Group Report：Red Wine Quality

第三組

梁湘梅 40541204S 、沈林緯 40541131S

- **梁湘梅的報告總結**

  分類器的選擇上我們使用了 Random Forest Classifier、Stochastic Gradient Descent Classifier、Support Vector Classifier(SVC)三種分類器。而三種分類器經過交叉驗證所得到的 Accuracy 分別為：Random Forest Classifier 為 91.2%；Stochastic Gradient Descent Classifier 的為 85.8%；Support Vector Classifier(SVC)為 90.8%。其中 Random Forest Classifier 的 Accuracy 值是最好的，因此我們最後會選用 Random Forest Classifier 進行分類。

- **沈林緯的報告總結**

  分別對 Decision Tree 與 Random Forest 這兩種分類器的效能進行分析。對 Decision Tree 來說隨著深度的增加對於訓練資料的分類越來越接近完美，但是對於測試資料來說反倒是深度在 10 之後準確度不如以往，另外，在不同深度下的 Decision Tree 執行時間，在深度=20 之後幾乎沒什麼變化。對 Random Forest 來說，首先，因為 Random Forest 本身就是透過多個 tree 來進行擬和，因此所得到的分數普遍比 decision tree 高，接著，random forest 在不同 tree 數量下，隨著 tree 數量上升，執行時間也會上升。所以儘管 random forest 普遍上的分類效果比 decision tree 好，但是其花費的時間也很驚人。

- **統整觀察**

  在資料的前處理上，梁湘梅的報告中並沒有對資料做降維的處理。

  我們最常通過降維以減少極端值、雜訊所造成的誤差與成本，提高精度、效率。而在這次的實作中，降維的主要目的是將去除冗餘資訊，以提高處理速度。通過資料分析，我們發現，固定的酸度、pH、密度，這三個屬性與 Quality 的關係並不高，也因此在沈林緯的報告中，最後取了 8 個屬性。透過降維，能使我們在分類的時候更能專注在對於 Quality 有影響的 Attribute 上。

- **結論**

  對於降維，梁湘梅的報告中在前處理上遺忘了降維的步驟。而在沈林緯的報告中，使用降維的主要目的是為了提高處理速度。

- **深刻檢討**

  事實上，我們兩份報告的相關性實在不高，唯一的共通性大概就是使用同一份資料集進行實作。

  由於我們這次的報告是等兩人的個人報告各自做完後，才進行討論。一方面是兩人都很忙碌，沒辦法空出時間一起做報告；另一方面是想說兩人各自做報告，相異性比較高，最後可以討論的不同點比較多。但真正進行討論時才發現雙方所注重的面向不同，兩份報告幾乎是完全不一樣的，一份是對於 Random Forest Classifier、Stochastic Gradient Descent Classifier、Support Vector Classifier (SVC)三種分類器的 Accuracy 比較；另一份則是探討 Decision Tree 與 Random Forest 這兩種分類器的效能，導致我們的小組報告無法進行統整討論。

  經過這次的報告後老師對我們的回饋，還有在聽過其他組別的報告後，我們檢討出為何

我們的報告會無法統合，原因在於我們沒有在最開始時就決定好要討論的主題。比起用同一個資料集進行分析，更重要的是要分析後討論的方向。我們應該提前選擇好要討論的方向是 Accuracy 還是效能，兩人分別負責兩至三組模組的分析。又或者是，對於同樣的兩至三組模組，一人進行 Accuracy 的探討；另一人進行效能的探討，最後在合併找出最佳的分類器。