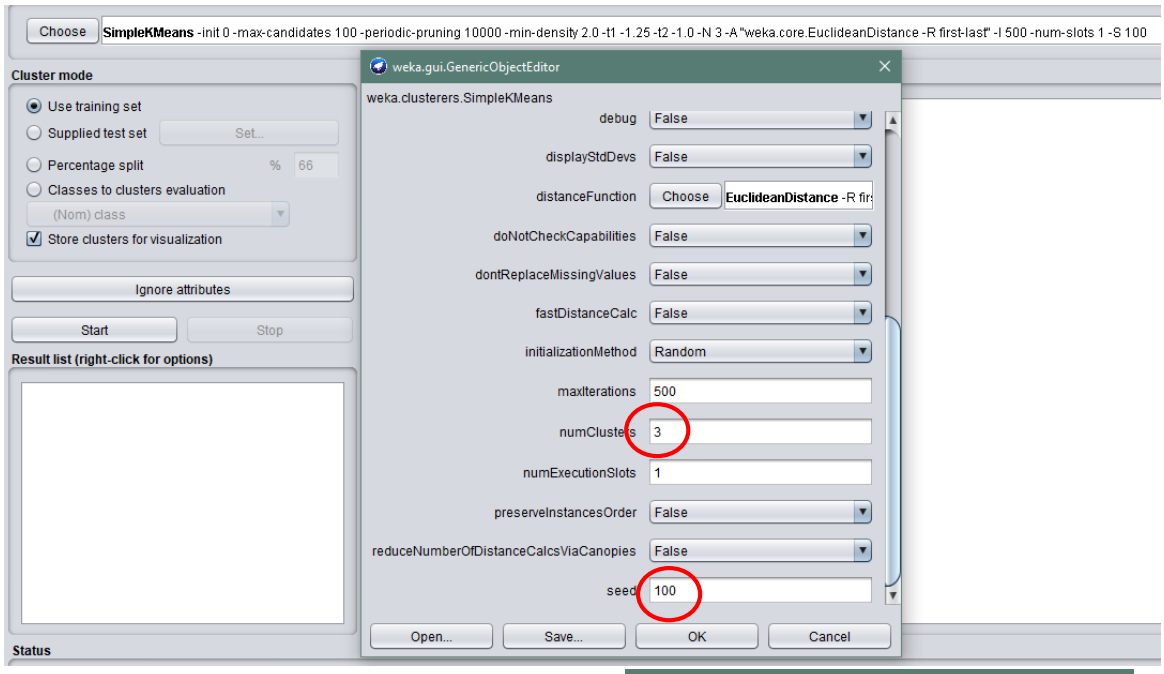
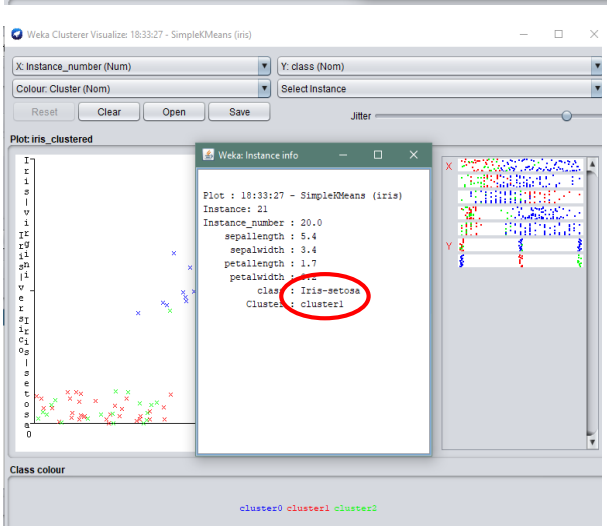
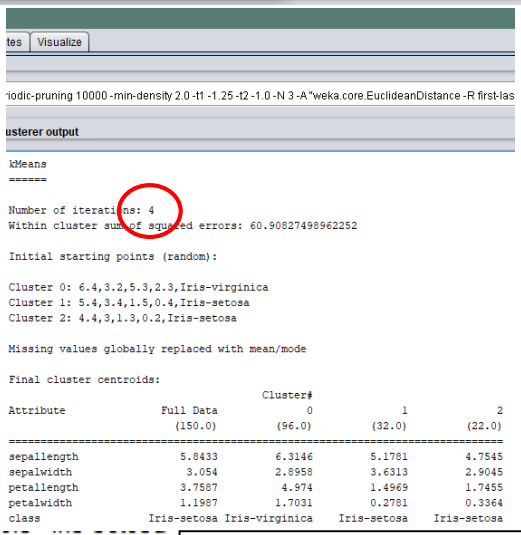
1. 利用 Weka 對 iris. arff 進行 Unsupervised Clustering，使用 Simple K-Means 演算法進行分群，產生的群不可以大於 3 群，在過程中對重要步驟截圖並加以說明，並回答以下問題：

(a) 設定 Use training set 且 k 值為 3、Seed 為 100，此結果 K-Means run 了幾次結束？Instance_number=20 的資料其所歸類的群是否與原始資料的分類相同？從結果中你認為哪些 data 是 Outlier？(20%)





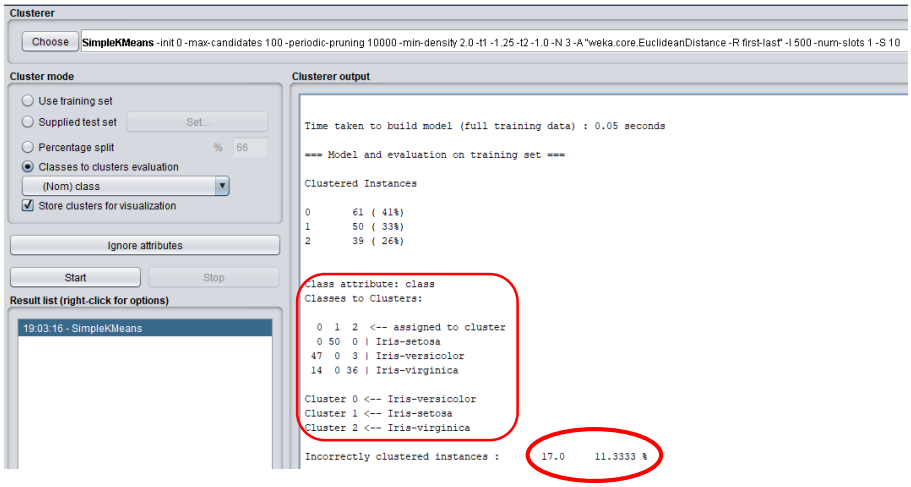


20	5.1	3.8	1.5	0.3	Iris-setosa

The original data of “Instance_number = 20”.

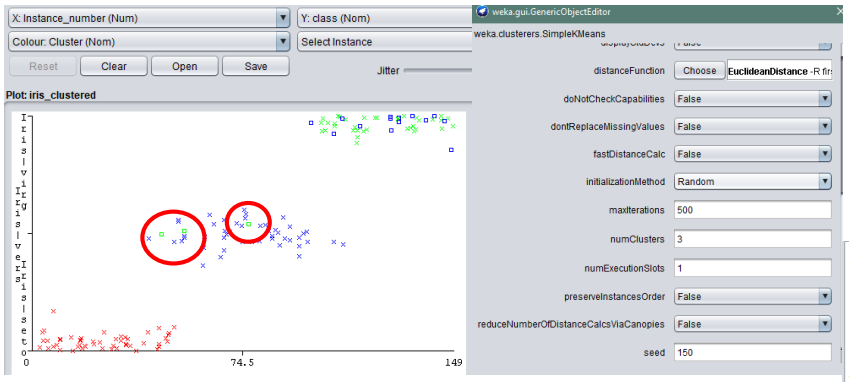
According to the screenshots, “Number of iterations = 4” indicates that K-Means had run for four times. The data which “Instance_number = 20” has the same class as its original data after clustering with K-Means.

(b) 設定 Classes to clusters evaluation 為 Class，且將 k 值設定為 3，試著說明原始資料與 預測群之間的關係。 (10%)

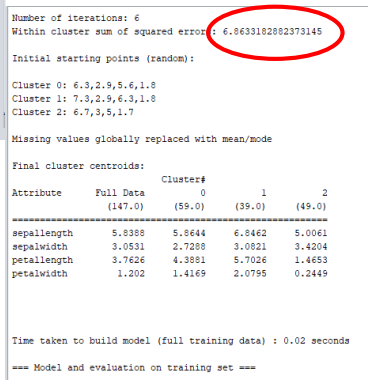


這邊被分成 cluster 0, cluster 1, cluster 2, 三個預測群。其中有三筆原爲 Iris-versicolor 的 instances 被分至 Iris-virginica, 有 14 筆原爲 Iris-virginica 的 instances 被分至 Iris-versicolor 。左圖下方表示有總共 17 筆 Incorrectly clustered instances, 其錯誤分群率為 11.3333%。

(c) 請嘗試調整 K-means 演算法的參數，找出最好的分群模型。請紀錄並截圖所使用的 設定 (例如: 使用哪些屬性、初始中心為何或是移除了哪些離異值) 。 (20%)



將右圖紅色圈起來三個 instances 刪除，分別是 instance_number = 50, 52, 77。將 seed 改爲 150。



最後呈現出來的 sum of squared errors = 6.8633182882373145，為嘗試多次後最小值。

(d) 承 (c) 小題，請問此結果分了幾群 ?並說明各群的特色。 (10%)

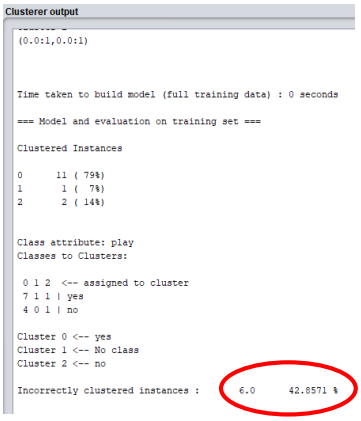
經過 clustering 之後，分成三個群，cluster0 的 class 全為 Iris-versicolor，cluster2 全爲 Iris-setosa，最後 cluster1 有部分 Iris-virginica 以及 Iris-versicolor。

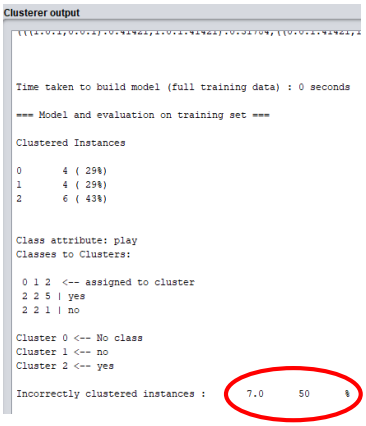
2. 用 Weka 軟體對 weather. nominal. arff 建立 HirarchicalClusterer ，選擇 “Classes to clusters evaluation” ，設定 numClusters = 3, distanceFunction = EuclideanDistance，在過程中對 重要步驟截圖並加以說明，並回答以下問題：

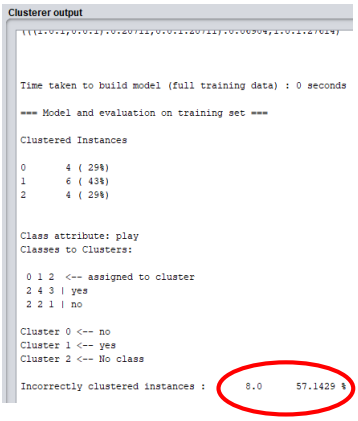
(a) 請嘗試使用以下的聚合判定方式進行聚合：Single、Complete、Average，請解釋三個方法的差別。 (15%)

先說明一下，single 取得是點與點之間的最小值，complete 是取最大值，而 average 則是取平均。

(b) 承上題，請問哪個聚合方式的效果最差？哪個聚合方式的效果最佳？(10%)







由左到右依序是 simple-linkage, complete-linkage, average-linkage。可以從 incorrectly clustered instances 中發現 simple-linkage 的錯誤率為 42.8571%，為最小所以效果最佳。而 average-linkage 的錯誤率為 57.1429%，為最大所以效果最差。

(c) 請列出 weather. nominal. arff 的 First Iteration (如下圖)。(記錄在 Word 上或手算拍照附圖皆可) (15%)

以下列出 I1~I5 的 first iteration。

No.	1: outlook	2: temperature	3: humidity	4: windy
	Nominal	Nominal	Nominal	Nominal
1	sunny	hot	high	FALSE
2	sunny	hot	high	TRUE
3	overcast	hot	high	FALSE
4	rainy	mild	high	FALSE
5	rainy	cool	normal	FALSE