

# 資料探勘研究與實務-作業 5(Spark)

## 第三組

0753431 吳伯揚 0753425 李嘉晨

0753440 吳肇堉 0753423 劉奕辰

### i. 截圖與步驟描述



#### ■ 安裝 Virtue Box 與相關網路、環境設定

```

hduser@master: ~
hduser@master:~$ ls
anaconda2
Anaconda2-2.5.0-Linux-x86_64.sh
examples.desktop
hadoop-2.7.7.tar.gz
homework
hduser@master:~$

```

homework.zip  
pythonwork  
spark-2.0.2-bin-hadoop2.7.tgz  
wordcount  
下載

公共  
圖片  
影片  
文件  
桌面

模板  
音樂

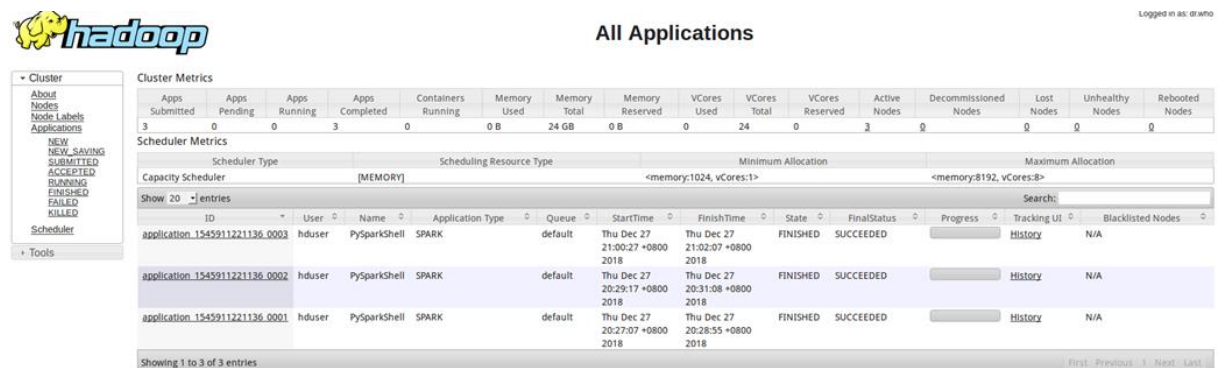
#### ■ 在虛擬機 Master 上 安裝 Hadoop, Spark, Anaconda 等相關套件

```

hduser@master:~$ hadoop fs -ls
Found 3 items
drwxr-xr-x - hduser supergroup          0 2018-12-27 21:02 .sparkStaging
drwxr-xr-x - hduser supergroup          0 2018-12-27 19:48 test
drwxr-xr-x - hduser supergroup          0 2018-12-27 20:24 wordcount
hduser@master:~$ hadoop fs -ls test
Found 4 items
-rw-r--r--  3 hduser supergroup        1366 2018-12-27 19:15 test/README.txt
drwxr-xr-x - hduser supergroup          0 2018-12-27 19:17 test/etc
drwxr-xr-x - hduser supergroup          0 2018-12-27 20:15 test/homework
-rw-r--r--  3 hduser supergroup        1366 2018-12-27 19:15 test/test1.txt
hduser@master:~$ hadoop fs -ls test/homework
Found 3 items
-rw-r--r--  3 hduser supergroup        64170 2018-12-27 19:48 test/homework/final
.csv
-rw-r--r--  3 hduser supergroup          10 2018-12-27 20:15 test/homework/t.txt
-rw-r--r--  3 hduser supergroup         1411 2018-12-27 19:48 test/homework/test.
py
hduser@master:~$

```

- 建立 HDFS 目錄並放入資料集(CSV)



The screenshot shows the Hadoop All Applications web interface. On the left is a navigation menu with options like Cluster, About, Nodes, Node Labels, Applications, NEW, NEW SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED, and Scheduler. The main content area is titled 'All Applications' and shows 'Cluster Metrics' at the top. Below this is a table of applications. The table has columns for ID, User, Name, Application Type, Queue, StartTime, FinishTime, State, FinalStatus, Progress, Tracking UI, and Blacklisted Nodes. Three applications are listed, all with a state of 'FINISHED' and 'SUCCEEDED'.

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI	Blacklisted Nodes
application_1545911221136_0003	hduser	PySparkShell	SPARK	default	Thu Dec 27 21:00:27 +0800 2018	Thu Dec 27 21:02:07 +0800 2018	FINISHED	SUCCEEDED	<div></div>	History	N/A
application_1545911221136_0002	hduser	PySparkShell	SPARK	default	Thu Dec 27 20:29:17 +0800 2018	Thu Dec 27 20:31:08 +0800 2018	FINISHED	SUCCEEDED	<div></div>	History	N/A
application_1545911221136_0001	hduser	PySparkShell	SPARK	default	Thu Dec 27 20:27:07 +0800 2018	Thu Dec 27 20:28:55 +0800 2018	FINISHED	SUCCEEDED	<div></div>	History	N/A

- 其他三台虛擬機為 data node 並執行程式

## ii. 作業目標對應之結果

```

accuracy = 0.6447368421052632
deathprecision = 0.4222222222222222
deathrecall = 0.5671641791044776
aliveprecision = 0.7898550724637681
aliverecall = 0.6770186335403726

```

- Accuracy, Recall, Precision 結果

```
DecisionTreeModel classifier of depth 5 with 10 nodes
If (feature 6 <= 0.0)
  If (feature 7 <= 0.0)
    If (feature 23 <= 0.0)
      If (feature 8 <= 0.0)
        If (feature 21 <= 0.0)
          Predict: 1.0
        Else (feature 21 > 0.0)
          Predict: 0.0
      Else (feature 8 > 0.0)
        Predict: 0.0
    Else (feature 23 > 0.0)
      If (feature 0 <= 42.0)
        If (feature 0 <= 29.0)
          Predict: 0.0
        Else (feature 0 > 29.0)
          Predict: 1.0
      Else (feature 0 > 42.0)
        If (feature 2 <= 0.0)
          Predict: 0.0
        Else (feature 2 > 0.0)
          Predict: 0.0
  Else (feature 7 > 0.0)
    If (feature 0 <= 0.0)
      If (feature 4 <= 0.0)
        If (feature 1 <= 0.0)
          Predict: 0.0
        Else (feature 1 > 0.0)
          Predict: 1.0
      Else (feature 4 > 0.0)
        Predict: 0.0
    Else (feature 0 > 0.0)
      If (feature 13 <= 0.0)
        If (feature 24 <= 0.0)
          Predict: 0.0
        Else (feature 24 > 0.0)
          Predict: 0.0
      Else (feature 13 > 0.0)
        Predict: 1.0
  Else (feature 6 > 0.0)
    If (feature 0 <= 50.0)
      If (feature 23 <= 0.0)
        If (feature 14 <= 0.0)
          If (feature 0 <= 29.0)
            Predict: 0.0
```

■ 產出決策樹模型(預測結果)

```

Predict: 0.0
Else (feature 0 > 29.0)
  Predict: 0.0
Else (feature 14 > 0.0)
  If (feature 0 <= 3.0)
    Predict: 1.0
  Else (feature 0 > 3.0)
    Predict: 0.0
Else (feature 23 > 0.0)
  If (feature 0 <= 14.0)
    If (feature 0 <= 6.0)
      Predict: 0.0
    Else (feature 0 > 6.0)
      Predict: 1.0
  Else (feature 0 > 14.0)
    If (feature 0 <= 33.0)
      Predict: 0.0
    Else (feature 0 > 33.0)
      Predict: 0.0
Else (feature 0 > 50.0)
  If (feature 4 <= 0.0)
    If (feature 2 <= 0.0)
      Predict: 1.0
    Else (feature 2 > 0.0)
      Predict: 0.0
  Else (feature 4 > 0.0)
    If (feature 10 <= 0.0)
      If (feature 0 <= 58.0)
        Predict: 0.0
      Else (feature 0 > 58.0)
        Predict: 0.0
    Else (feature 10 > 0.0)
      Predict: 1.0

```

■ 產出決策樹模型(預測結果) (續上頁)

### iii. 討論

此次分析與第一次作業相同，以預測腳色是否死亡為目的，其中 0 代表存活；1 代表死亡，

模型主要是預測腳色是否會死亡，其中 Accuracy 有 64%，並且由死亡的 Precision(42%)以及 Recall(56%)可以看出模型似乎沒那麼好，並沒有太符合我們的需求。

當中原因可能是因為資料與死亡的關聯性沒有那麼大，導致預測腳色死亡時並不那麼準確。

而從存活的 Precision(78%)以及 Recall(67%)都遠大於死亡的 Precision(42%)以及 Recall(56%)可以得知由於大部分的腳色都存活，死亡的人數較少，所以模型傾向預測腳色存活使得存活的 Precision 以及 Recall 較高。