.982 0.025 0.976 0.982 0.979 0.957 0.979 0.968 mal
0.975 0.018 0.981 0.975 0.978 0.957 0.979 0.969 fem
Weighted Avg. 0.979 0.021 0.979 0.979 0.979 0.957 0.979 0.968
=== Confusion Matrix ===
a b <-- classified as
537 10 | a = male
(Nom) label
Start
Stop
Result list (right-click for options)
16:59:32 - functions.SMO
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Line
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) label
Start
Stop
Result list (right-click for options)
16:59:32 - functions.SMO
Classifier output
Time taken to build model: 0.08 seconds
=== Evaluation on test split ===
Time taken to test model on test split: 0 seconds
=== Summary ===
Correctly Classified Instances 1054 97.8644 %
Incorrectly Classified Instances 23 2.1356 %
Kappa statistic 0.9573
Mean absolute error 0.0214
Root mean squared error 0.1461
Relative absolute error 4.2706 %
Root relative squared error 29.2224 %
Total Number of Instances 1077
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.982 0.025 0.976 0.982 0.979 0.957 0.979 0.968 mal
0.975 0.018 0.981 0.975 0.978 0.957 0.979 0.969 fem
Weighted Avg. 0.979 0.021 0.979 0.979 0.979 0.957 0.979 0.968
=== Confusion Matrix ===
a b <-- classified as
537 10 | a = male
Status
OK
Log
x 0