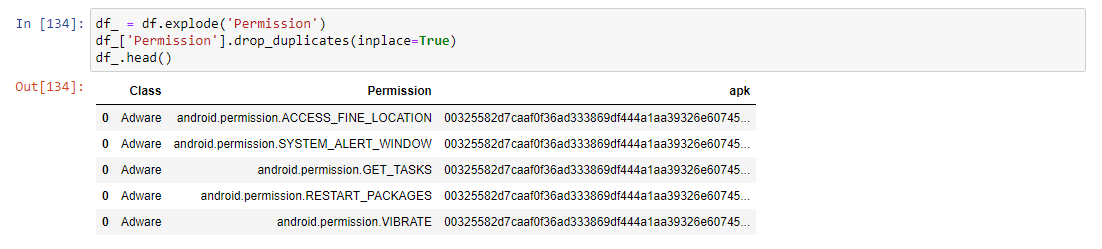
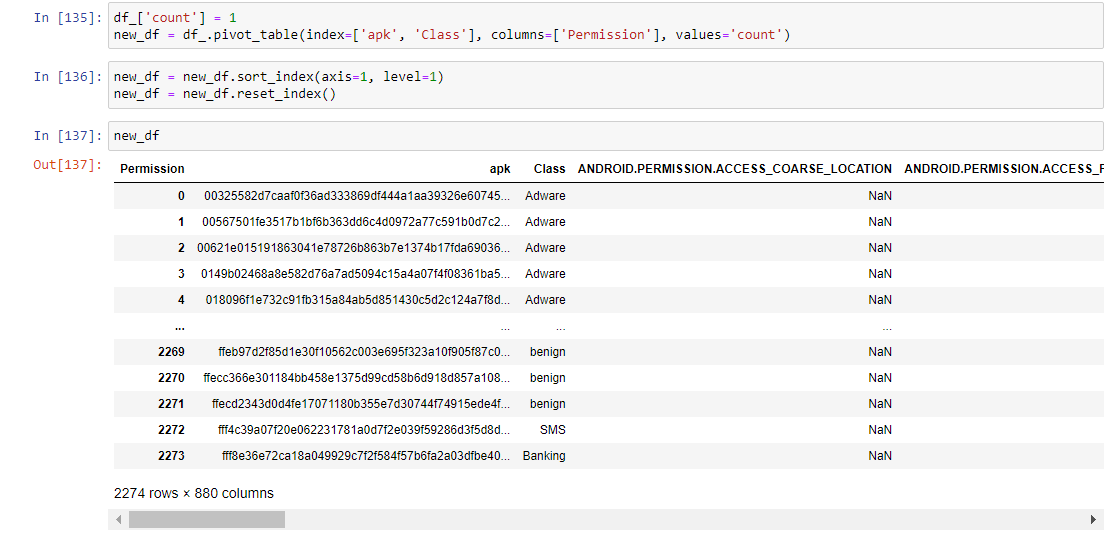
ETC\_HW8\_107403020 李泳輝

資料前處理

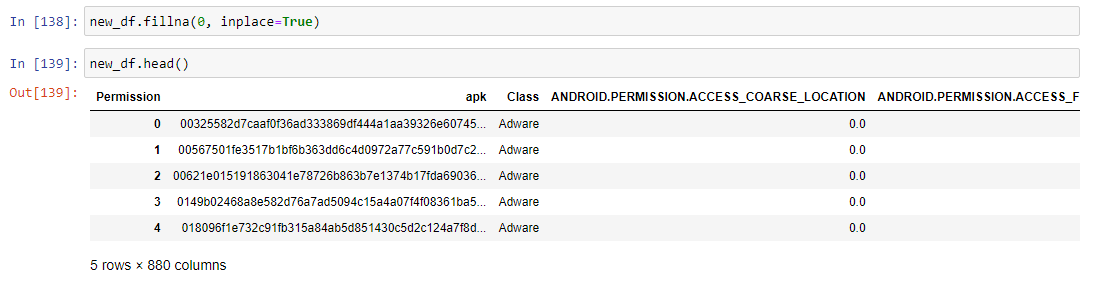
1. 使用explode將Permision的list展開



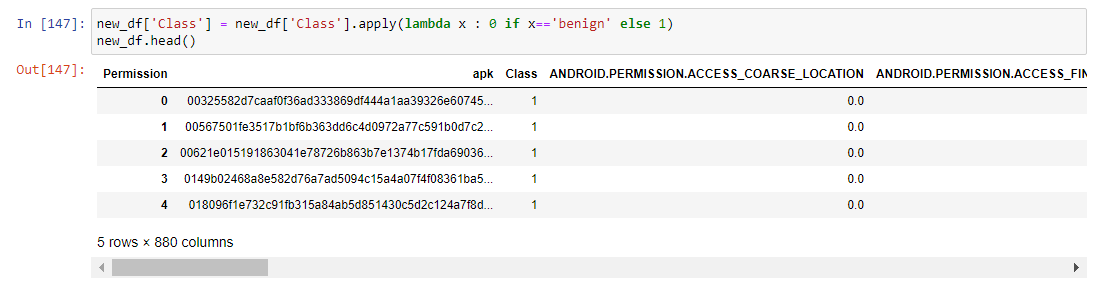
1. 根據apk和class將展開後的Permission重新分配到相應的欄位。



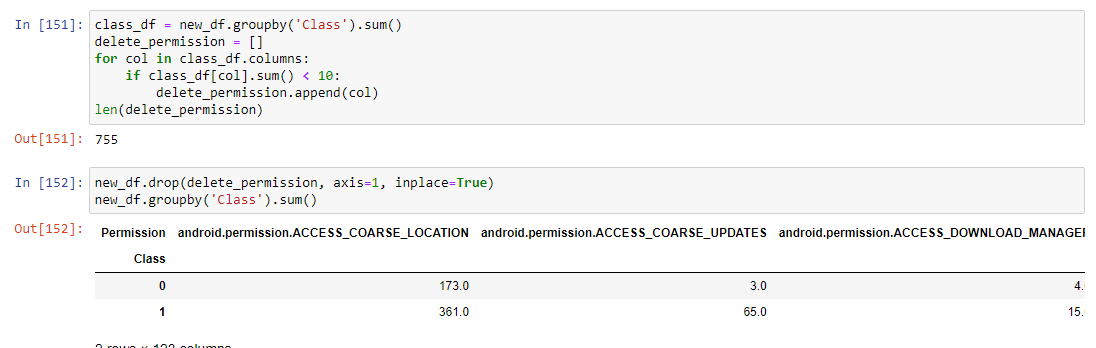
1. 填補空值



1. 將Class改成數值，benign為0，其餘為1



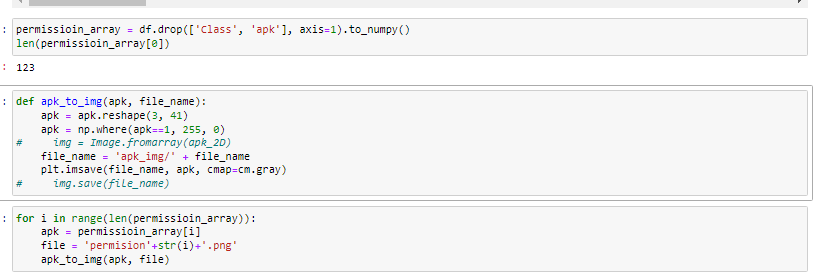
1. 因為欄位數太多，留下permission出現次數超過5的欄位



1. 將處理好的資料儲存



1. 將apk 轉圖片



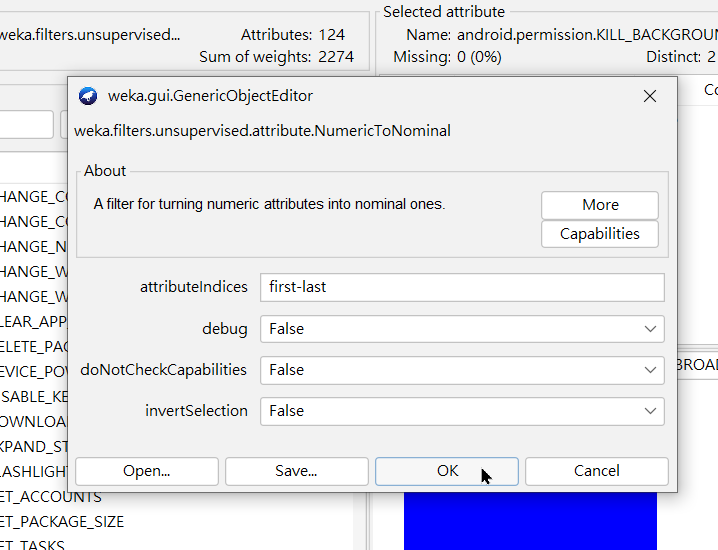
1. 將處理好的資料儲存



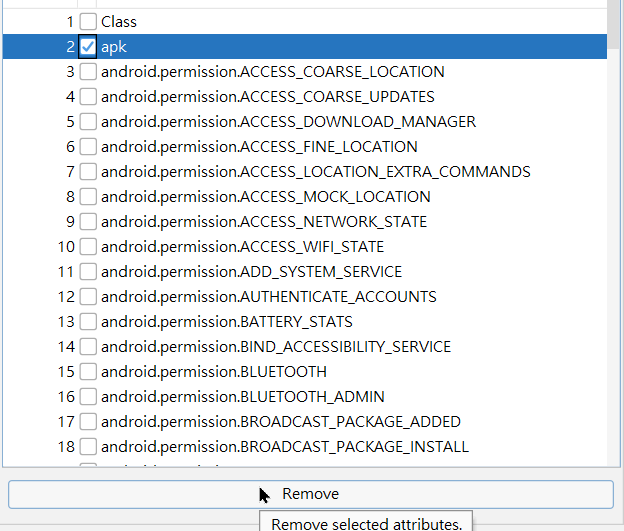
訓練模型

1. 第一類資料，用來訓練RandomForest、MLP

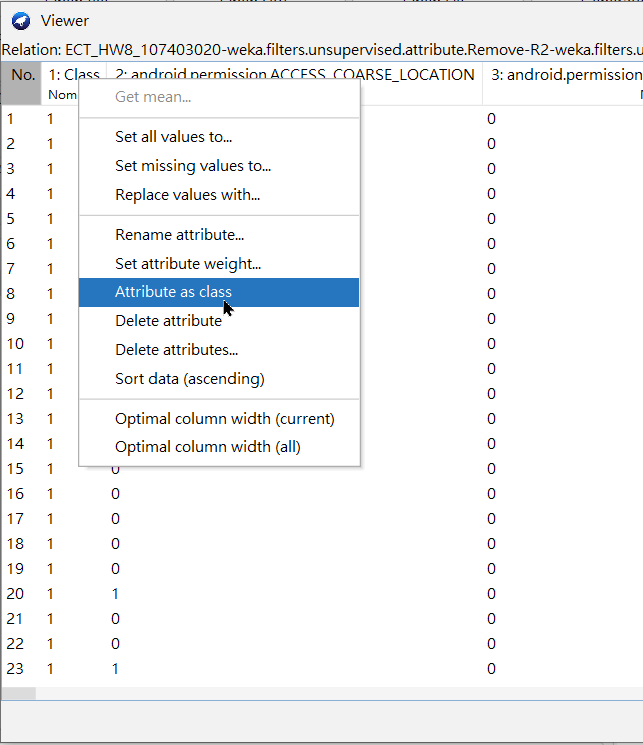
numeric\_to\_nominal



去除apk欄位

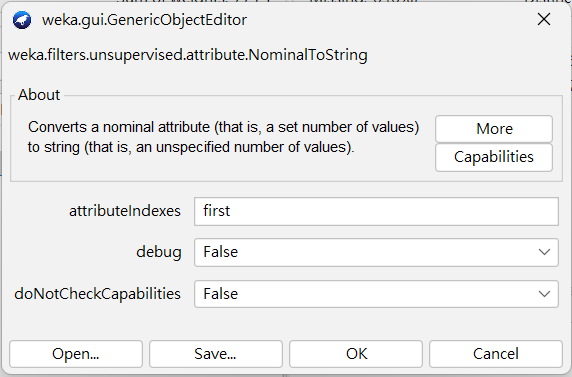


設定Class

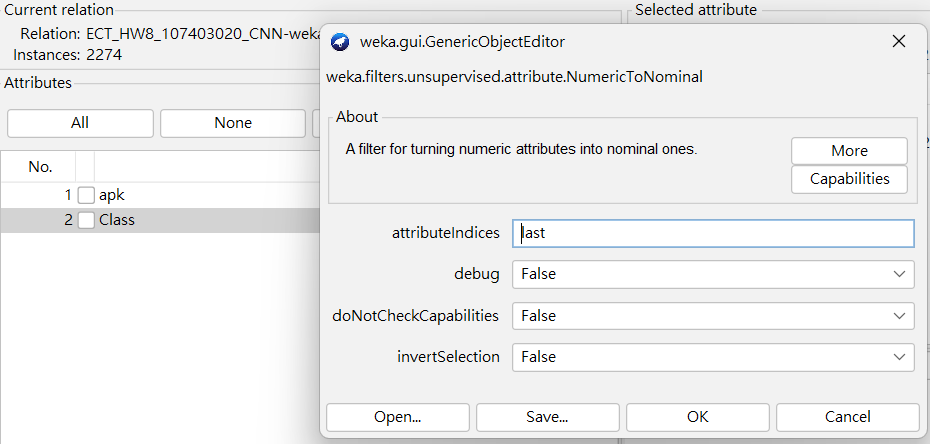


1. 第二類資料 apk, Class (apk放的是圖片位置)，用來訓練CNN

將apk轉string

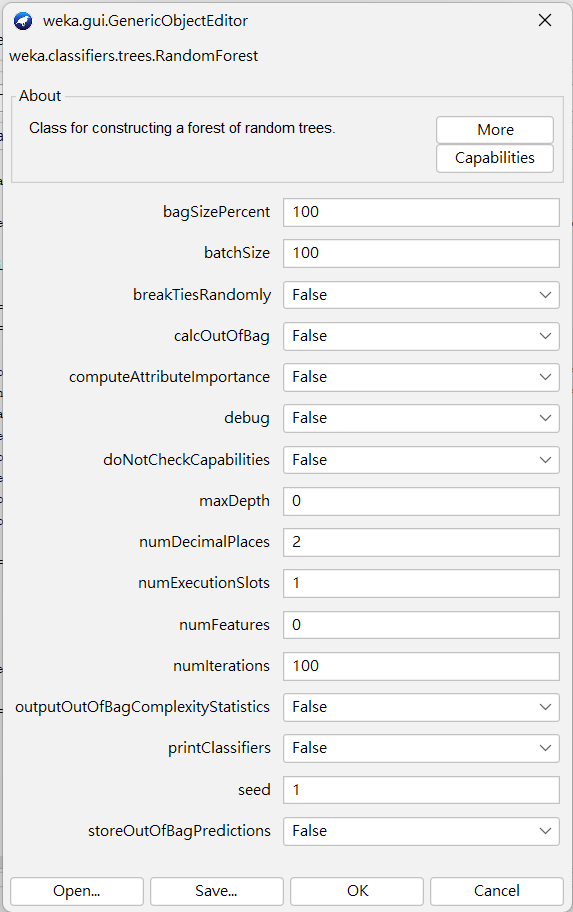


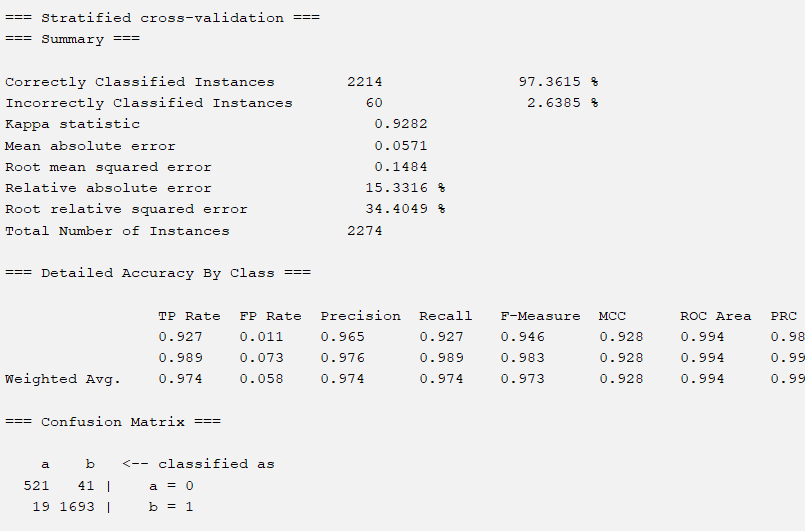
numeric\_to\_nominal



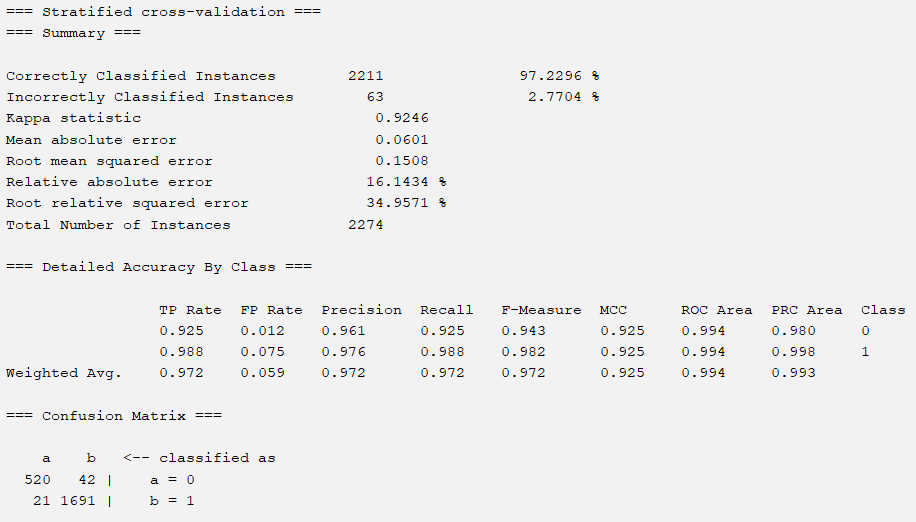
1 . RandomForest

超參數 採用系統預設值。使用10 Cross-validation來進行訓練。

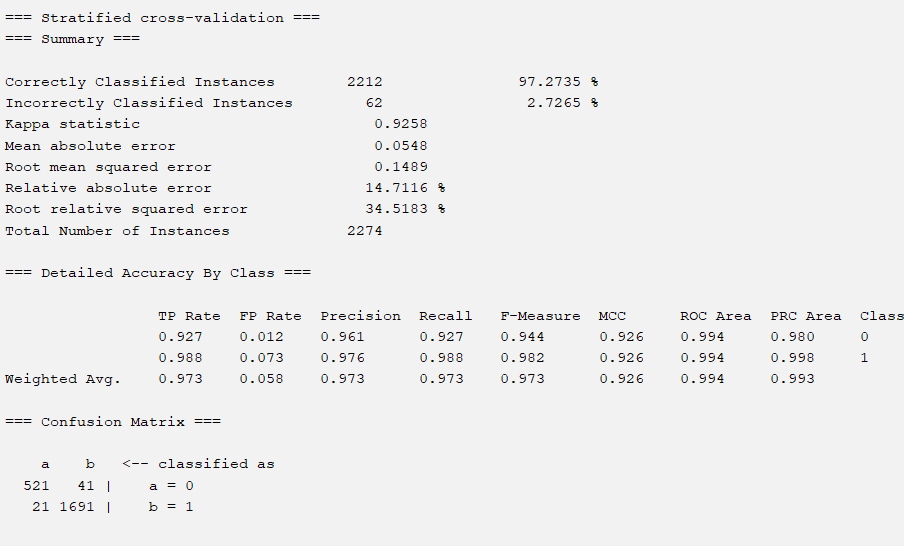




2. 將numFeatures設為 5，numFeatures是設定有多少的features用於random selection。



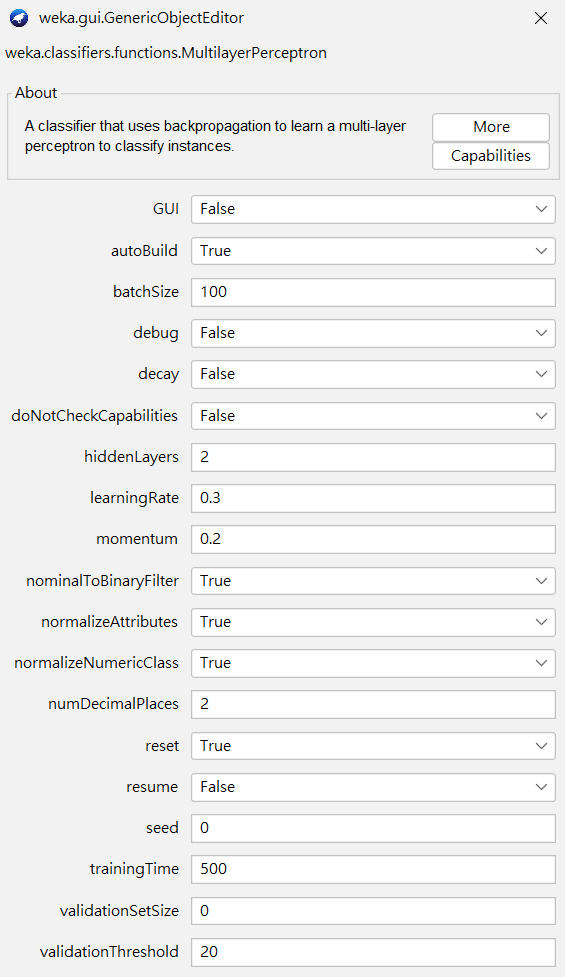
1. 將numFeatures設為 11

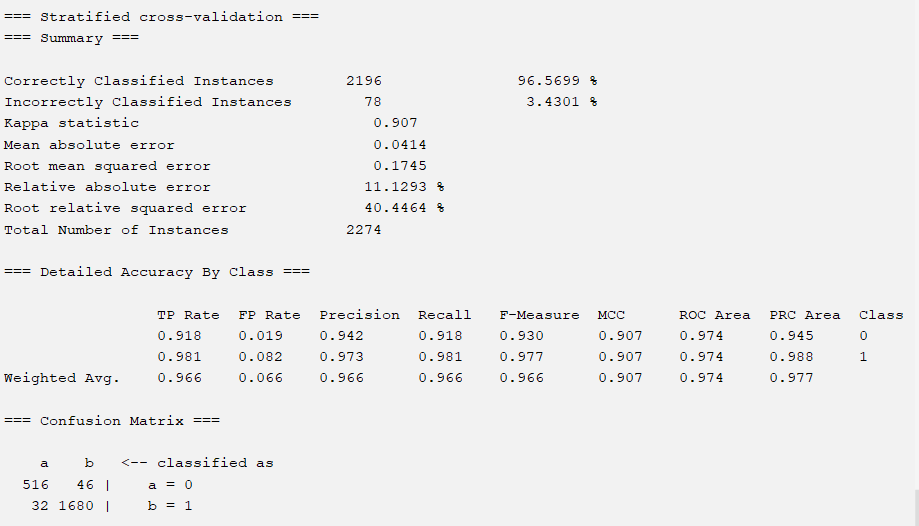


1. MLP

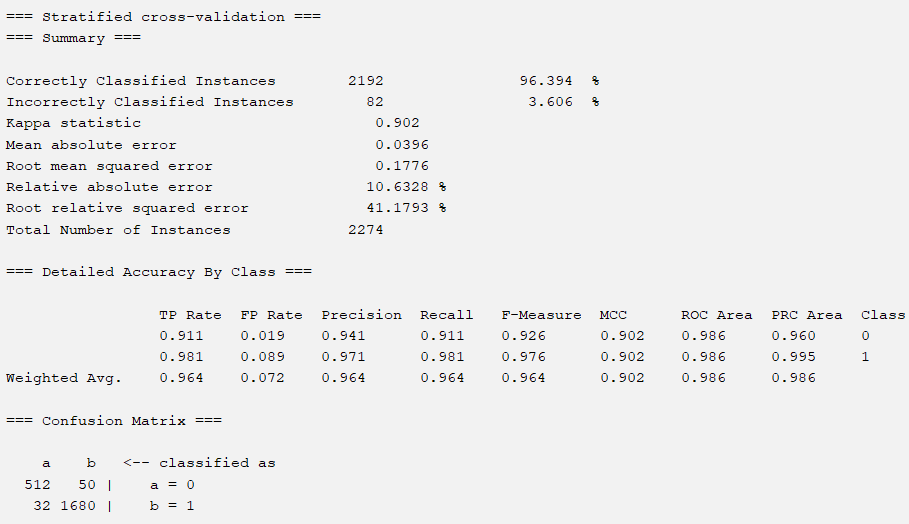
使用10 Cross-validation來進行訓練。

超參數 : 因為預設的模型hidden layer層數過多，訓練時間過久。將hidden layer改成2，也就是指有一層hidden layer，node個數為2。

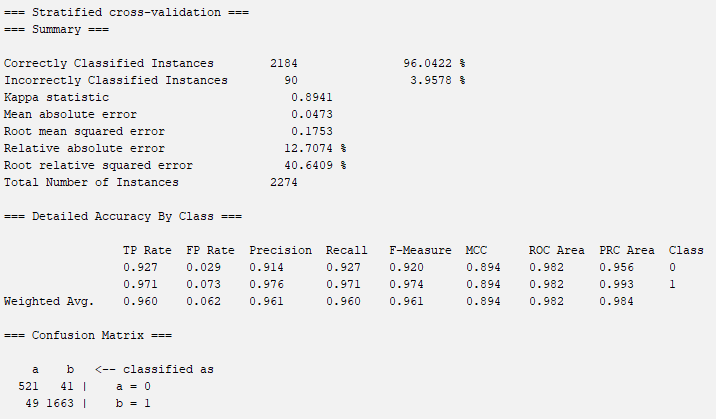




超參數 : hidden layer設為10，有一層hidden layer，node個數為10



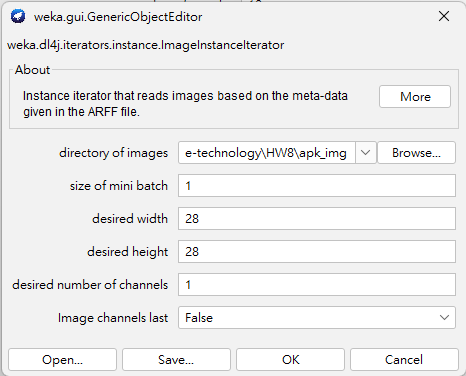
超參數 : hidden layer設為10, 2，代表有兩層hidden layer，第一層node個數為10，第二層為2



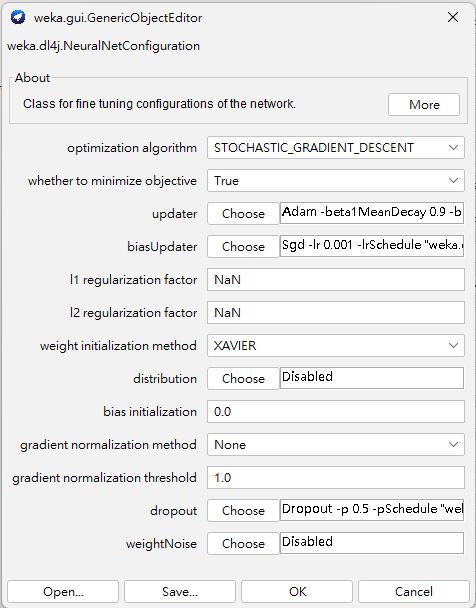
1. CNN

最大的問題是電腦效能，太複雜的模型電腦跑不動。CNN考慮到時間因素，因此使用percentage split來將資料拆成訓練集、測試集，其中訓練集佔66%。

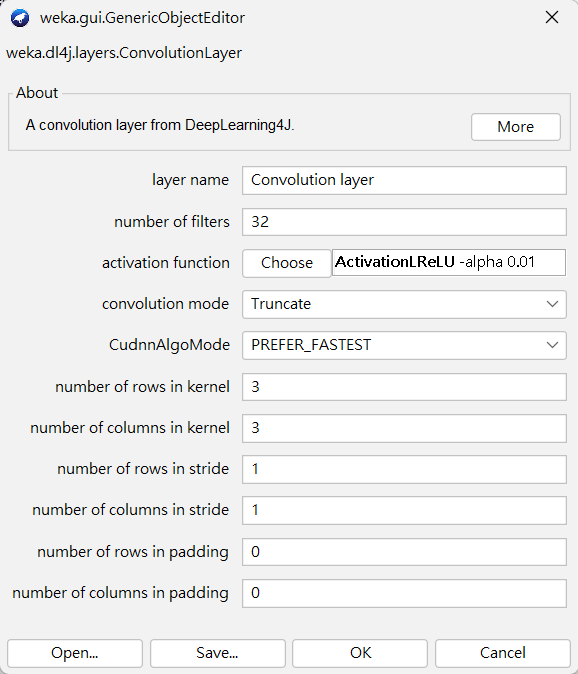
instance iterator選擇ImageInstanceIterator，讓weka會根據apk中放的檔名，到選定的資料夾讀取圖片。



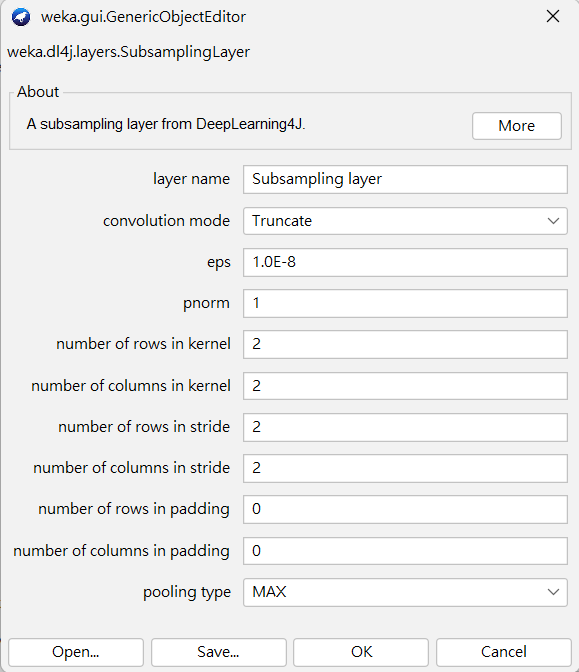
設定dropout 訓練時每次會隨機丟棄0.5%的神經節點。



* convolution layer的nOut是filiter的個數，activation設定成ReLU。在此的捲積層kernel均是3X3，stride均設為1。



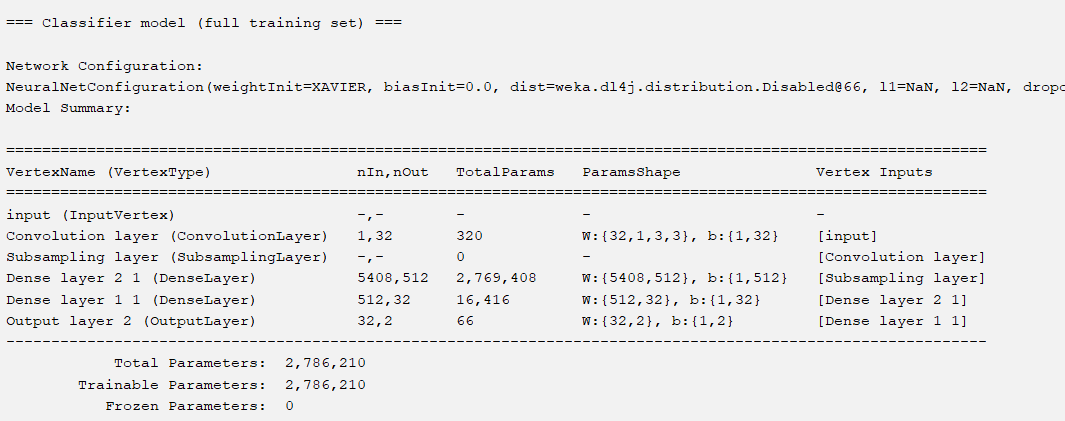
* subsampling設定為Max，會將filiter得到的feature map做壓縮。kernel均設為2X2，stride為2。

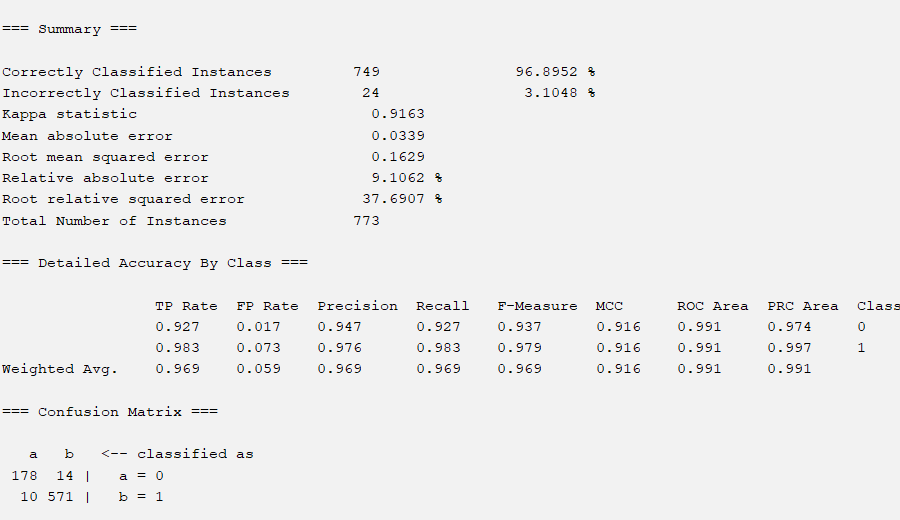


* 第一層的Dense會把上一層所有的feature map展開，排成一列。activation設定成ReLU。
* 最後一層的output layer的Loss function設定成LossNegativeLogLikelihood。

模型一:

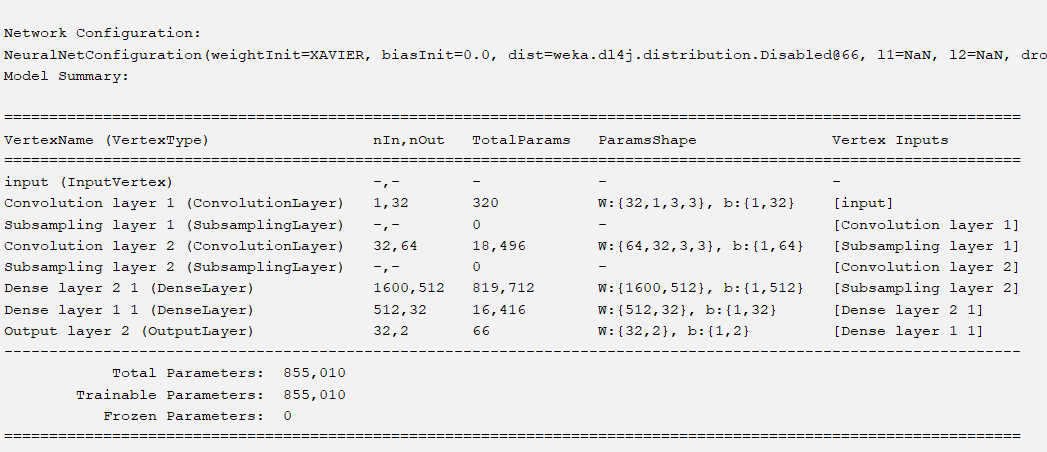
一層convlution:

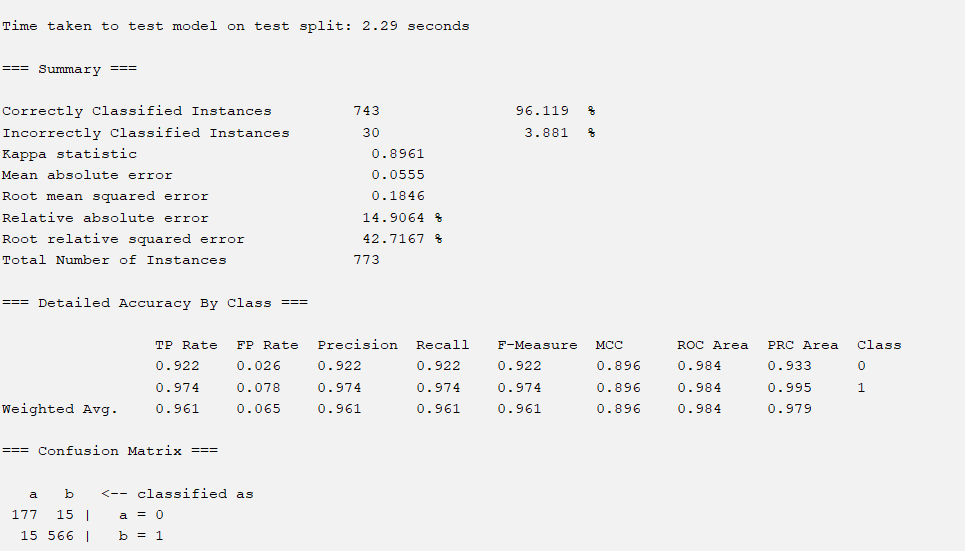




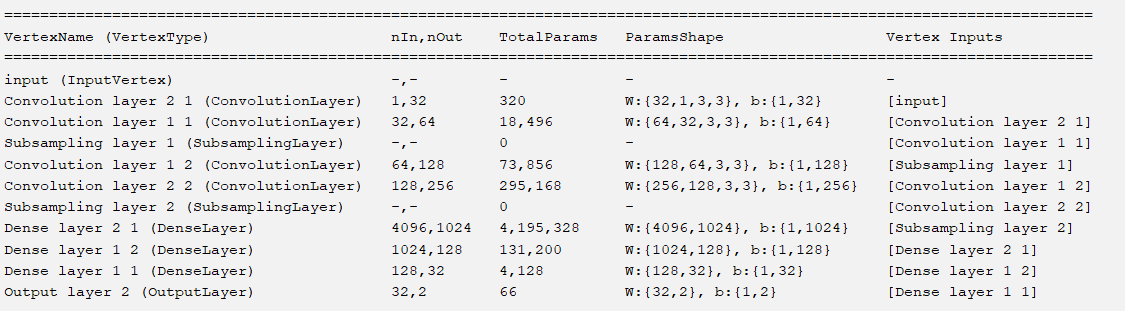
模型二:

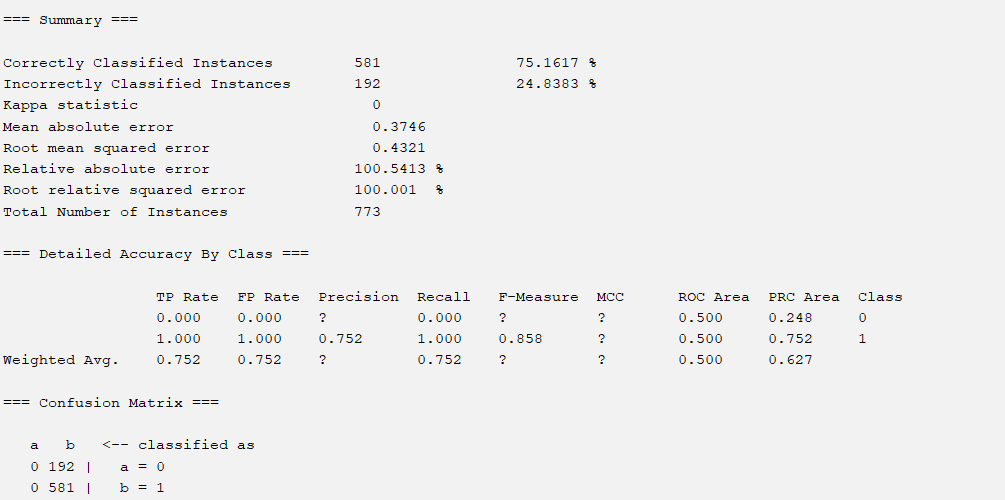
兩層 convolution





模型三:





評估報告

 分析結果，需有圖表統整同一個模型不同超參數的表現以及各模型的表現

1. 模型比較

|  |  |  |  |
| --- | --- | --- | --- |
|  | RandomForest | MLP | CNN |
| accuracy | 97.3615% | 96.5699% | 96.8952 |

1. Random Forest

|  |  |
| --- | --- |
| numberFeatures | accuracy |
| 0 | 97.2615% |
| 5 | 97.2296% |
| 11 | 97.2735% |

1. MLP

|  |  |
| --- | --- |
| hidden Layer | accuracy |
| 2 | 96.5699% |
| 10 | 96.394% |
| 10, 2 | 96.0422% |

1. CNN

|  |  |
| --- | --- |
|  | accuracy |
| 模型一 | 96.8952% |
| 模型二 | 96.119% |
| 模型三 | 75.1617% |

* 使用的方法的優缺點

1. RandomForest好處是執行快速且準確率高，但可讀性差。
2. MLP則是簡單好懂，雖然hidden layer沒設很高，但執行的效果還算不錯，但hidden layer設太高，電腦執行時間久。
3. CNN準確率不錯、參數共享，但執行的時間久，且容易overfitting。

* 是否有改進之處

前處理的部分，可以嘗試刪除多個欄位，嘗試找到最適合的組合。

雖然這次實驗CNN模型複雜，對準確度沒有太大的幫助，反而造成準確度下降的問題。推測這可能跟資料量和overfitting有關。因此推測可以使用Dropout或regulization等技術，來避免overfitting。

* 最後會選擇哪一個模型

最後選擇RandomForest。考慮到電腦的執行時間，以及準確度，RandomForest訓練出的結果明顯優於MLP和CNN。