Competition1

指導教授：李政德　教授

組員:王子豪 H24031346 周逸平 H24036087

一、模型介紹(Description)

我們使用了隨機森林模型(Random Forest)對 TF-IDF(term frequency - inverse document frequency)矩陣進行參數篩選，後利用支援向量機(Support Vector Machine)中的 Support Vector Regression 模型使用篩選後的參數對資料進行分類。

1. 隨機森林模型：

隨機森林是一個包含多個決策樹的分類器，並且其輸出的類別是由個別樹輸出的類別的眾數而定。其中權重部分所使用的資訊獲利(Information Gain)方法主要利用一資料集的熵值(entropy)與該資料集被某一參數分群時的熵值之差以判斷該參數的重要性。

TF-IDF：

一種用於資訊檢索與文字挖掘的常用加權技術。tf-idf 是一種統計方法，用以評估一字詞對於一個檔案集或一個語料庫中的其中一份檔案的重要程度。字詞的重要性隨著它在檔案中出現的次數成正比增加，但同時會隨著它在語料庫中出現的頻率成反比下降。

支持向量迴歸模型：

支援向量機模型(Support Vector Machine，SVM)是將例項表示為空間中的點，這樣對映就使得單獨類別的例項被儘可能寬的明顯的間隔分開。然後，將新的例項對映到同一空間，並基於它們落在間隔的哪一側來預測所屬類別。支持向量迴歸模型（Support Vector Regression，SVR）是使用 SVM 來擬合曲線，做迴歸分析。與分類的輸出是有限個離散的值所不同之處在於，回歸模型的輸出在一定範圍內是連續的。

二、步驟說明：

1. 資料前處理：
   a. 將 training data 中的 text 項與 testing data 中的 text 項合併後取 tfidf。
   b. 將 a.取完的 tf-idf 矩陣分割為 training data 的 tf-idf 矩陣與 testing data 的 tf-idf 矩陣

2. 參數篩選—隨機森林模型：
   將 train data 之 tf-idf 矩陣放入隨機森林模型之中，其權重計算方式使用「資訊獲利」進行計算。依此方式將特徵參數減少至 2000 個。

3. 配適模型—SVR 模型：
   將 2.中選擇的 2000 個參數放入 SVR 模型之中，其中 Cost 設定為 10，Gamma 設定為 2。

4. 預測輸出：
   將 testing data 的 tf-idf 矩陣放入配適好的 SVR 模型進行預測，並將預測結果中，小於 1 的令為 1，大於 5 的令為 5，後取四捨五入。
   透過以上步驟對測試資料的 stars 進行預測，最終預測準確度為 RMSE 得 0.8587，ACC 得 0.4428。

Github url:
https://github.com/isthereanyusernameNOTtaken/data_sciense/tree/master/report2