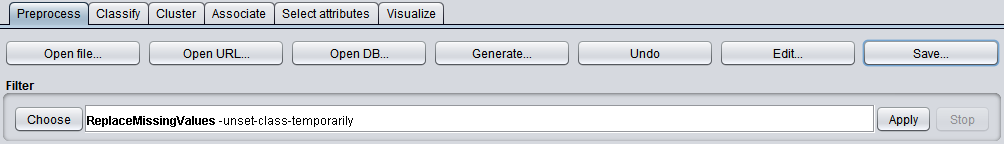
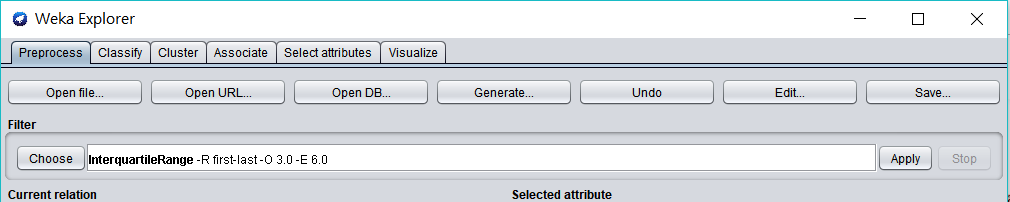
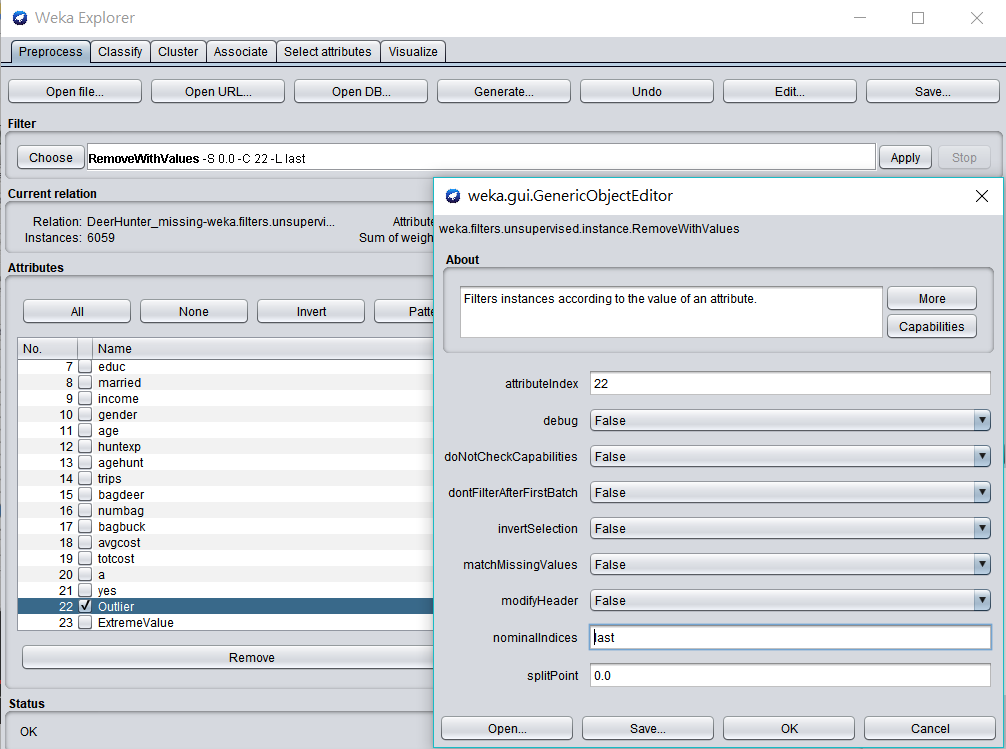
2019 ECT 作業八

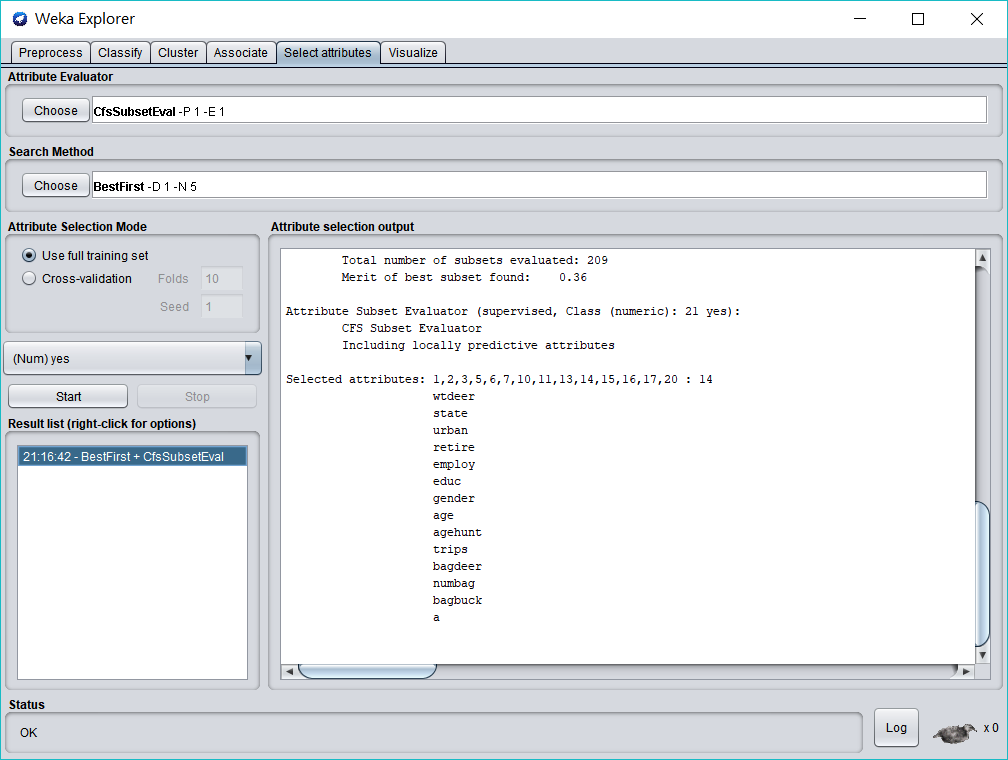
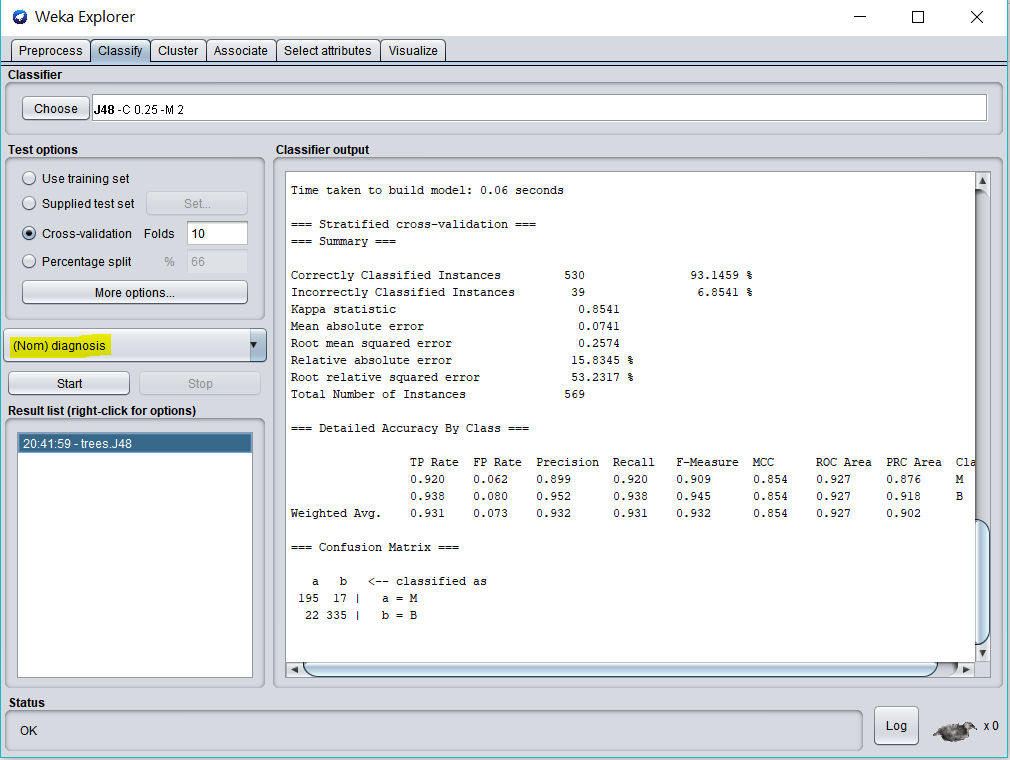
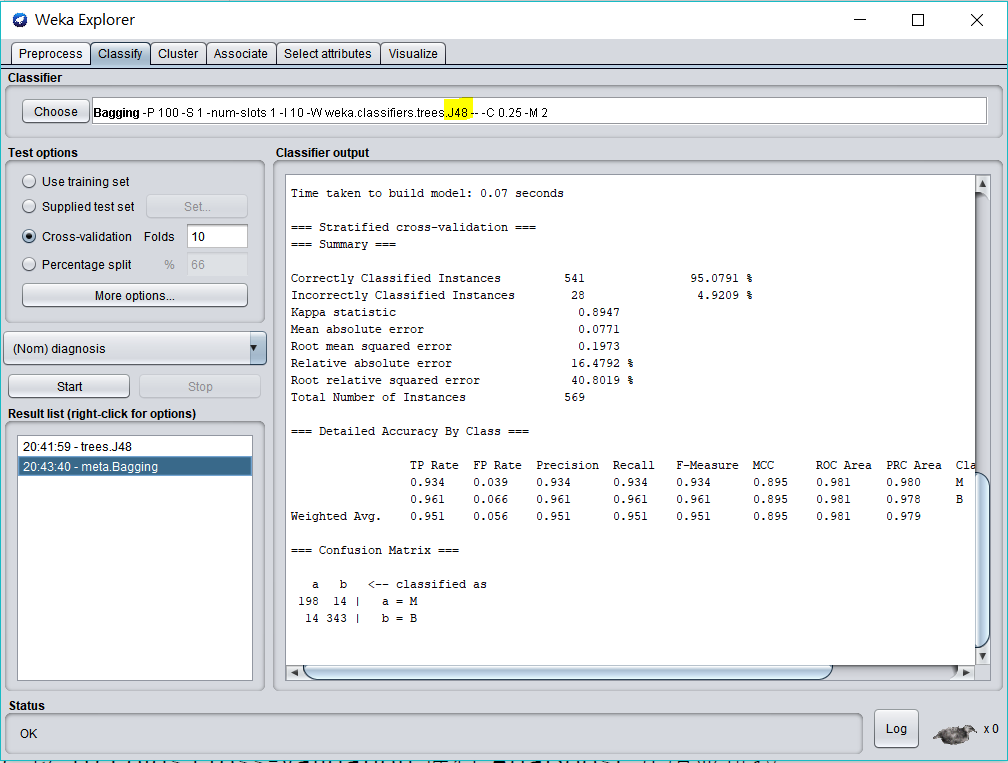
* 1. **一、 利⽤Weka 對DeerHunter.arff 進⾏前處理，依序完成以下步驟及問題：**
  2. **(a). Replace Missing Value，需列出補上的值為何(10%)**
  3. 

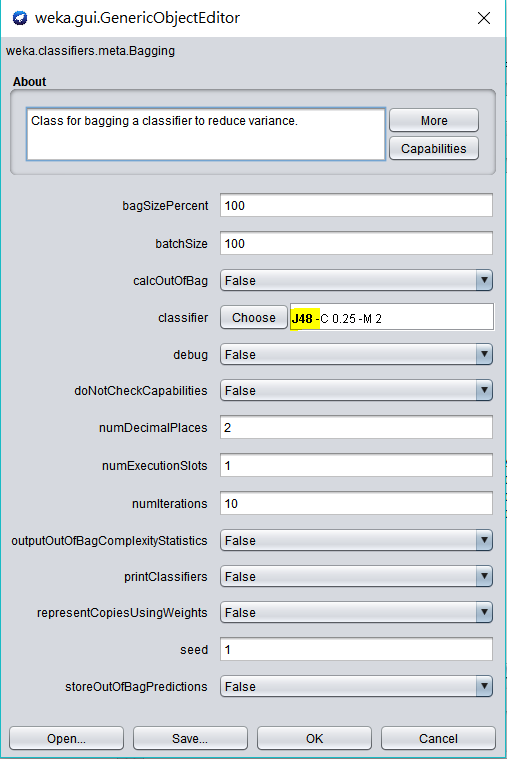
1. 在Preprocess的Filter選取「weka/filters/unsupervised/attribute/ReplaceMissingValues」
2. 點選「Apply」，可將MissingValue值變更為MeanValue
   1. **(b). Outlier Detection& Remove (10%)**
   2. 
3. 在Preprocess的Filter選取「weka/filters/unsupervised/attribute/InterquartileRange」
4. 點選「Apply」，會多出兩個Attribute(Outlier與ExtremeValue)

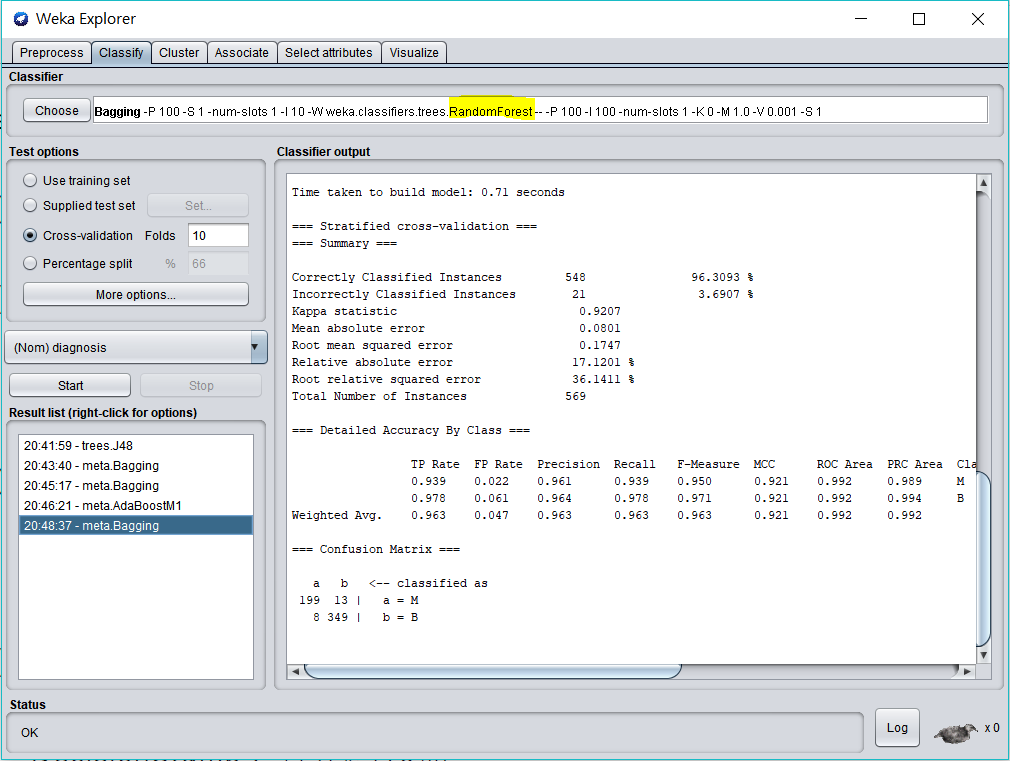


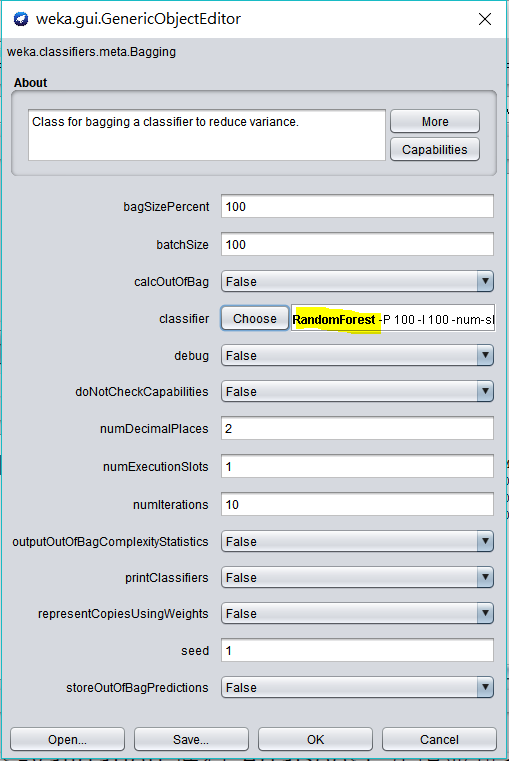
1. 在Preprocess的Filter選取「weka/filters/unsupervised/instance/RemoveWithValues」
2. 在參數調整內，將「attributeIndex」設為「22」(Outlier的index編號)，「nominalIndices」設為「last」(Outlier為yes的instance)，並點選「OK」。

(此步驟的目的在於判別若為Outlier的instance，則刪除)

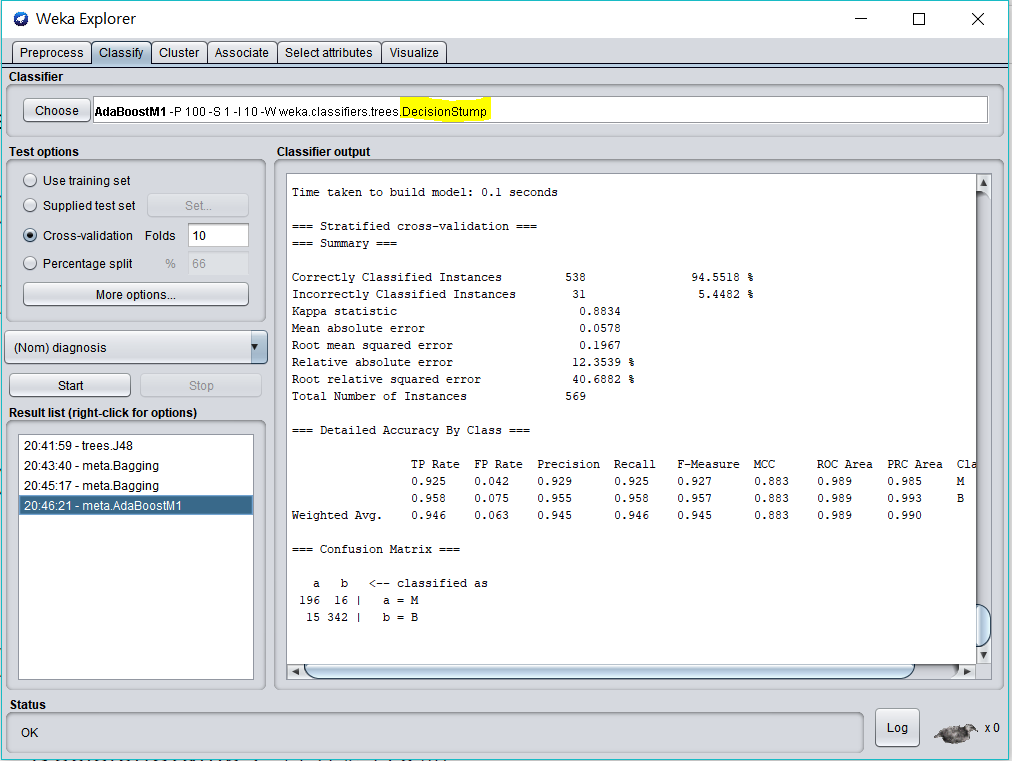
1. 點選「Apply」，完成。
   1. **(c). Attribute Selection，請篩選出適合的屬性(10%)**
   2. 
2. 在Select attributes的Attribute Evaluator選取「weka/attributeSelection/CfsSubsetEval」、在Search Method選取「weka/attributeSelection/BestFirst」
3. 將屬性設為「(Num)yes」
4. 點選「Start」
5. 可由「Attribute selection output」得適合的屬性為「wtdeer」、「state」、「urban」、「retire」、「employ」、「educ」、「gender」、「age」、「agehunt」、「trips」、「bagdeer」、「numbag」、「bagbuck」、「a」，共14個
   1. **二、 請用weka對BreastCancer.csv對目標diagnosis進行Ensemble learning並與未使用的結果進行比較(請列出重要過程及適當說明)：**
   2. **(a). 以10 Folds cross-validation進行J48分類(5%)**
   3. 
6. 在Classifier中選取「J48」
7. Test options 使用 Cross-validation，並設定Folds為10
8. 選取屬性「(Nom)diagnosis」
9. 點選「Start」
   1. **(b). 以10 Folds cross-validation進行Bagging分類並選擇J48 classifier進行分類(10%)**
   2. 
10. 在Classifier中選取「Bagging」

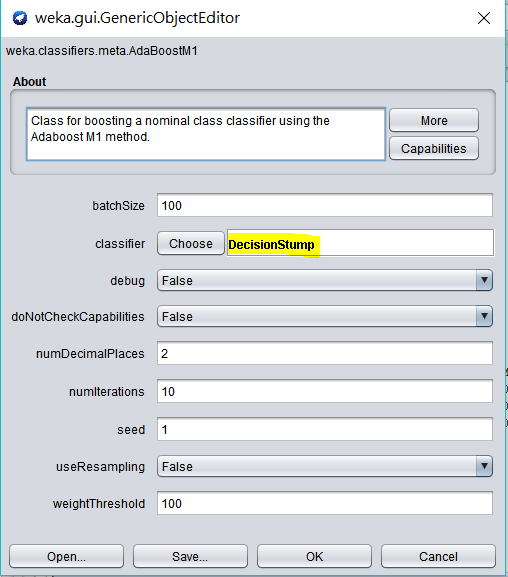


1. 在參數調整的地方，將classifier設為「J48」
2. Test options 使用 Cross-validation，並設定Folds為10
3. 選取屬性「(Nom)diagnosis」
4. 點選「Start」
   1. **(c). 以10 Folds cross-validation進行Bagging分類並選擇Randomforest進行分類(10%)**
   2. 



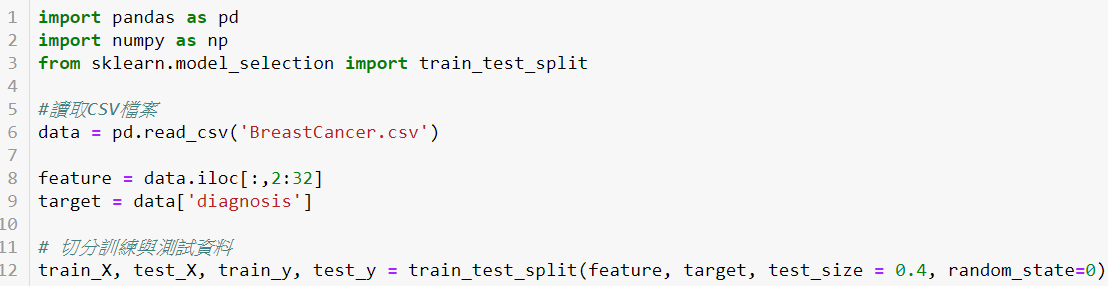
1. 在參數調整的地方，將classifier設為「RandomForest」
2. Test options 使用 Cross-validation，並設定Folds為10
3. 選取屬性「(Nom)diagnosis」
4. 點選「Start」
   1. **(d). 以10 Folds cross-validation進行AdaBoost分類並選擇DecisionStump進行分類(10%)**





1. 在Classifier中選取「AdaBoost」
2. 在參數調整的地方，將classifier設為「DecisionStump」
3. Test options 使用 Cross-validation，並設定Folds為10
4. 選取屬性「(Nom)diagnosis」
5. 點選「Start」

**三、 請用python對BreastCancer.csv對目標diagnosis進行Ensemble learning並與未使用的結果進行比較(請列出重要過程及適當說明)：**

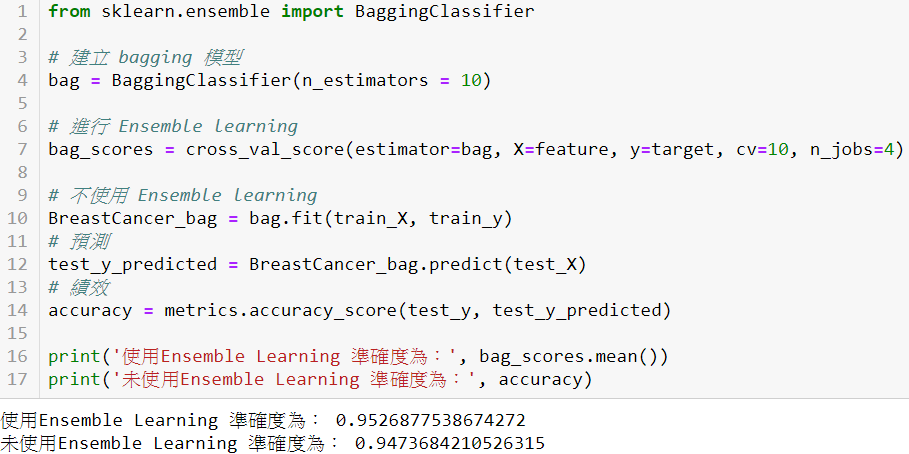


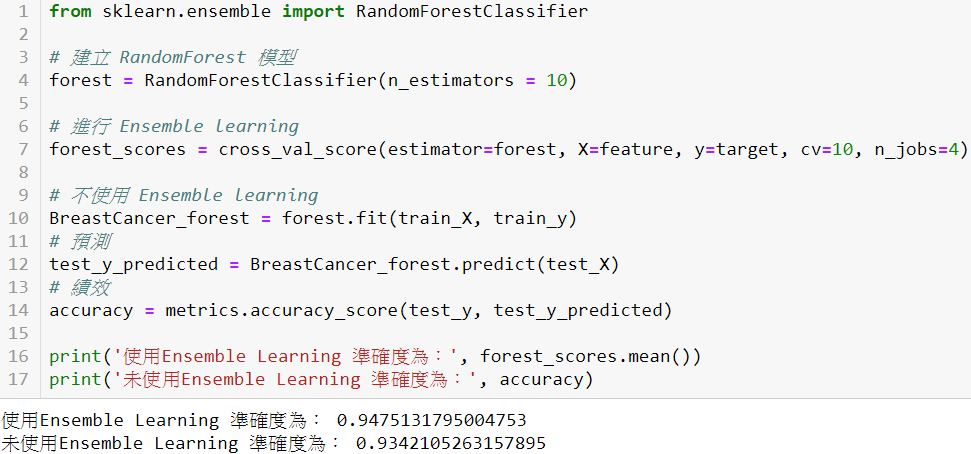
1. 首先將csv檔匯入
2. 將資料切割為feature與target
3. 切分為訓練資料與測試資料(用於未使用Ensemble learning的部分)

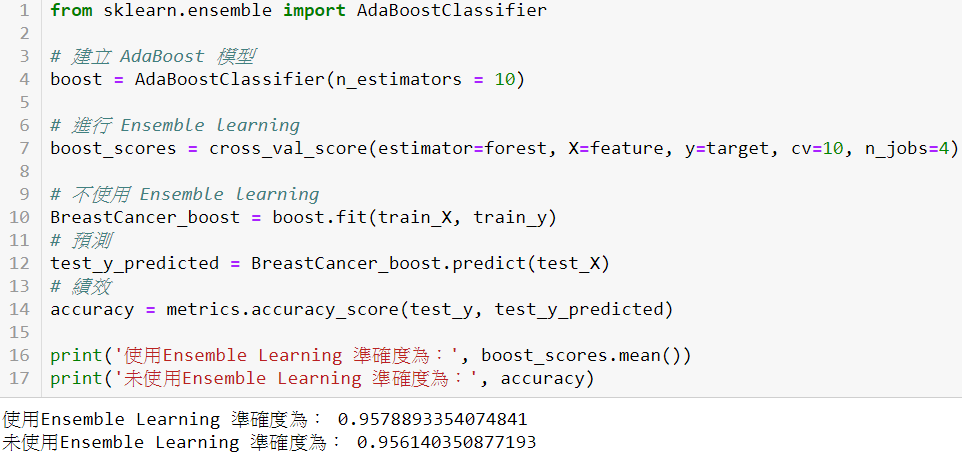
**(a). 以10 Folds cross-validation進行DecisionTreeClassifier分類(5%)**



**(b). 以10 Folds cross-validation進行BaggingClassifier，n\_estimators=10分類(10%)**



1. **(c). 以10 Folds cross-validation進行RandomForestClassifier分類(10%)**
2. 
3. **(d). 以10 Folds cross-validation進行AdaBoost,n\_estimators=10分類(10%)**



**結論：使用Ensemble learning通常會有較佳的準確度，通過組合多個模型，集成學習有助於提高機器學習效果。與單個模型相比，該方法允許產生更好的預測性能。**