**Big Data Programming**

**CSEE5590-0004/490-0004**

**Lab 4 Report**

Team : 6

Chandrasekhar Pentakota (class ID : 9)

Mustafa Bilal (class ID : 15)

**Links :**

<https://github.com/chanduhub/BigDataProgramming/wiki>

<https://youtu.be/Ud2dsdrKsYY>

https://github.com/bmian93/BigDataProgramming/tree/master/BigDataProgramming/Lab4

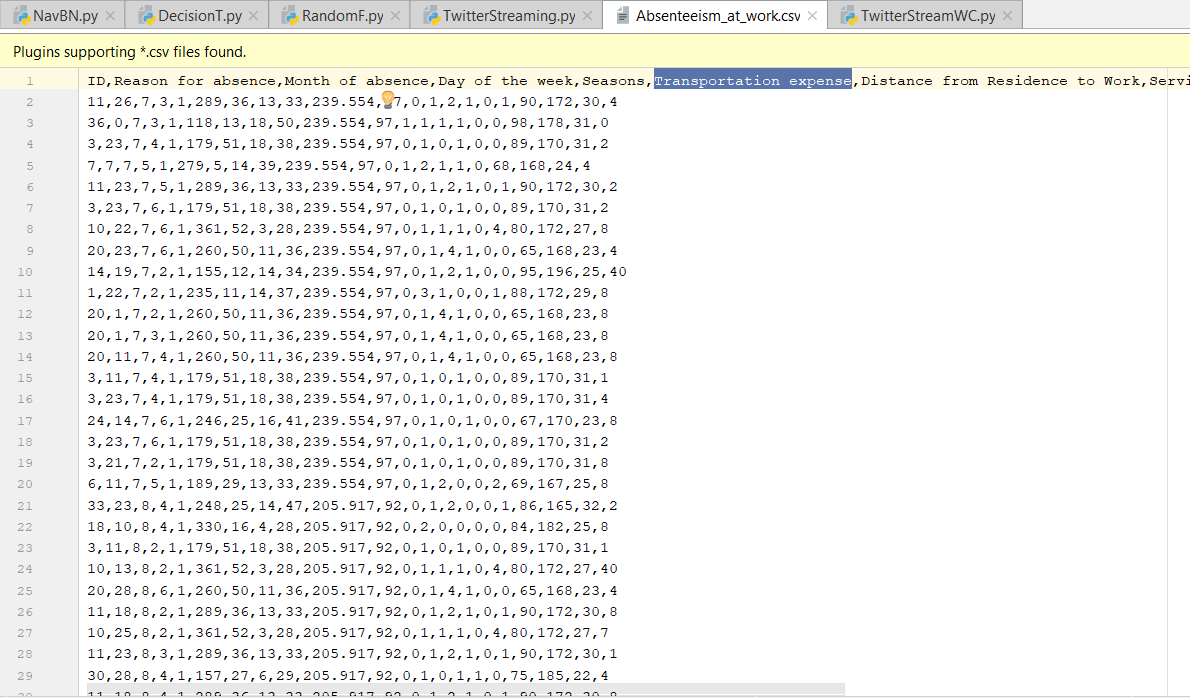
**Task 2**

**Objective :** Use the Classification Algorithms Naïve Bayes, Decision Tree and Random Forest for the same attribute classification.

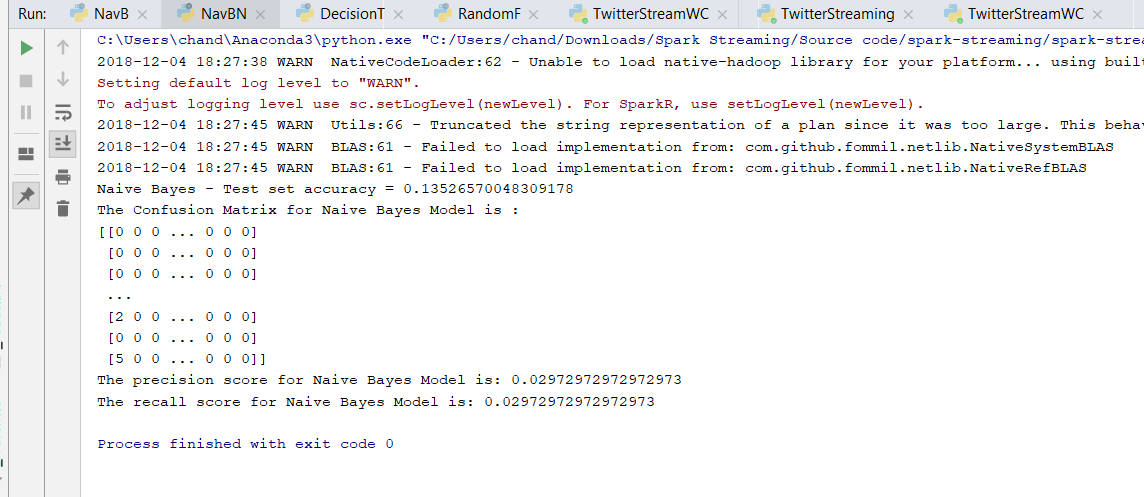
**Feautures :** Comparison of all the three classification algorithms and finding out the best one with the same attributes.

**Steps :**

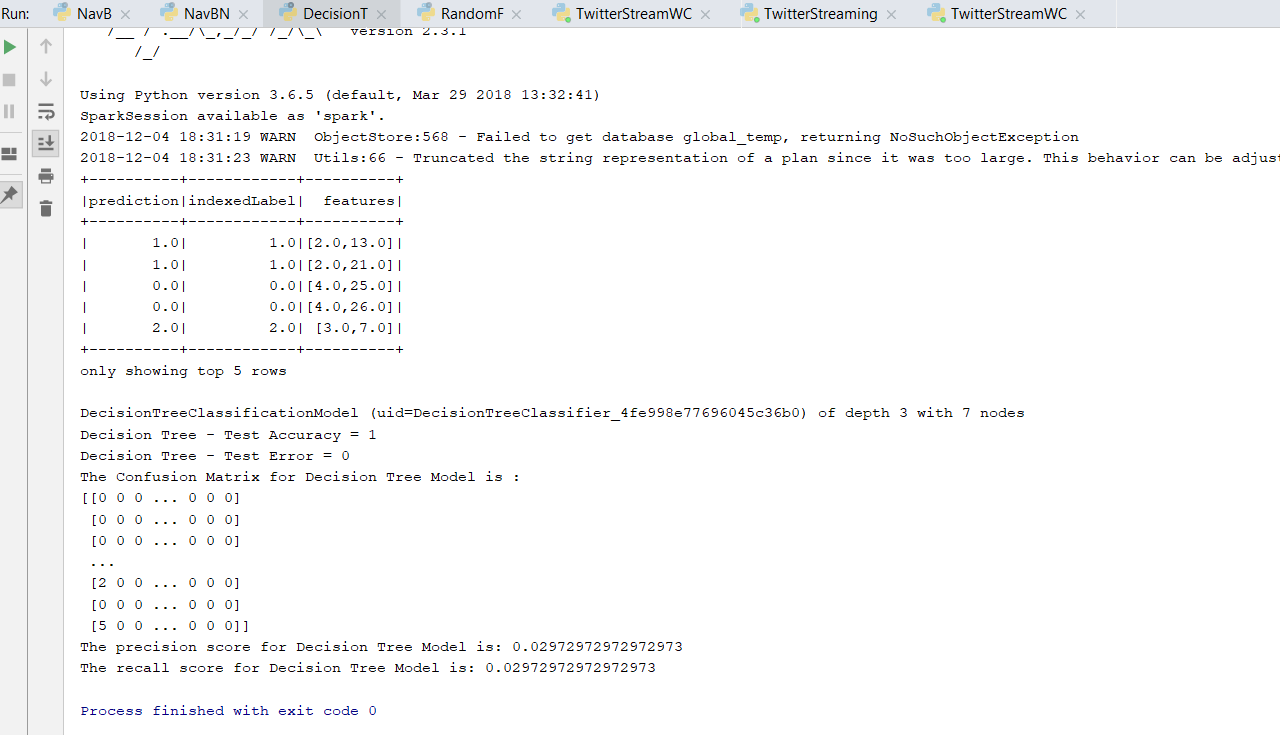
**Input –**

****

**Naïve Bayes Model Output –**



**Decision Tree Model Output –**



**Random Forest Model Output –**

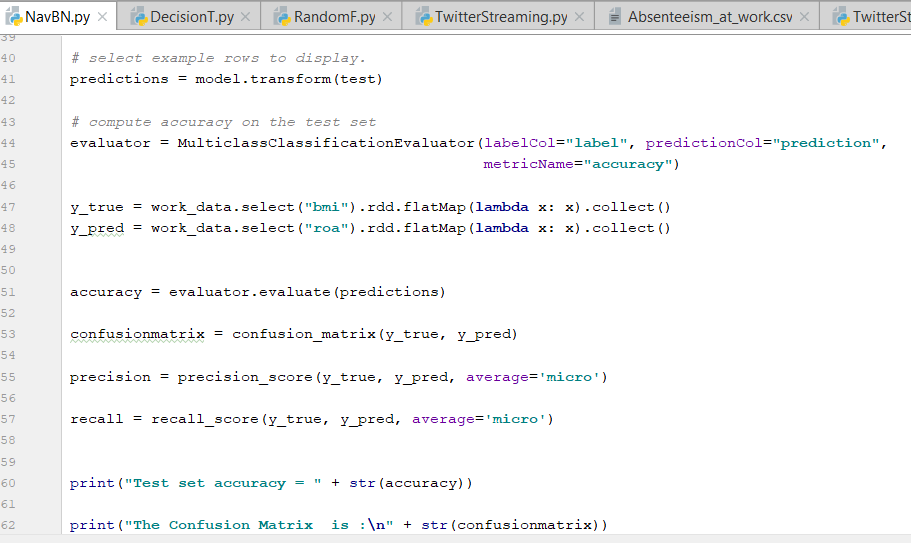


All the above three models used four features out of all which contributes for the well most prediction. The feautures include reason for absence, body mass index, age and transportation expense. The last one is included because people may think about their transportation expenses if they are not willing to go work on that day. Random forest model works best out of all the three models which got the highest accuracy for the model. The Confusion matrix, Accuracy based on FMeasure, Precision & Recall for all the algorithms is calculated and can be found from the output images.

**Code :**

**Naïve Bayes Model**





**Decision Tree Model**



**Random Forest Model**



**Task 3**

**Objective :** Perform Word Count on Twitter Streaming Data using Spark.

**Feautures :** This program takes streams of data from the twitter and the other program does word count on the collected streaming data.

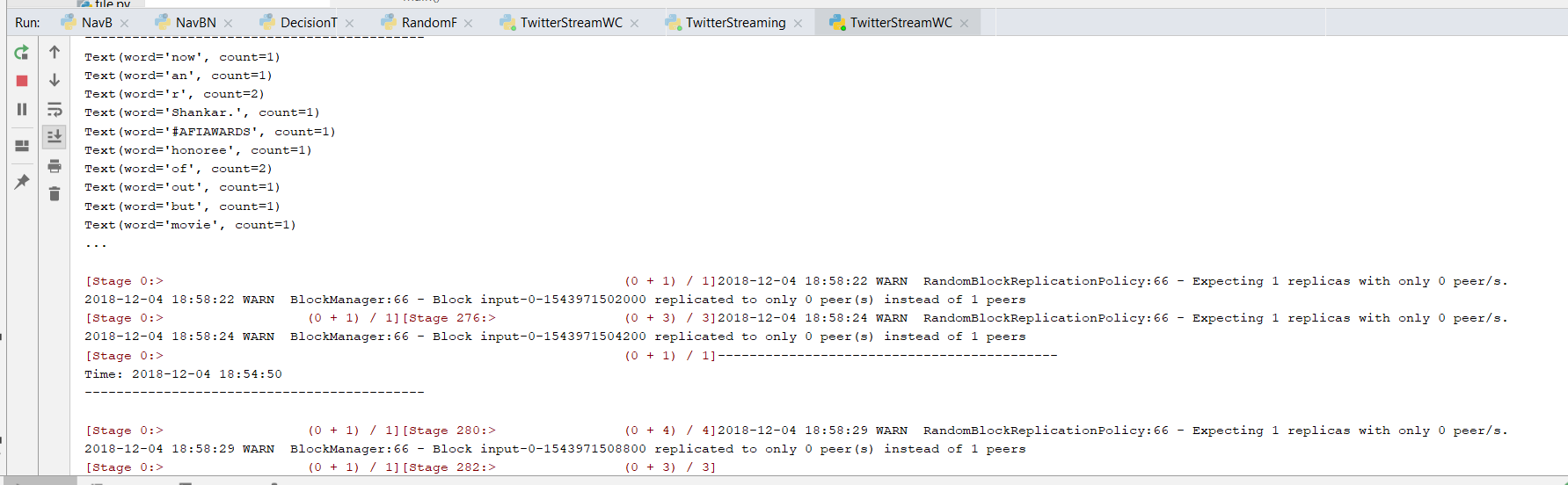
**Steps :**

**Input –**

The below data is live streaming data which is generated with the help of python code.



**Output –**



**Code :**

The below code is to get twitter streaming data using tweepy library in python using some unused port number on the system.



The below code is to run word count on the collected streaming data using map and reducebykey actions from the spark rdd.

