**DM期末專題 第11組**

**組員：陳欣儀0553442、李宗穎0653402、張雅婷0653413、黃偉嘉0653421**

一、Dataset：Police Killings

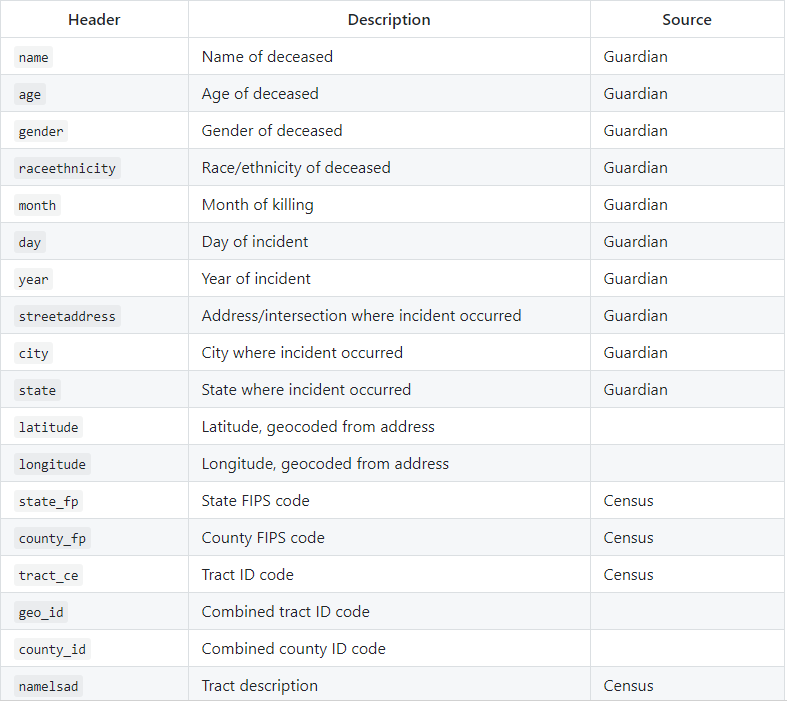
（資料來源：<https://github.com/fivethirtyeight/data/tree/master/police-killings>）

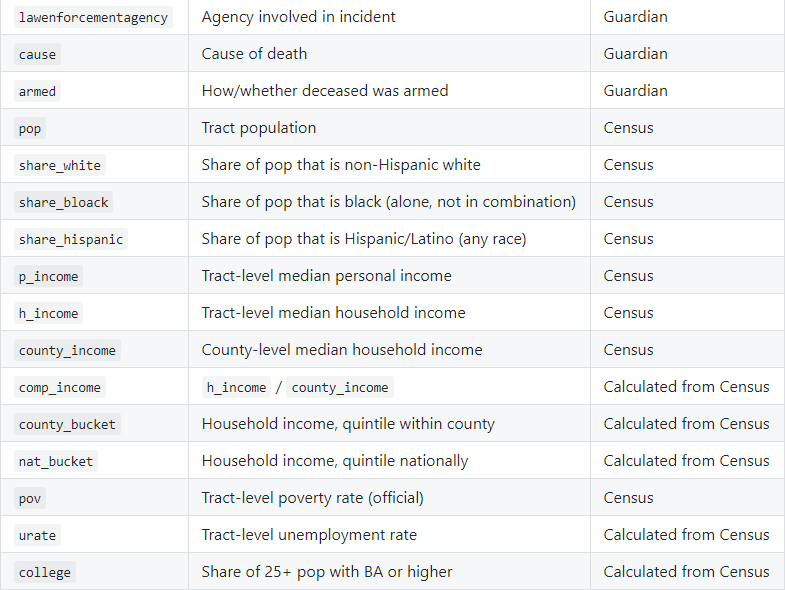
1. 資料集介紹、使用動機

此資料集的來源為Github，內容包含了2015年警方殺死美國人的故事背後的資料，主要是把衛報(Guardian)關於警察殺人事件的資料庫與美國社區調查的人口普查數據結合起來呈現，那我們這組的資料集使用動機是想**了解種族歧視是否為造成這些憾事的主因**，因此我們將把種族膚色作為訓練目標並且進一步預測結果。

1. 資料集說明

csv檔欄位的說明如下：





註：Guardian是衛報、Census是人口普查

二、分析工具

VirtualBox Ubuntu, Jupyter, Apache Spark

三、實作流程與評估方法（截圖說明、在spark執行的部分）

1. 實作流程為：

讀檔 (the-counted-2015.csv, the-counted-2016.csv)

-> 資料前處理（轉data type、種族膚色作為target value、spark dataframe）

-> 切檔案作為training, testing data (75%, 25%且轉成labeled RDD)

-> 訓練model -> 測試 model

2. 評估方法為：

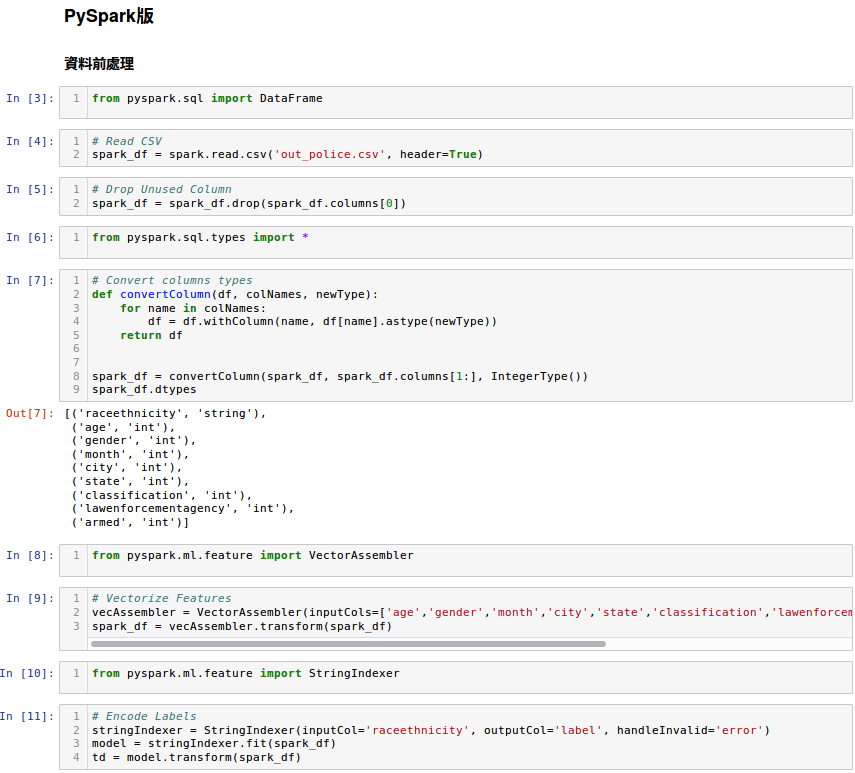
Gradient-Boosted Tree Classifier (GBT)

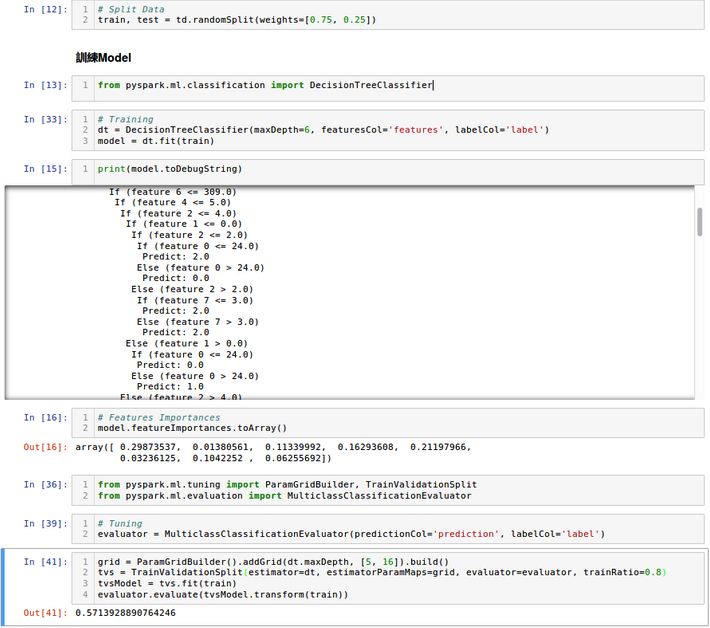
Decision Tree Classifier

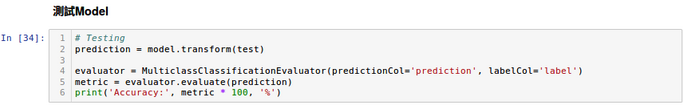
Random Forest Classifier

Naive Bayes

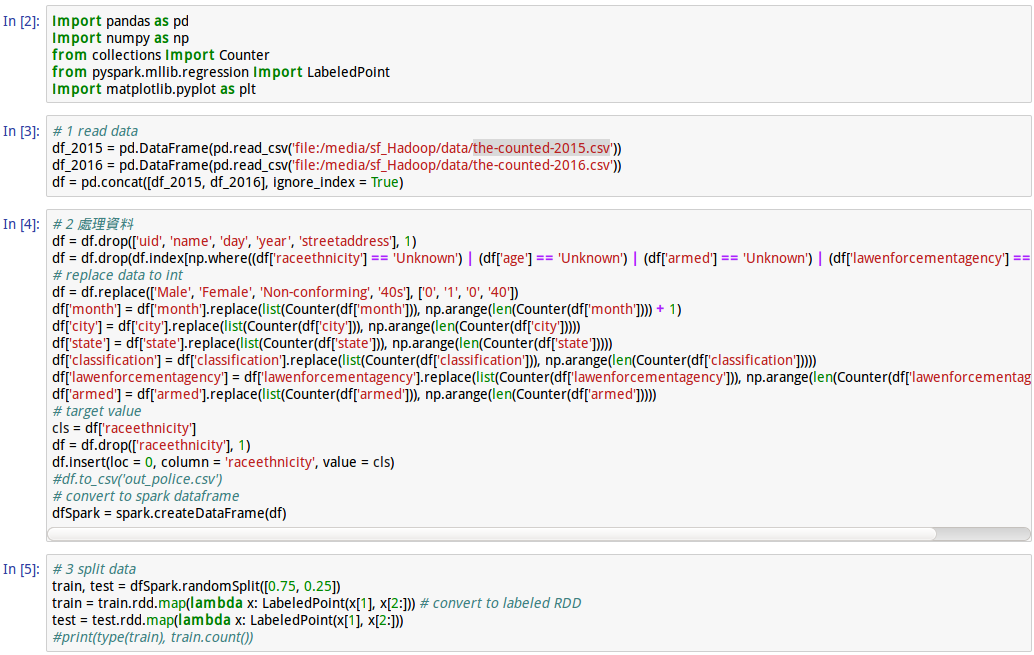
1. 過程截圖：
2. Decision Tree Classifier

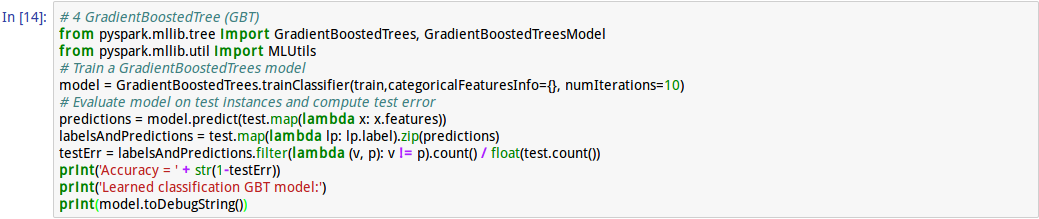






1. Gradient-Boosted Tree Classifier (GBT)



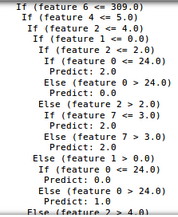


1. Random Forest Classifier
2. Naïve Bayes

四、預測結果比較

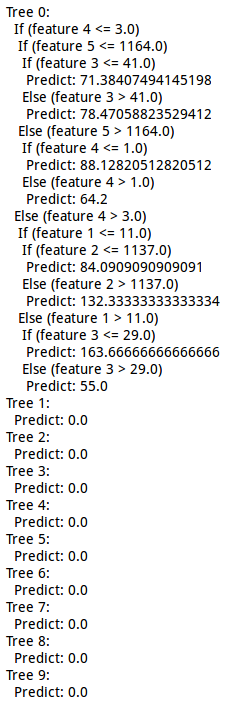
(1) Decision Tree

Accuracy: 50.485173602484764



(2) Gradient-Boosted Tree Classifier (GBT)





(3) Random Forest Classifier



(4) Naïve Bayes



五、結論