資料分析

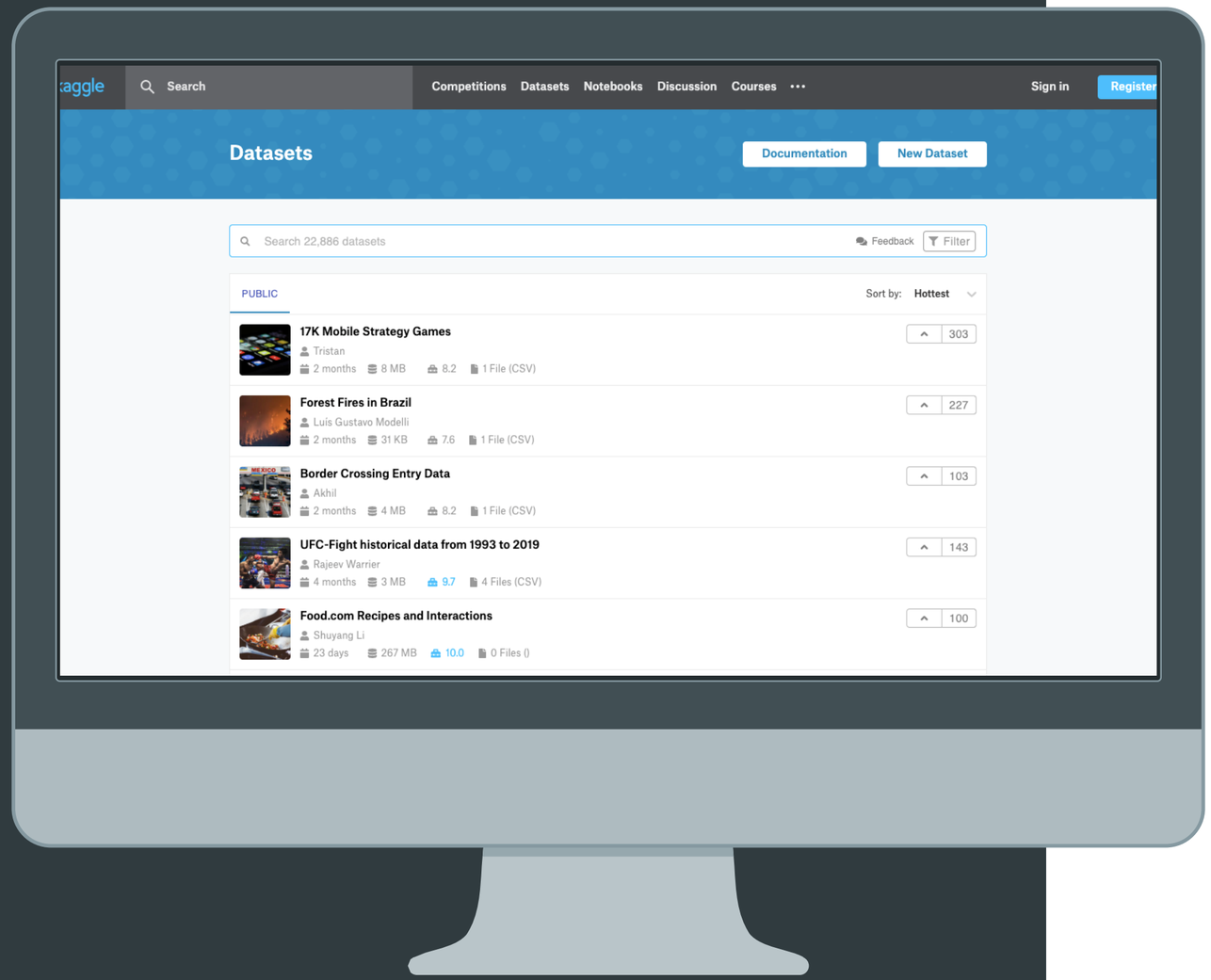
Gender Recognition by Voice

組員：

00557103游俊弘

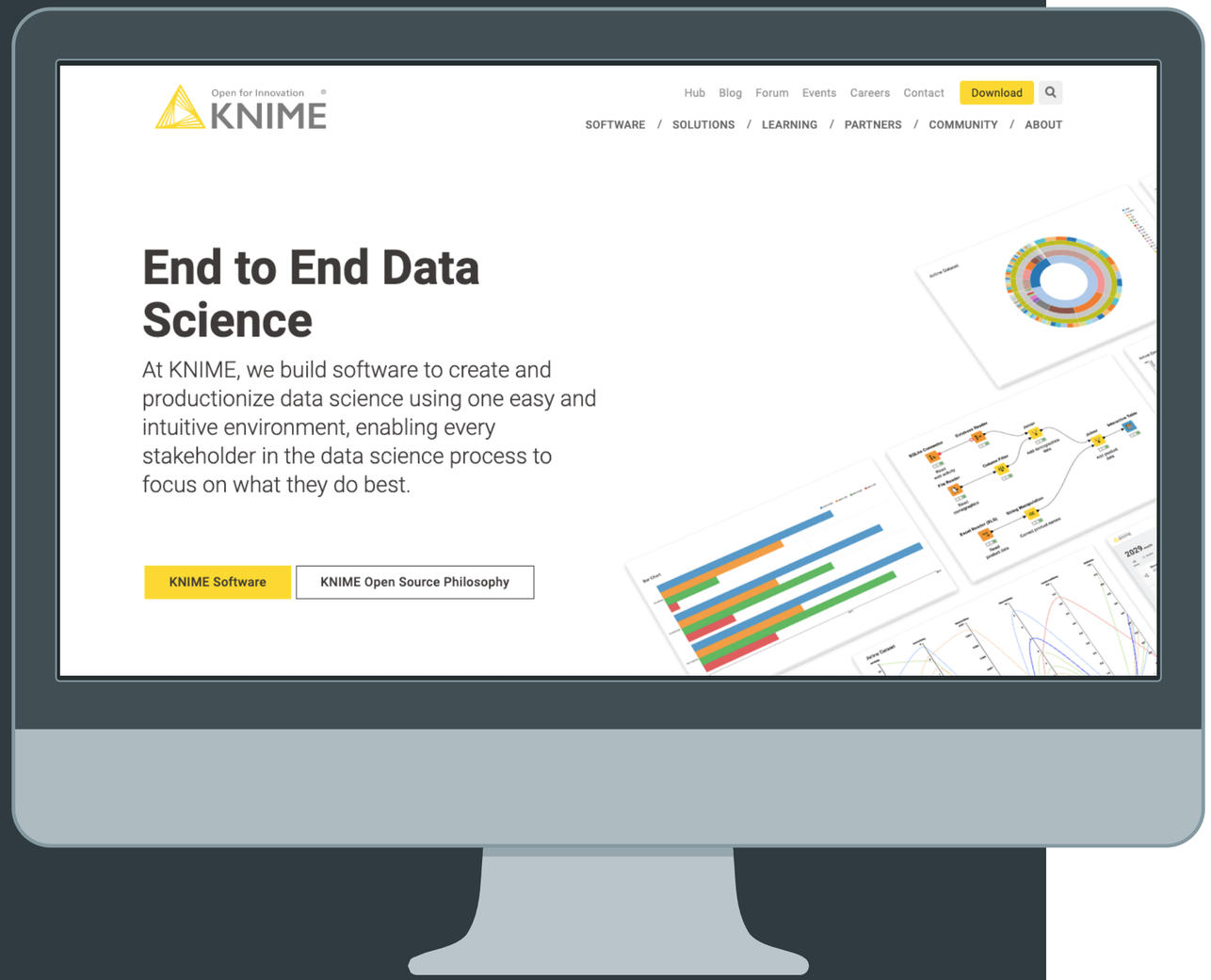
00657207林弈呈

00457131丁兆文



Use the dataset By Kaggle

https://www.kaggle.com/primaryobjects/voicegender



Analyze the dataset By KNIME

https://www.knime.com/

TABLE OF CONTENTS

SECTION 1

What's your data mining problem?

Gender Recognition by Voice and Speech Analysis

Identify a voice as male or female

- This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech.
- The dataset consists of 3,168 recorded voice samples, collected from male and female speakers.
- The voice samples are pre-processed by acoustic analysis in R using the seewave and tuneR packages, with an analyzed frequency range of 0hz-280hz

Question

■主要

1. 男性和女性聲音之間還有哪些其他特徵？
2. 可以發現男性和女性聲音在共鳴(resonance)上有區別嗎？

■次要

1. 可以從正常聲音中識別假音(falsetto)嗎？（為此可能需要單獨的數據集）
2. 數據中還有其他有趣的功能嗎？

“以人耳看來，用聲音決定性別能否依靠簡單的頻率來決策呢？”

SECTION 2

Data understanding

Dataset

data categories

- Dataset: 3618*21
- Filename: voice.csv
- Data based on acoustic properties of the voice and speech (frequence:kHz)
  - meanfreq、sd、median、Q25、Q75、IQR、skew、kurt、sp.ent、sfm、mode、centroid、meanfun、minfun、maxfun、meandom、mindom、maxdom、dfrange、modindx、class

Data Attribute-acoustic properties

//頻率以kHz為單位

- meanfreq:平均頻率
- sd:頻率標準差
- median:中位數頻率
- Q25、Q75:第1、3四分位數
- IQR:四分位距(Q75-Q25)
- skew:偏斜
- kurt:峰度
- sp.ent:頻譜熵
- sfm:頻譜平坦度
- mode: 頻率眾數
- centroid:頻率重心
- meanfun:跨聲學信號測得的基本頻率的平均值
- minfun:跨聲學信號測得的最小基頻
- maxfun:跨聲學信號測得的最大基頻
- meandom:整個聲信號測得的主頻的平均值
- mindom:跨聲學信號測得的最小主頻
- maxdom:跨聲信號測得的主頻率最大值
- dfrange:跨聲信號測得的主頻範圍
- modindx:調製指數。計算為相鄰基頻測量之間的累計絕對差除以頻率範圍
- class:男性或女性

Meandom(average dominant frequency)

- 口語中的頻率變化很大，更不用說整個句子了。頻率隨著語調而上升和下降，通常是在單詞和語音中傳達某些情感。這可能使查明準確的頻率變得困難。
- 可以通過使用每個語音樣本中測得的平均主導頻率。
- 關於性別，平均主導頻率的確具有統計學意義。因為皆是正值，因此這支持了基本假設，即頻率的增加與女性的語音分類相對應。

Fundamental frequency(基本頻率)

- males ranges from 100 to 150 Hz
- females ranges from 180 to 250 Hz

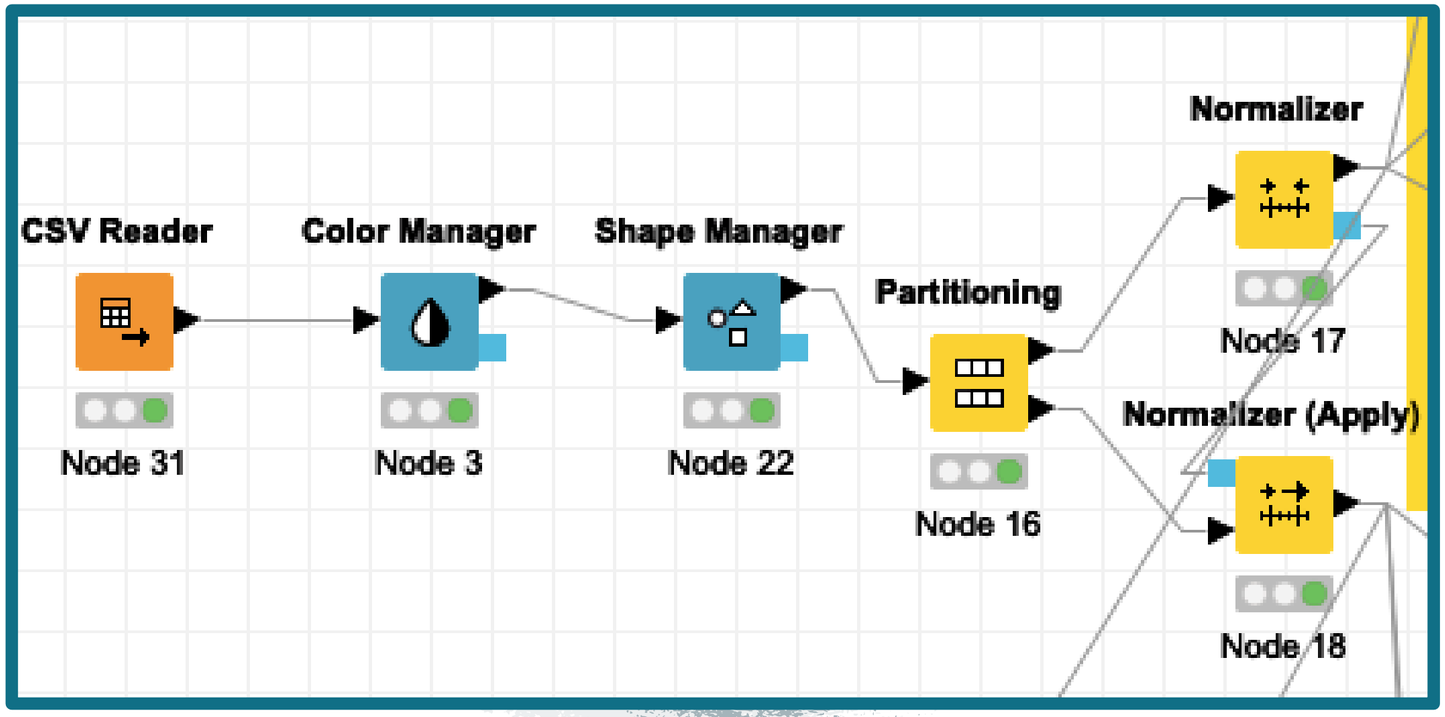
SECTION 3

Data preparation

Analysis

Use different algorithm

- 找到適合的演算法，達到精準的分析
- 用不同的演算法進行分析，例如
  a. Logistic Regression
  b. Random forest
  c. SVM
- 對dataset進行feature selection以及降維(using PCA)



Data
preparation

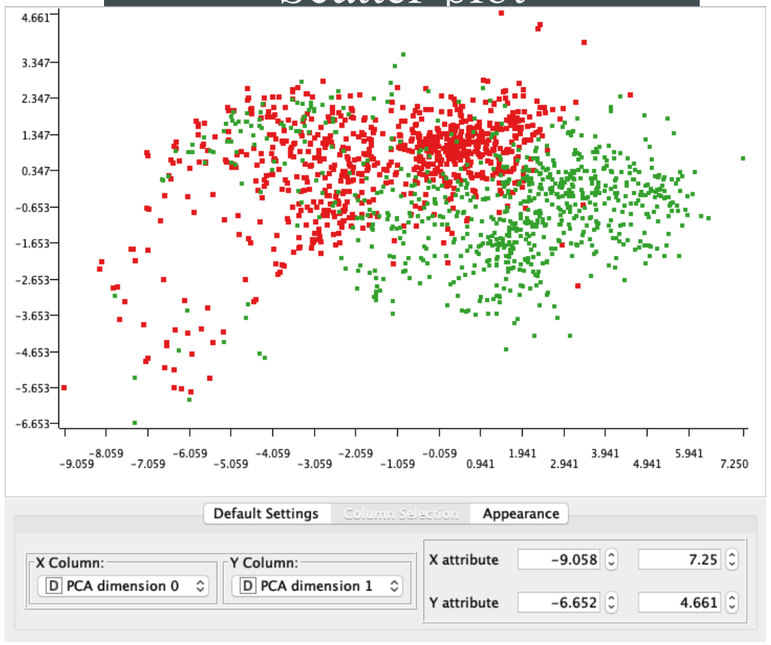
Import資料庫檔案(voice.csv) → 進行color/shape manager
→ partitioning(relative 50%) → Normallizer(using z-score)

SECTION 4

Result and Discussion

PCA Dimension Reduction to 2dim.

Scatter plot

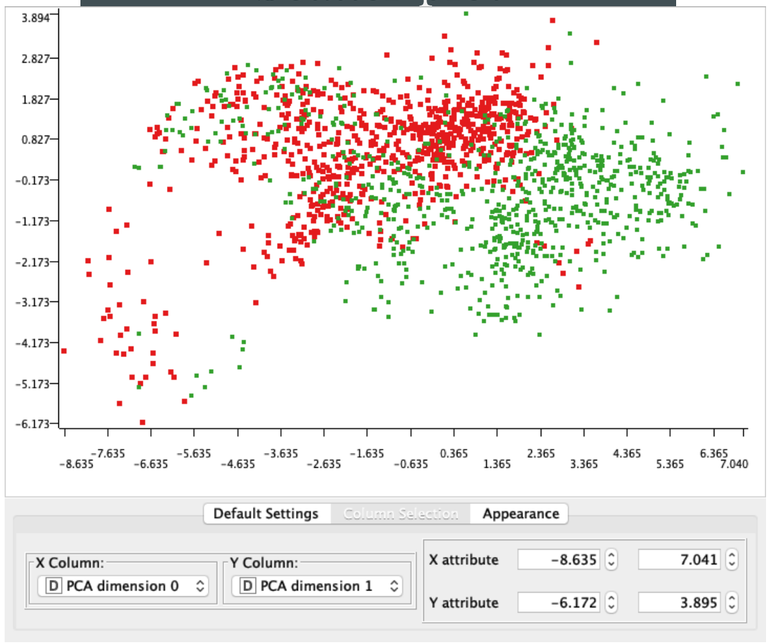


- PCA training

X:PCA0

Y:PCA1

Scatter plot



- PCA testing

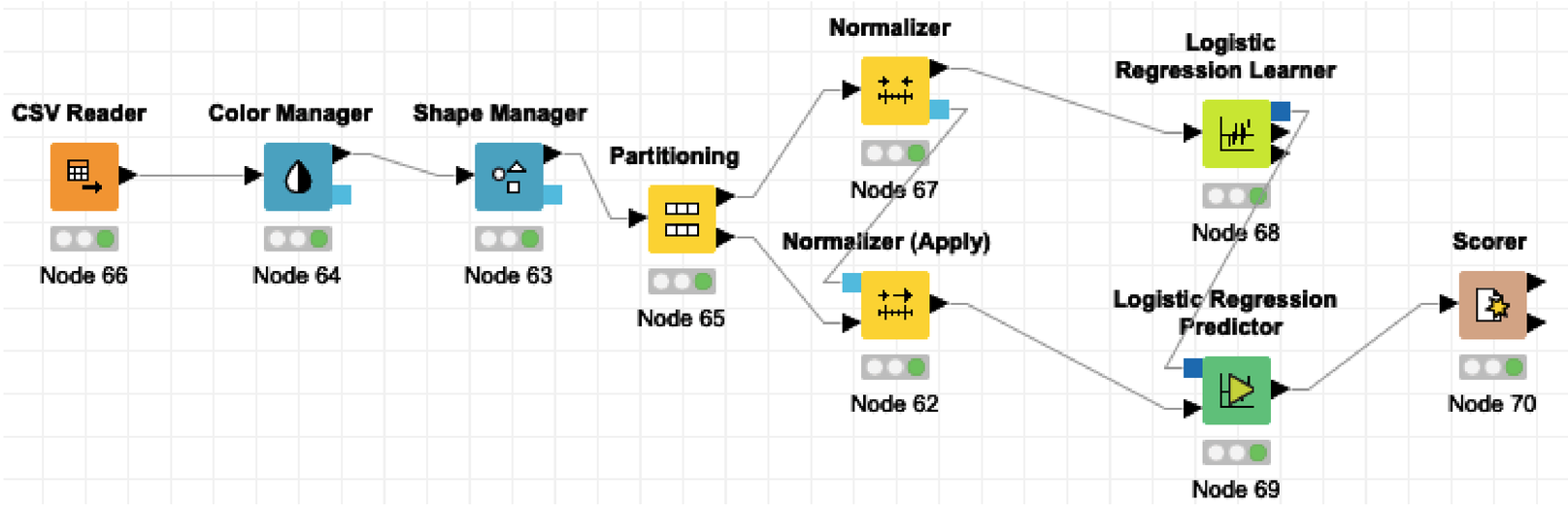
X:PCA0

Y:PCA1

Use Only meandom

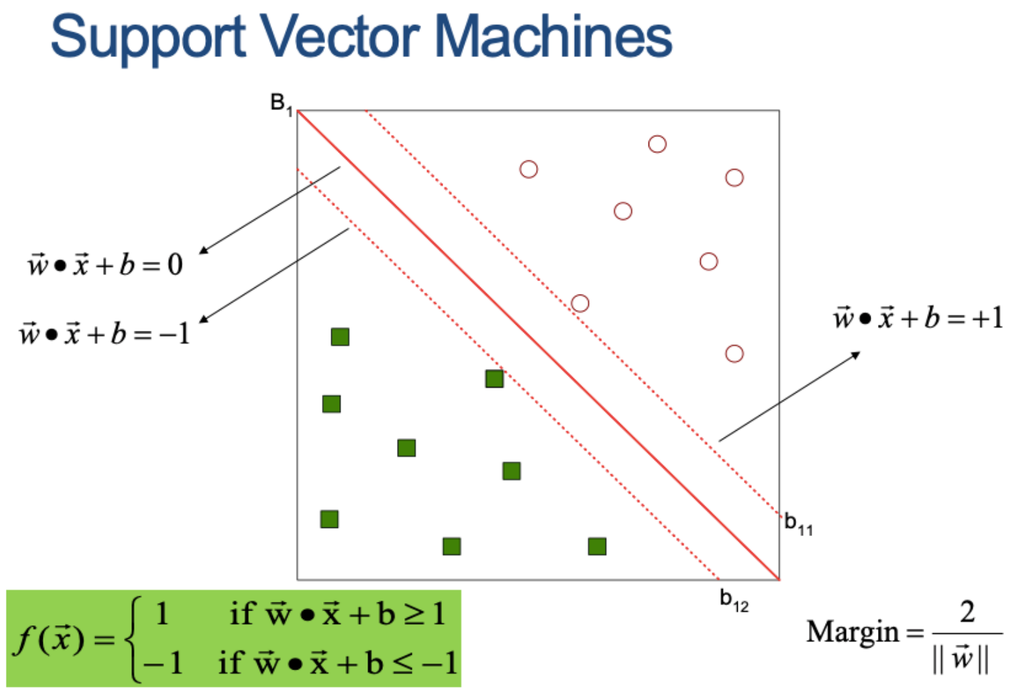
- 另外，我們對於單一頻率，只使用meandom，用邏輯回歸模型透過訓練，在測試上的準確性為56.313%。準確度達到一半左右，但仍遠不能準確檢測男性/女性聲音。

Correct classified: 892	Wrong classified: 692
Accuracy: 56.313 %	Error: 43.687 %
Cohen's kappa (κ) 0.126	



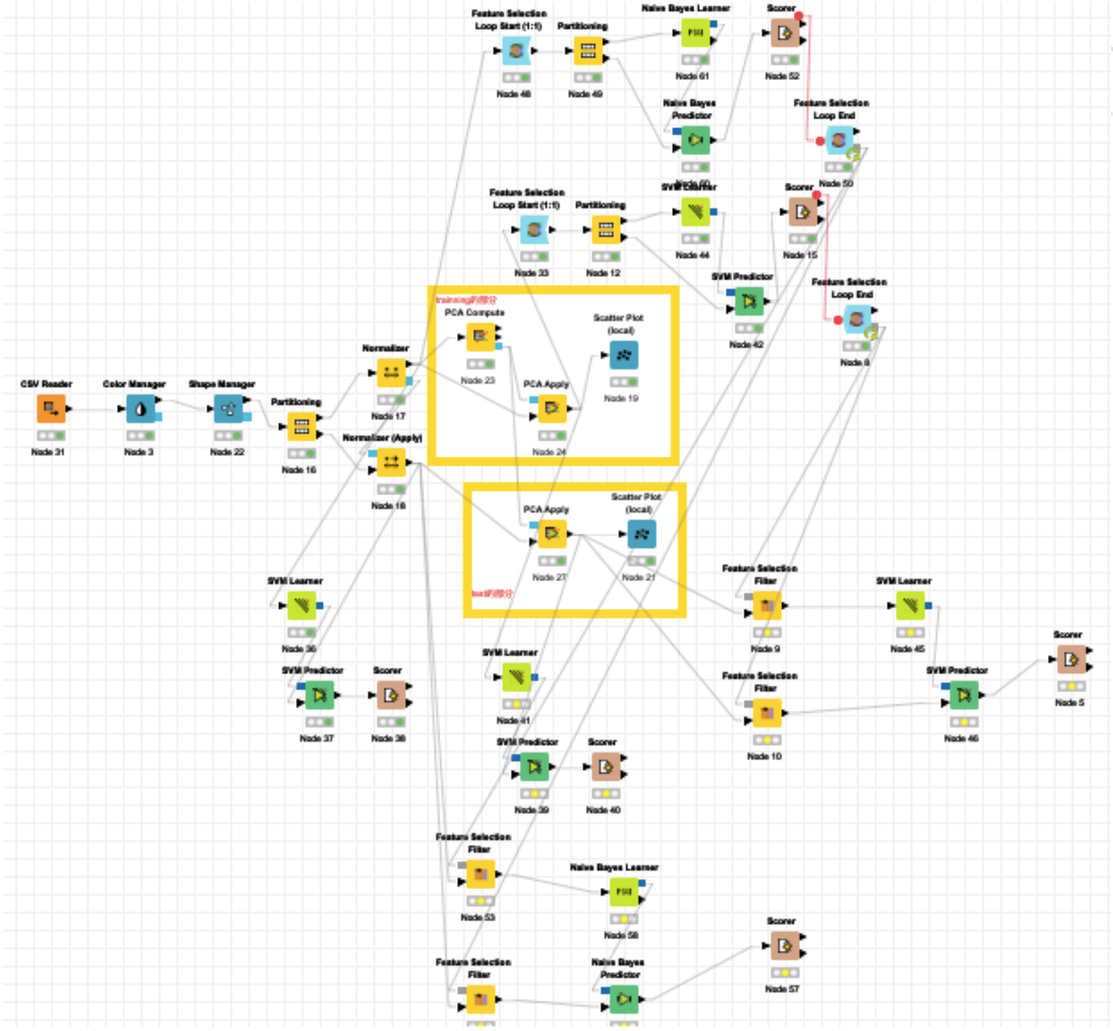
SVM

- SVM是一種監督式的學習(Supervised learning)方法，其概念非常簡單，就是找到一個決策邊界(decision boundary)讓兩類之間的邊界(margins)最大化，使其可以完美區隔開來。



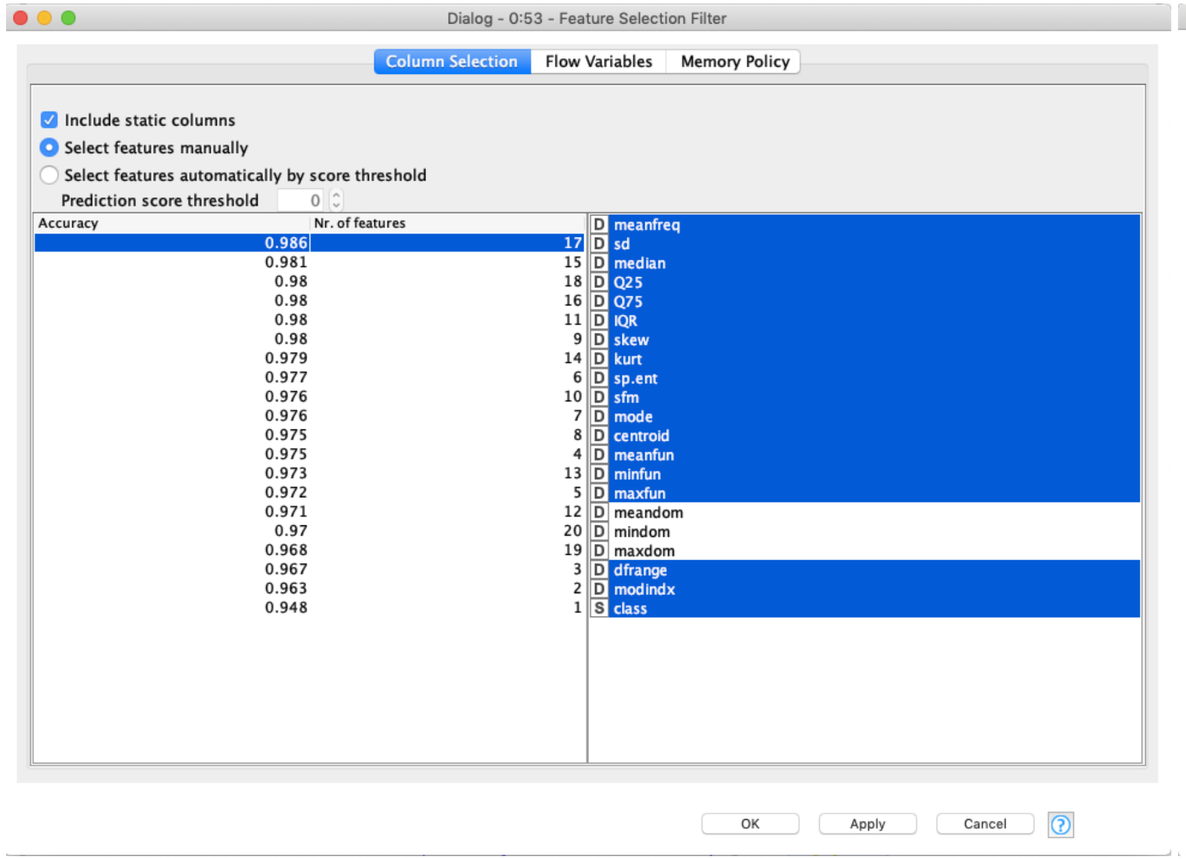
PCA	Feature selection	Accuracy(%) by SVM
×	×	97.475
×	✓	97.727
✓	×	97.475
✓	✓	97.033

SVM workflows

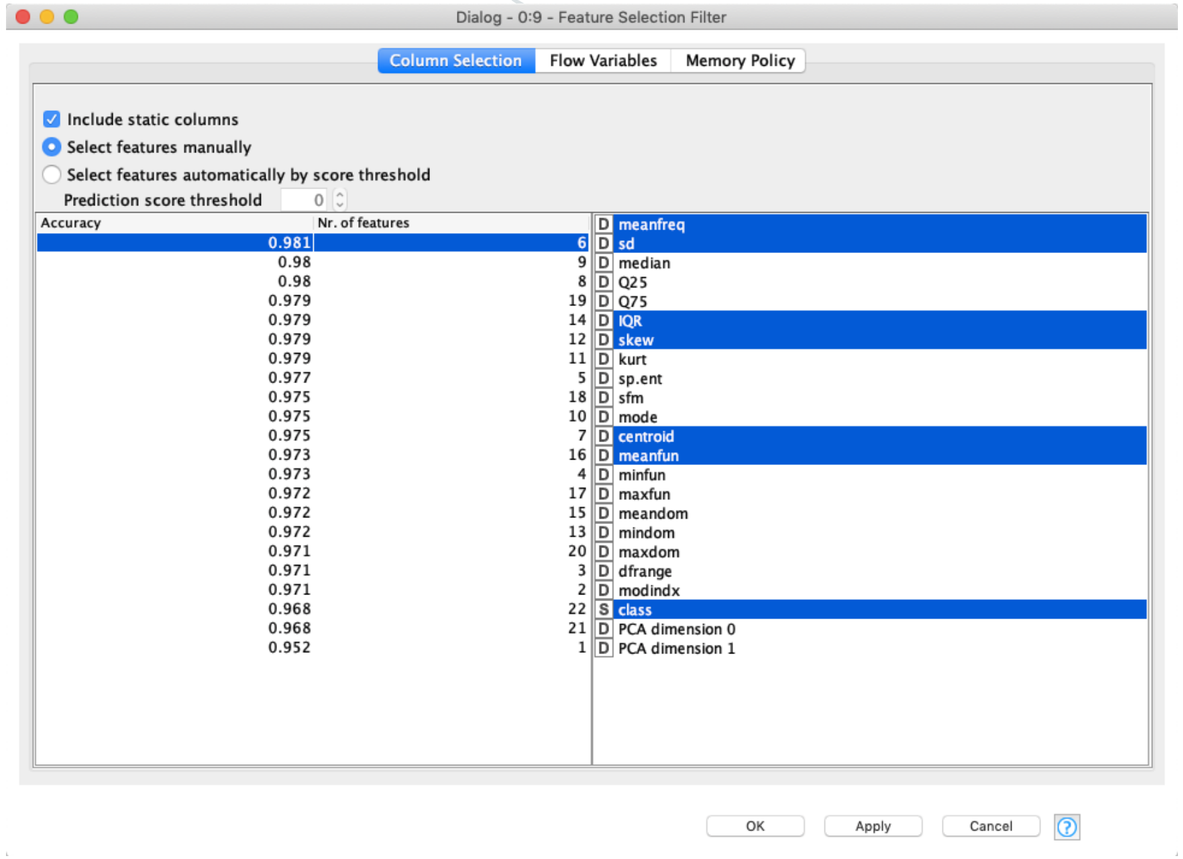


Feature selection for SVM

Feature selection



Feature selection + PCA



SVM Confusion matrix

None and PCA

class \ Pre...	male	female
male	768	24
female	16	776

Correct classified: 1,544 | Wrong classified: 40

Accuracy: 97.475 % | Error: 2.525 %

Cohen's kappa (κ) 0.949

Feature selection

class \ Pre...	male	female
male	773	19
female	17	775

Correct classified: 1,548 | Wrong classified: 36

Accuracy: 97.727 % | Error: 2.273 %

Cohen's kappa (κ) 0.955

Feature selection + PCA

class \ Pre...	male	female
male	769	23
female	24	768

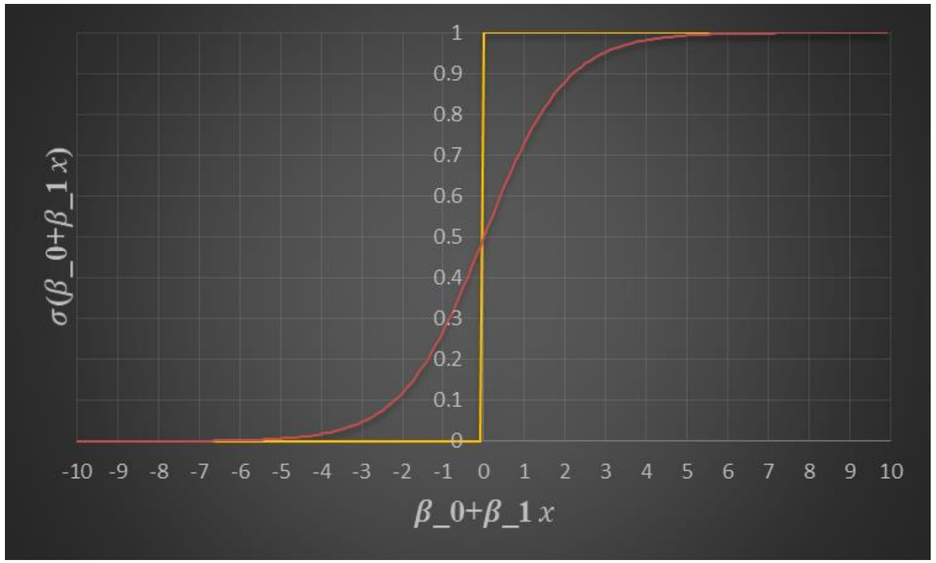
Correct classified: 1,537 | Wrong classified: 47

Accuracy: 97.033 % | Error: 2.967 %

Cohen's kappa (κ) 0.941

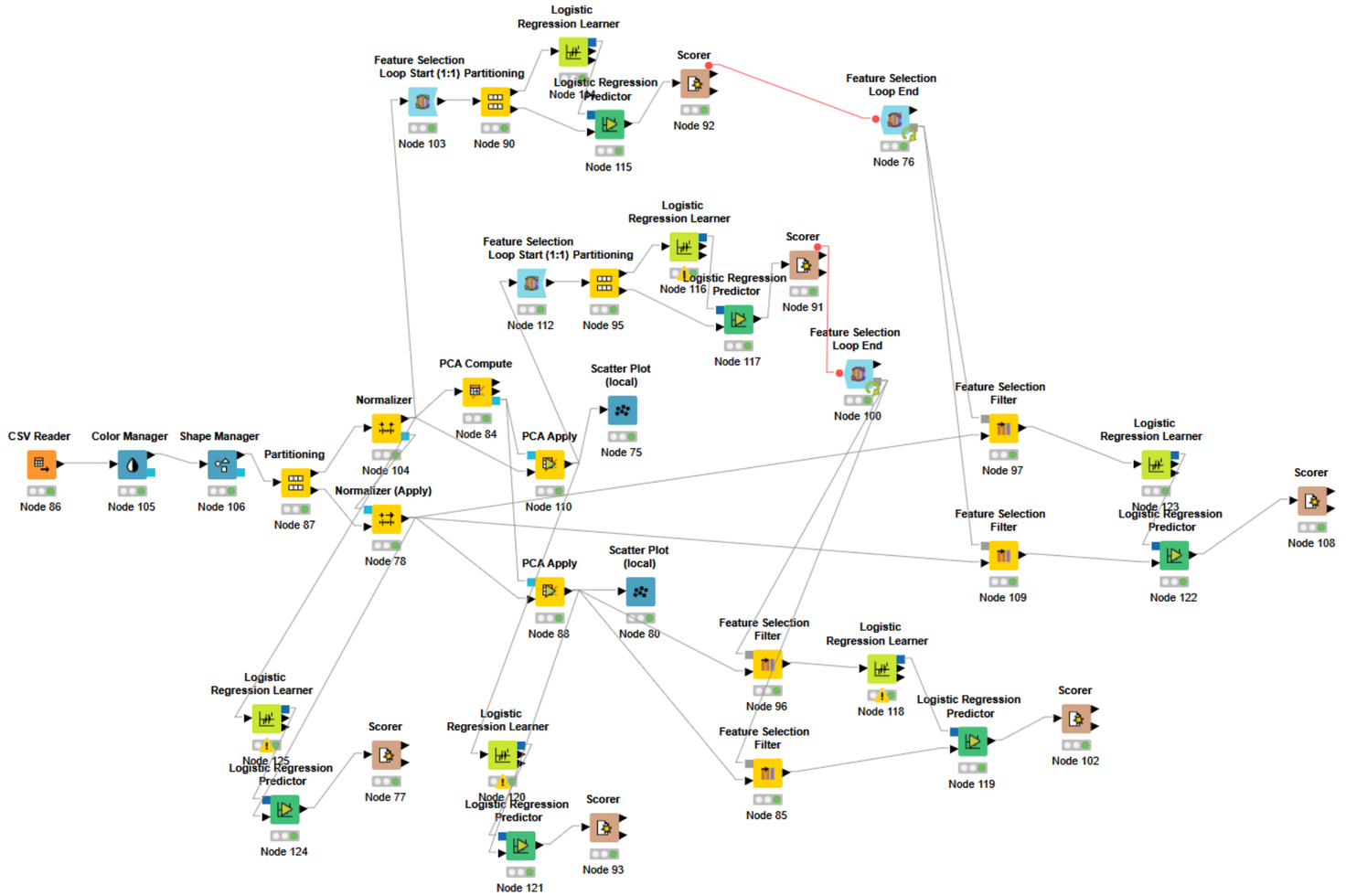
Logistic Regression

- 一種統計模型，其基本形式是使用logistic函數對二進制因變量進行建模，儘管存在許多更複雜的擴展。在回歸分析中，邏輯回歸是在評估邏輯模型的參數



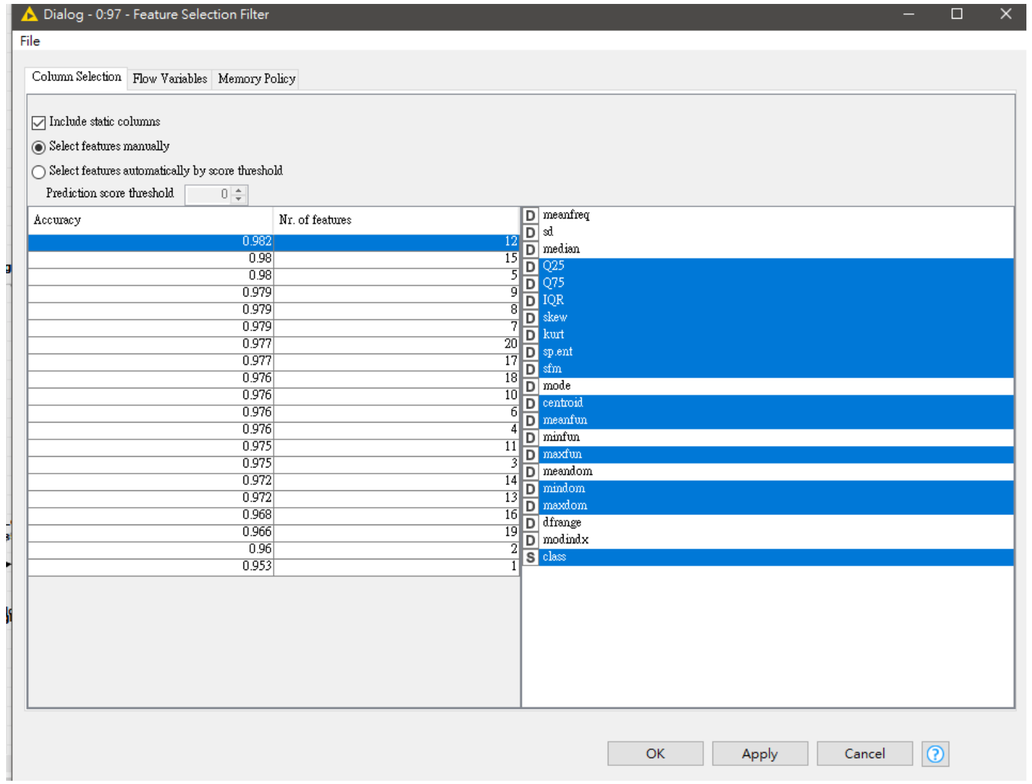
PCA	Feature selection	Accuracy(%) by Logistic Regression
×	×	96.465
×	✓	96.402
✓	×	96.843
✓	✓	97.412

Logistic Regression workflows

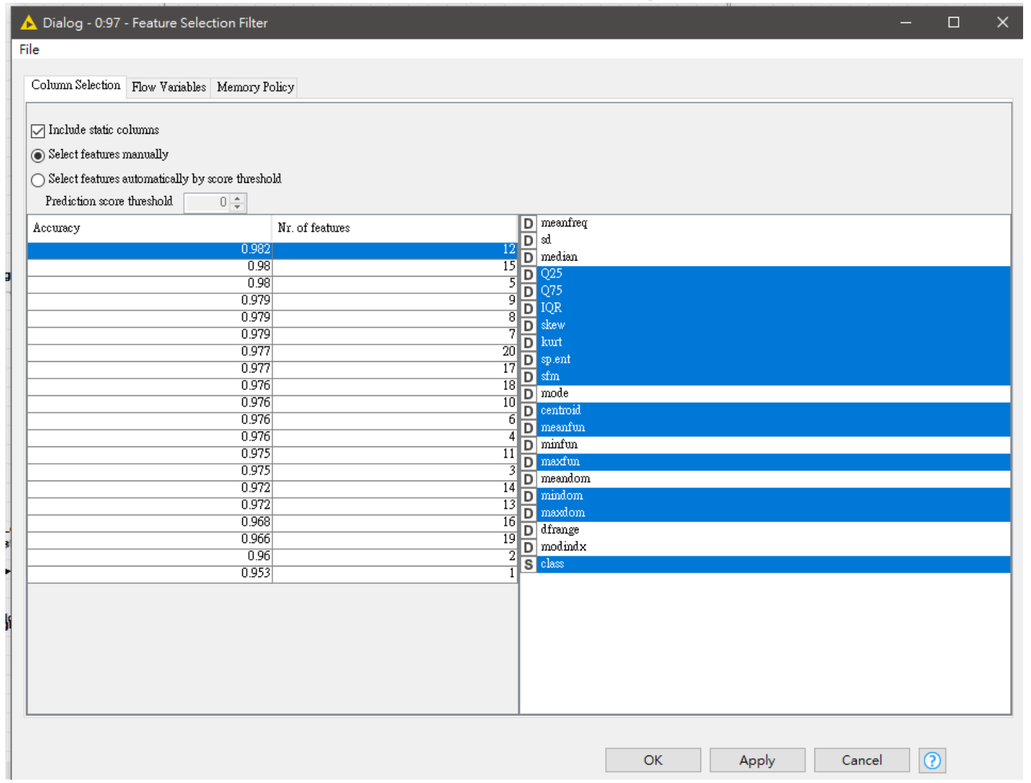


Feature selection for Logistic Regression

Feature selection



Feature selection + PCA



Logistic Regression Confusion matrix

None

class \ Predi...	male	female
male	773	19
female	37	755

Correct classified: 1,528
Wrong classified: 56
Accuracy: 96.465 %
Error: 3.535 %
Cohen's kappa (K) 0.929

PCA

class \ Predi...	male	female
male	774	18
female	39	753

Correct classified: 1,527
Wrong classified: 57
Accuracy: 96.402 %
Error: 3.598 %
Cohen's kappa (K) 0.928

Feature selection

class \ Predi...	male	female
male	773	19
female	31	761

Correct classified: 1,534
Wrong classified: 50
Accuracy: 96.843 %
Error: 3.157 %
Cohen's kappa (K) 0.937

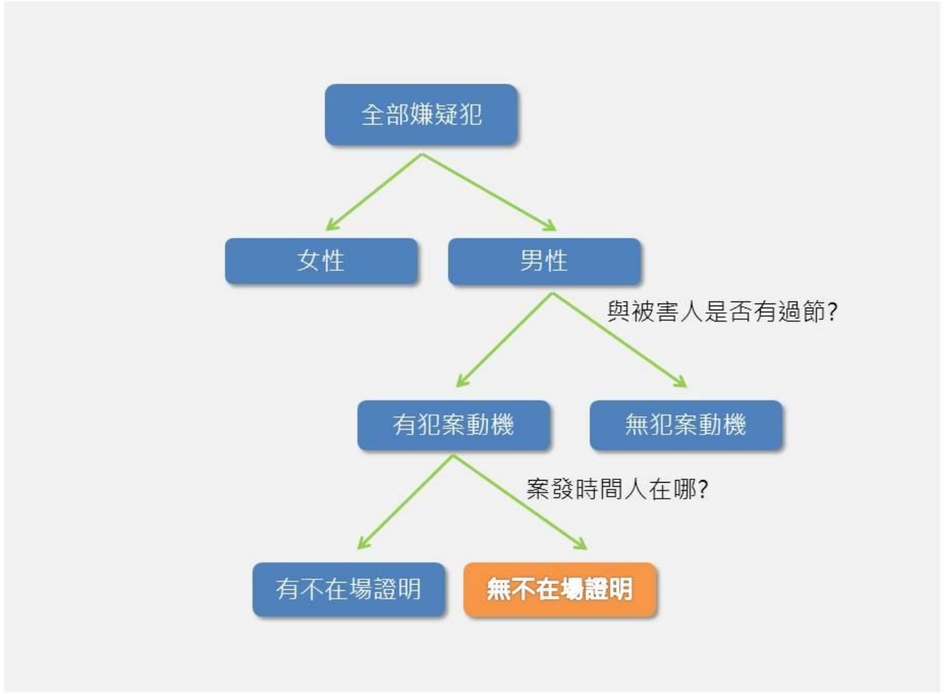
Feature selection + PCA

class \ Predi...	male	female
male	775	17
female	24	768

Correct classified: 1,543
Wrong classified: 41
Accuracy: 97.412 %
Error: 2.588 %
Cohen's kappa (K) 0.948

Random Forest

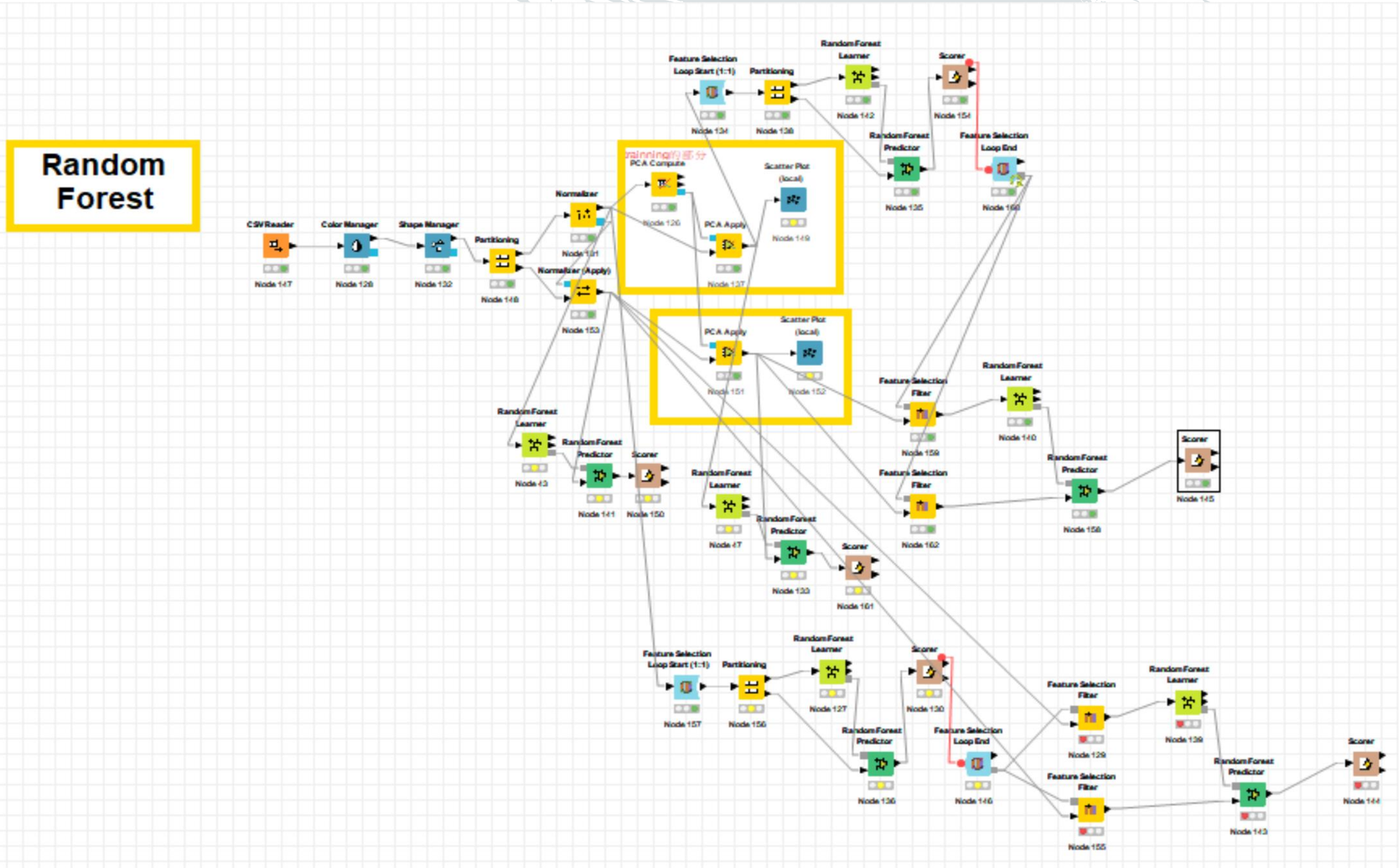
- 隨機森林其實就是進階版的決策樹，這句話簡單說明，就是很多顆樹加起來可以變成一座森林。



決策樹

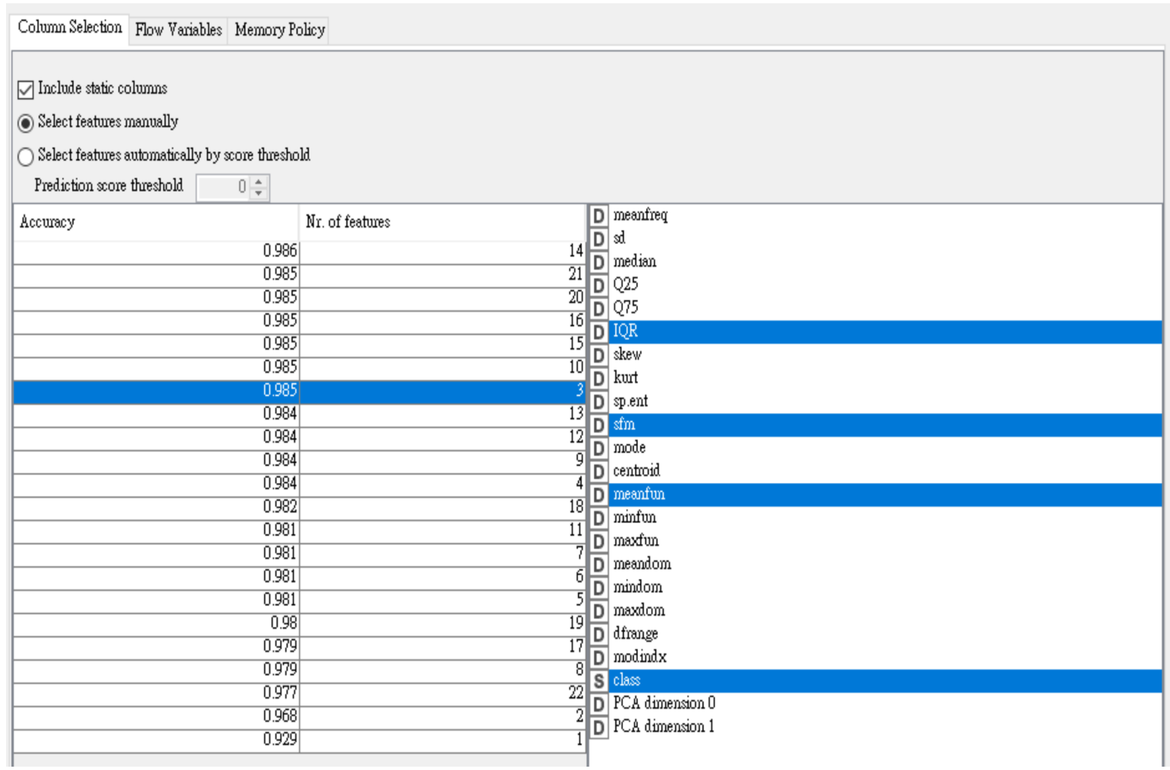
PCA	Feature selection	Accuracy(%) by Random Forest
✕	✕	97.917
✕	✓	100
✓	✕	97.917
✓	✓	100

Random Forest workflows

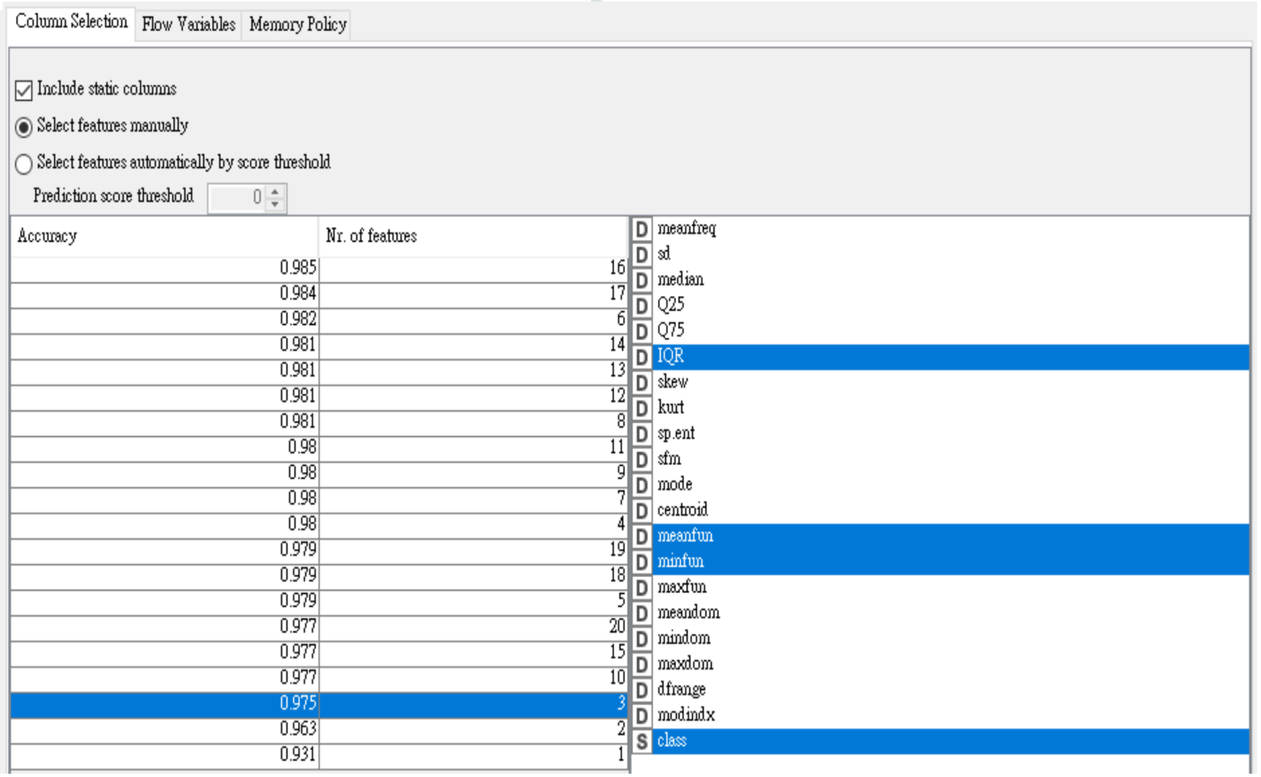


Feature selection for Random Forest

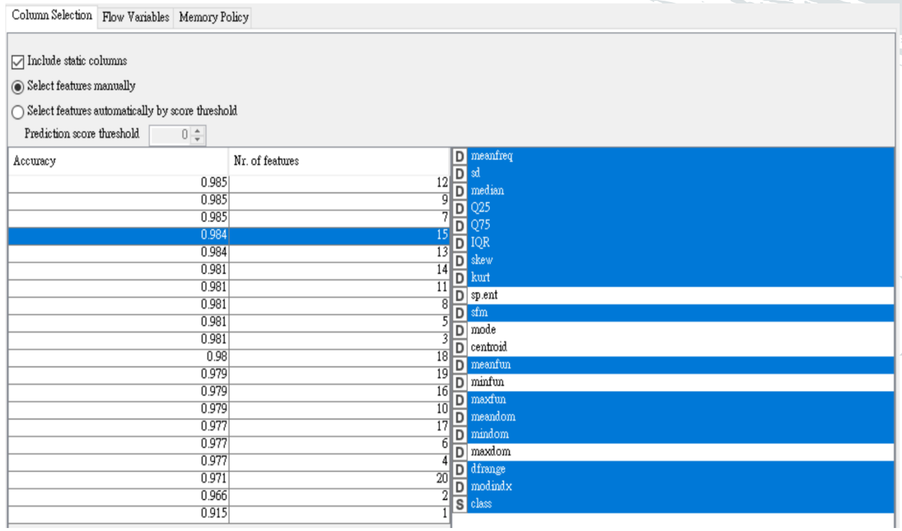
Feature selection

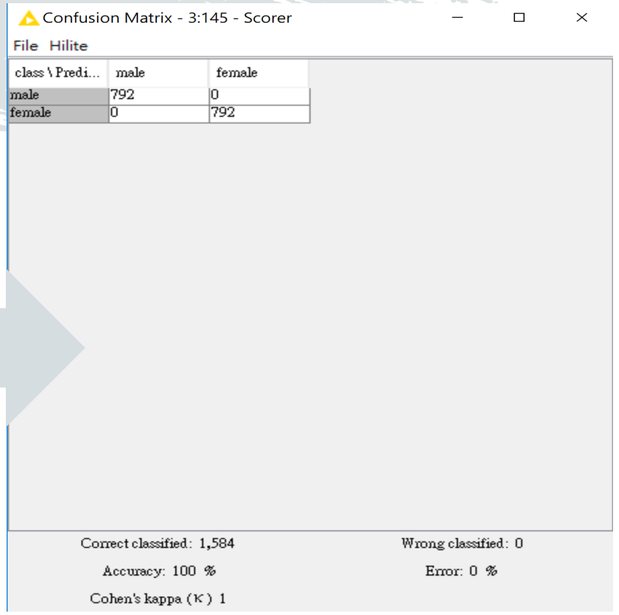


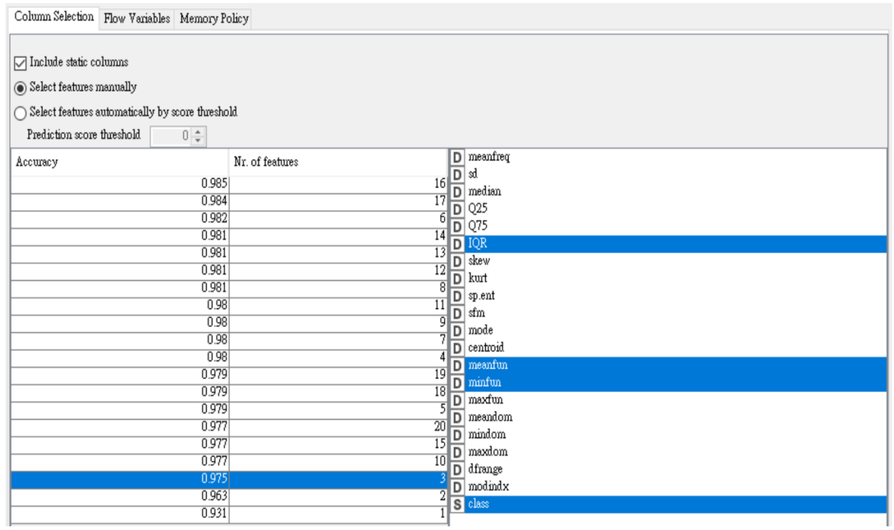
Feature selection + PCA

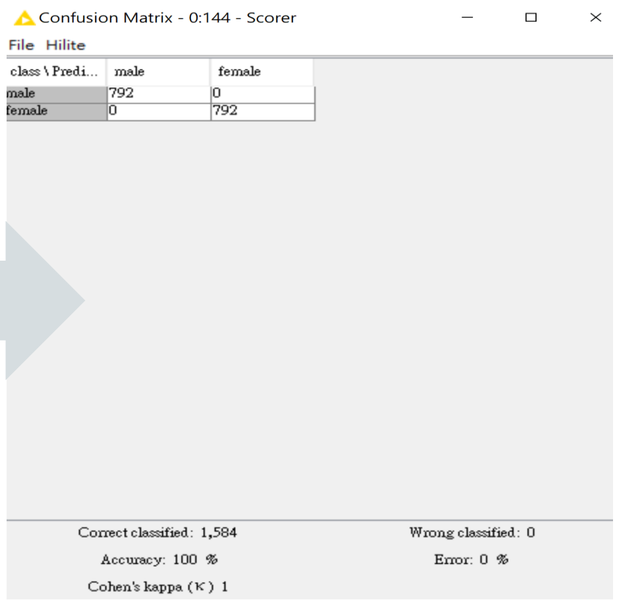


Feature selection for Random Forest





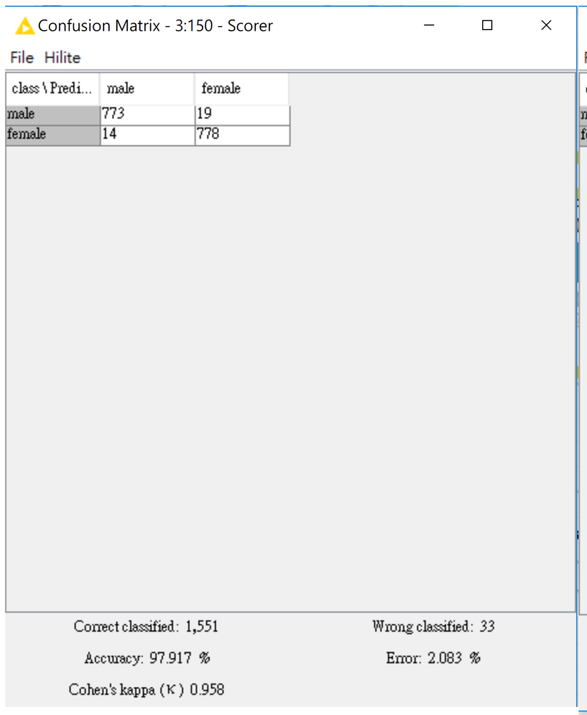




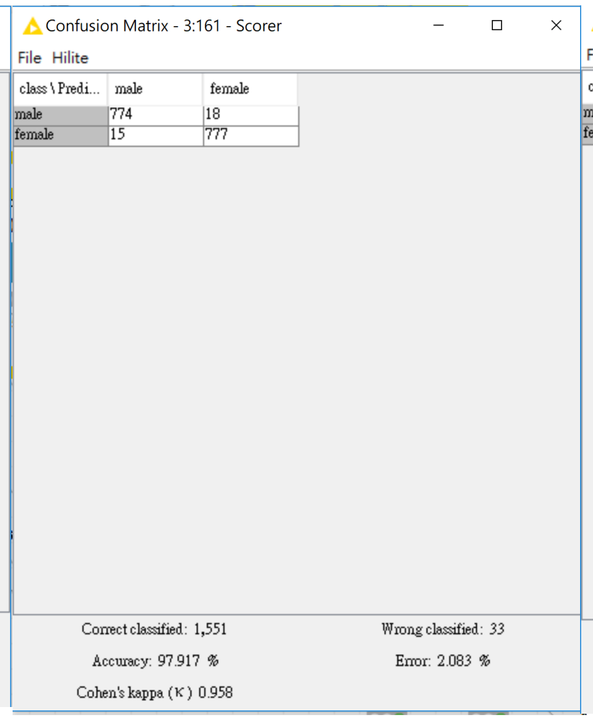
★在Random Forest 中使用Feature Selection以及Feature Selection +PCA，不論有幾個Feature，都不會影響到最後算出來的精準度，結果都為100%

Random Forest Confusion matrix

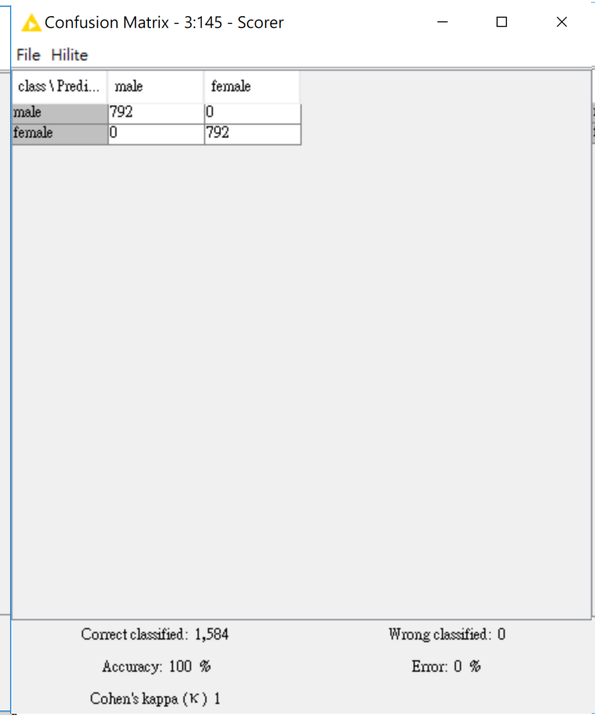
None



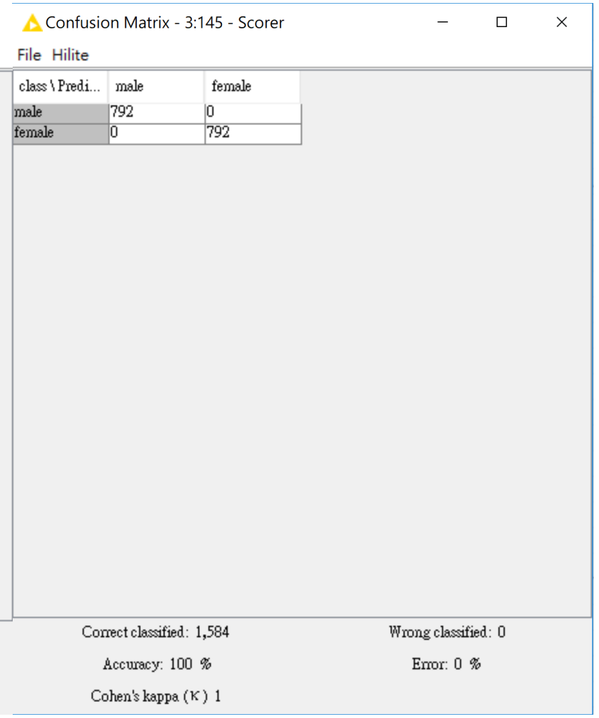
PCA



Feature selection



Feature selection + PCA



Result and Discussion

- 不管是透過PCA,feature selection，其準確度相差不遠
- 不管是透過SVM,random forest,logistic regression，其準確度相差不遠，若使用random forest精準度可以高達100%
- 在Random Forest 中使用Feature Selection或是Feature Selection +PCA，不論有幾個Feature，都不會影響到最後算出來的精準度
- 對於男性如果發出較高頻率的聲音(假音)，有可能造成誤判
- Partitioning多少與結果並無影響，我們有嘗試過使用不同的partitioning，重新execute了10多次，進行分析後與relative 50%的準確度相差不遠
- 本次使用的dataset僅僅是世界77億人口中的一小部分，如果可以加大資料集，那訓練結果也會更好，但也只是接近100%，仍還是需要其他的features來準確辨別性別

SECTION 5

Reflection

Reflection

https://github.com/king87515/knime_voice_project

- 透過這次project，對於KNIME的操作更熟悉
  - Ex: CSV reader,PCA,color/shape manager,plot,feature selection,etc.
- Kaggle裡面的資料集很多，也有很多問題在討論區得到解答，裡面的reference也很用幫助
- 對一些machine learning的演算法更加了解，以及如何使用與操作
- 原本對於一些統計方面的東西不夠了解，藉由這次機會了解了更多
- 一直對於資料處理、分析這塊很有興趣，是一門學問很大、很淵博、實用的工具
- 了解到一些聲學性質與名詞

References

- https://www.kaggle.com/primaryobjects/voicegender
- http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/
- https://github.com/primaryobjects/voice-gender
- http://notebookpage1005.blogspot.com/2018/03/random-forest.html

THANK YOU!