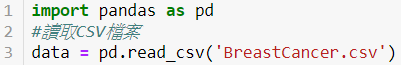
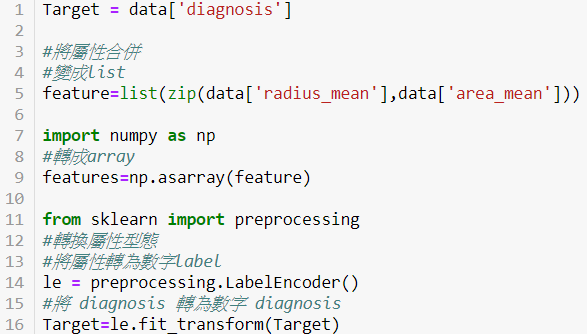
**2019 ECT 作業六**

1. **請用 python 依照步驟對 BreastCancer.csv 進行 KNN 及 KMeans 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。(80%)**

首先，引入需要使用的套件與讀取csv檔。



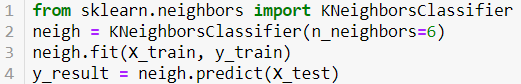
1. **將 radius\_mean 及 area\_mean 切為 feature，diagnosis 切為 target**



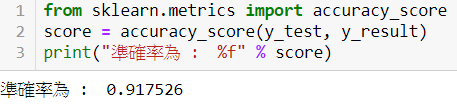
1. **切分資料及與訓練集，設 test\_size=0.34，random\_state=5**



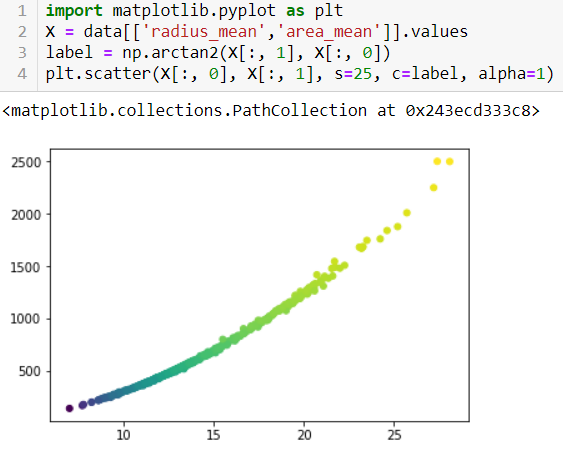
1. **用 KNeighborsClassifier 進行 KNN 分類，n\_neighbors 設為 6**



1. **用 metrics 算出此模型對於測試集的準確度**



1. **運用 matplotlib 中的 scatter 圖，x 軸設為 radius\_mean，y 軸設為area\_mean，c 設為 label，印出測試分類圖形。**



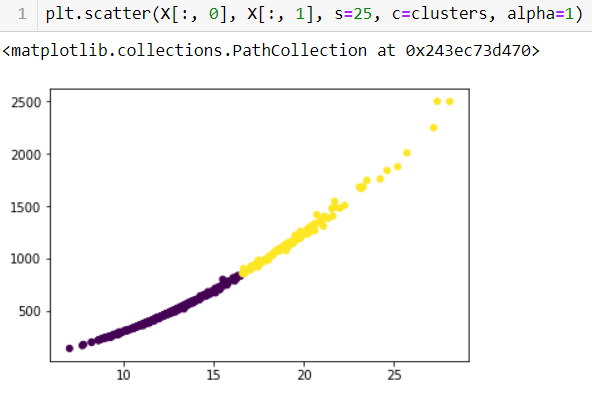
1. **用 cluster.Means 設 n\_clusters=2**



1. **用 fit\_predict 對切分的 feature 進行預測**



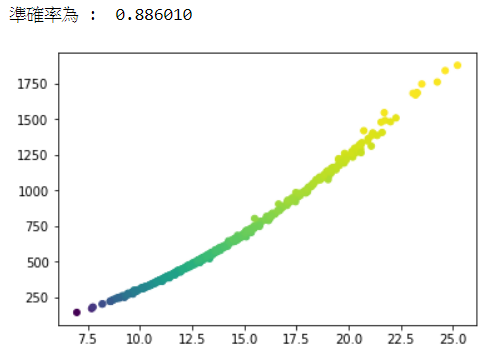
1. **運用 matplotlib 中的 scatter 圖，x 軸設為 radius\_mean，y 軸設為area\_mean，c 設為分群結果，印出分類圖形。**



1. **移除 area\_mean 中大於 2000 的資料**



1. **重複上述 a~e 的動作，同時回答問題：在此案例中移除與分布較遠的資料是否有達到更好的效果?**



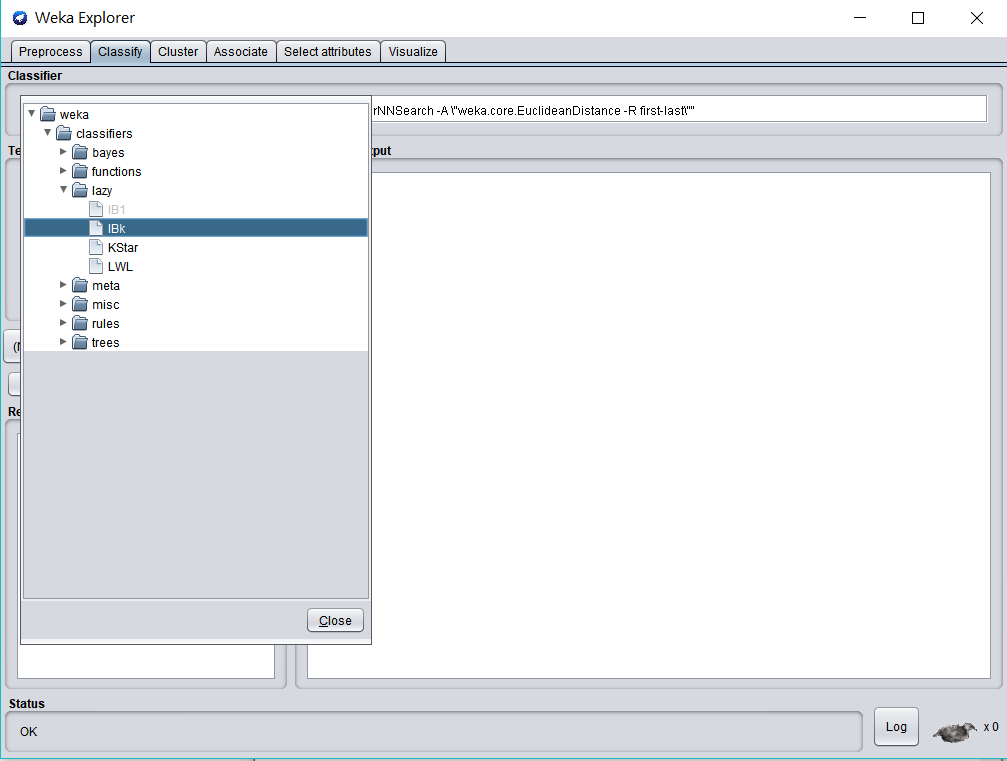
並沒有達到更好的效果。準確率從0.9175下降至0.886

1. **請用 weka 對 BreastCancer.csv，進行 IBK(knn) k 設為 6 及simplekMeans 進行分析，Percentage split 設為 66%，截圖並附上過程及準確率。(20%)**

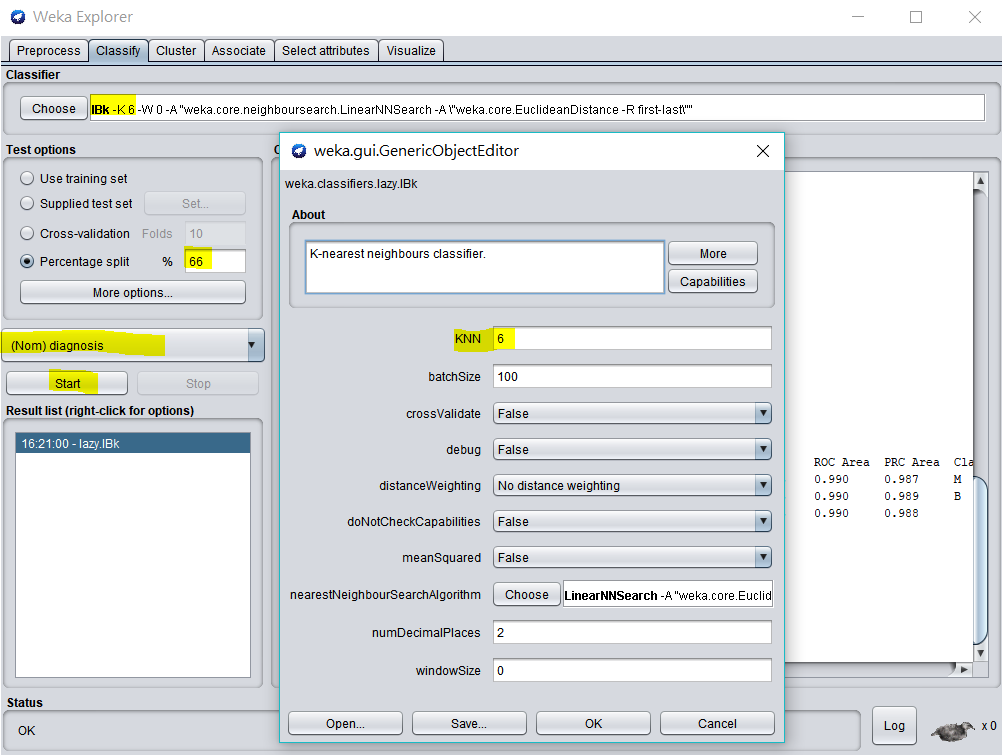
首先在Weka中開啟BreastCancer.csv。

**IBK(knn) 分析**

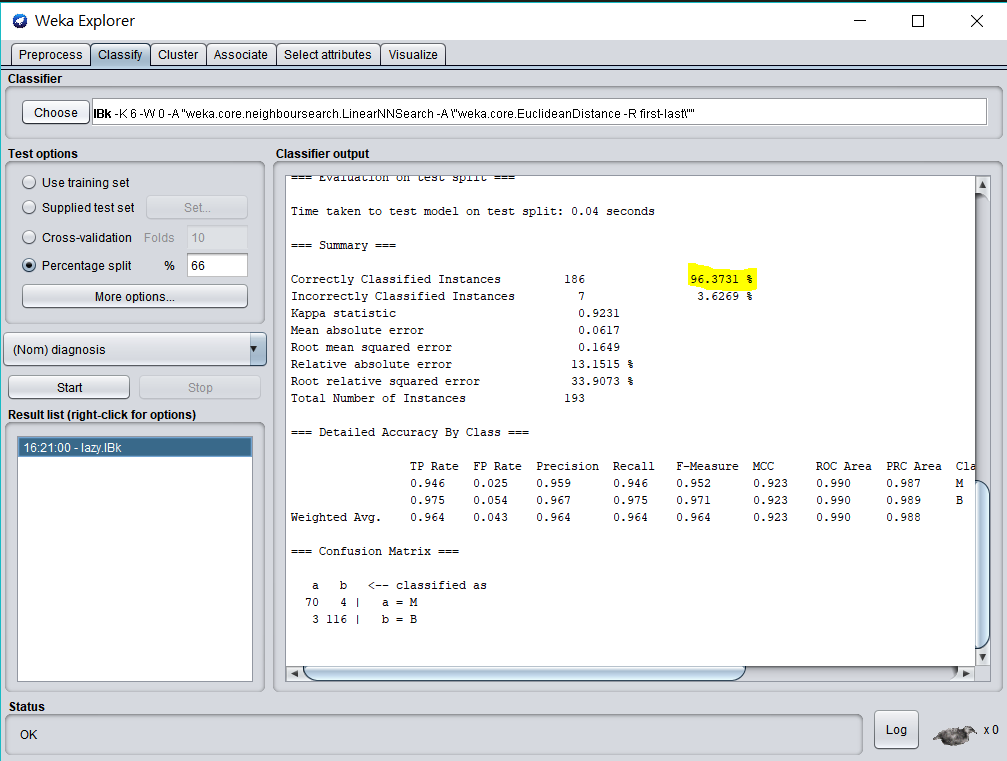
1. 在Classify面板中，在Classifier選擇「weka / classifiers / lazy/ IBK」。

****

1. 點選上方的參數設定，將「KNN」設置成6，在「Test options」中選取「Percentage split」，並設定為66%；選擇預測「(Nom)diagnosis」，並點選「Start」。

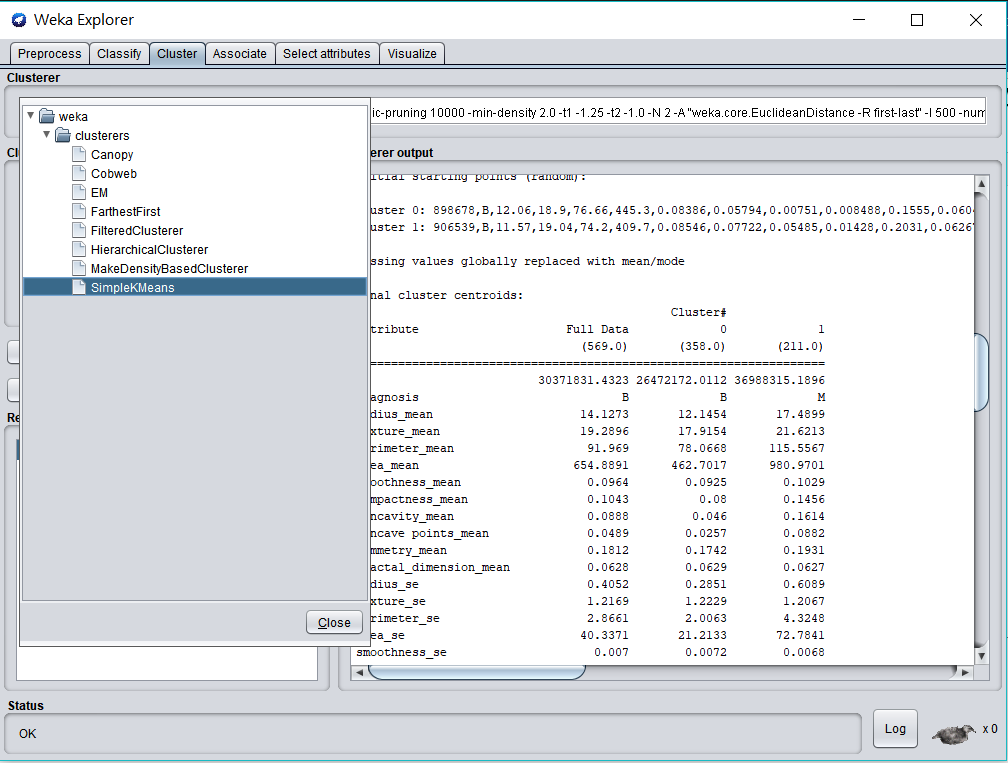


1. 可得準確度為 96.3731 %



**simplekMeans 分析**

1. 在Cluster面板中，在Cluster選擇「weka / clusters/ **simplekMeans**」。



1. 在「Test options」中選取「Percentage split」，並設定為66%，並點選「Start」。

可得Clustered Instance，有40%的instance被歸類到0，60%的instance被歸類到1。

