**NAME: HARISH CHANDRA JYOSHI**

**CLASS ID: 06**

**TEAM ID: 04**

**UMKC EMAIL ID:**[**hjddh@mail.umkc.edu**](mailto:hjddh@mail.umkc.edu)

**NAME: ATLURI VENKATA AKHILA KRISHNA**

**CLASS ID: 01**

**TEAM ID: 04**

**UMKC EMAIL ID:**[**vagq2@mail.umkc.edu**](mailto:vagq2@mail.umkc.edu)

**Video link: https://youtu.be/mGymAaEUWu4**

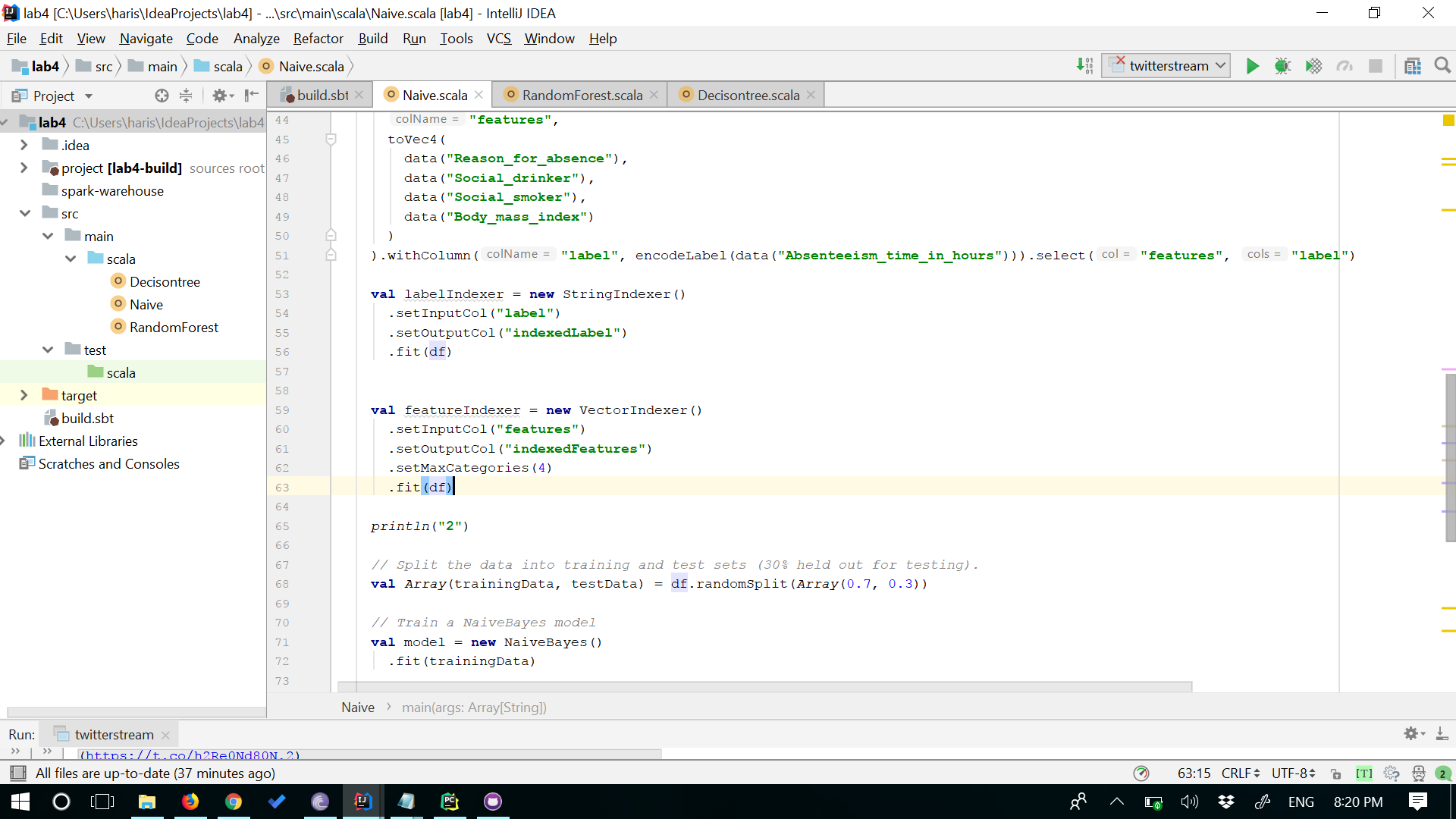
**Task 1: Implementing Spark Mlib classification algorithms on a data set.**

DataSet used: <https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work>

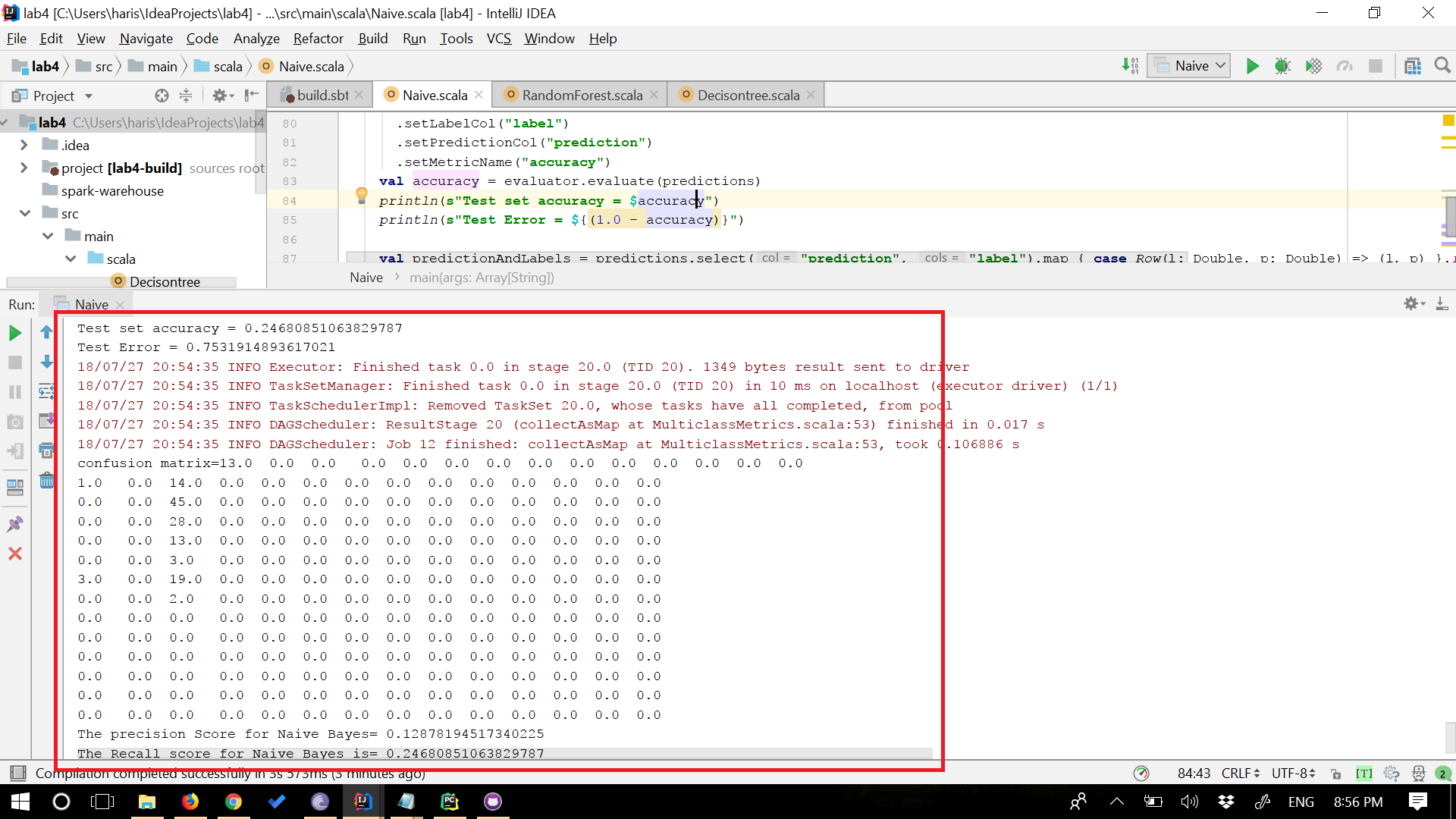
The input data is split into 70% for training and 30% for testing. Classification is performed based on the columns 'Reason for absence', 'social Drinker', 'Social smoker', 'Body mass Index', 'Absenteeism at work'.

**1.Naive Bayes:**

It's classification task based on Baye's theorem.

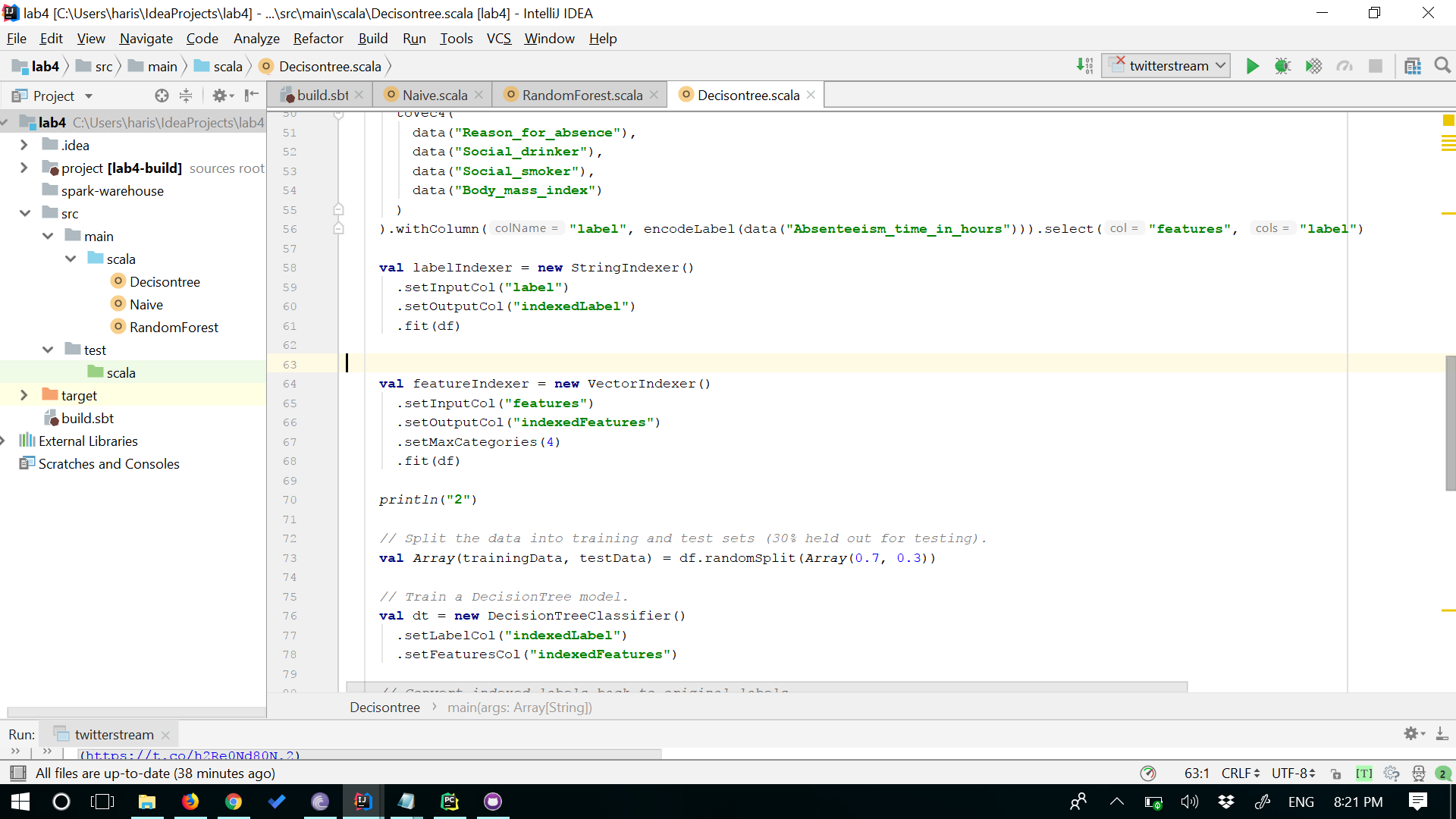


**Output after Running the Naive Bayes algorithm.**

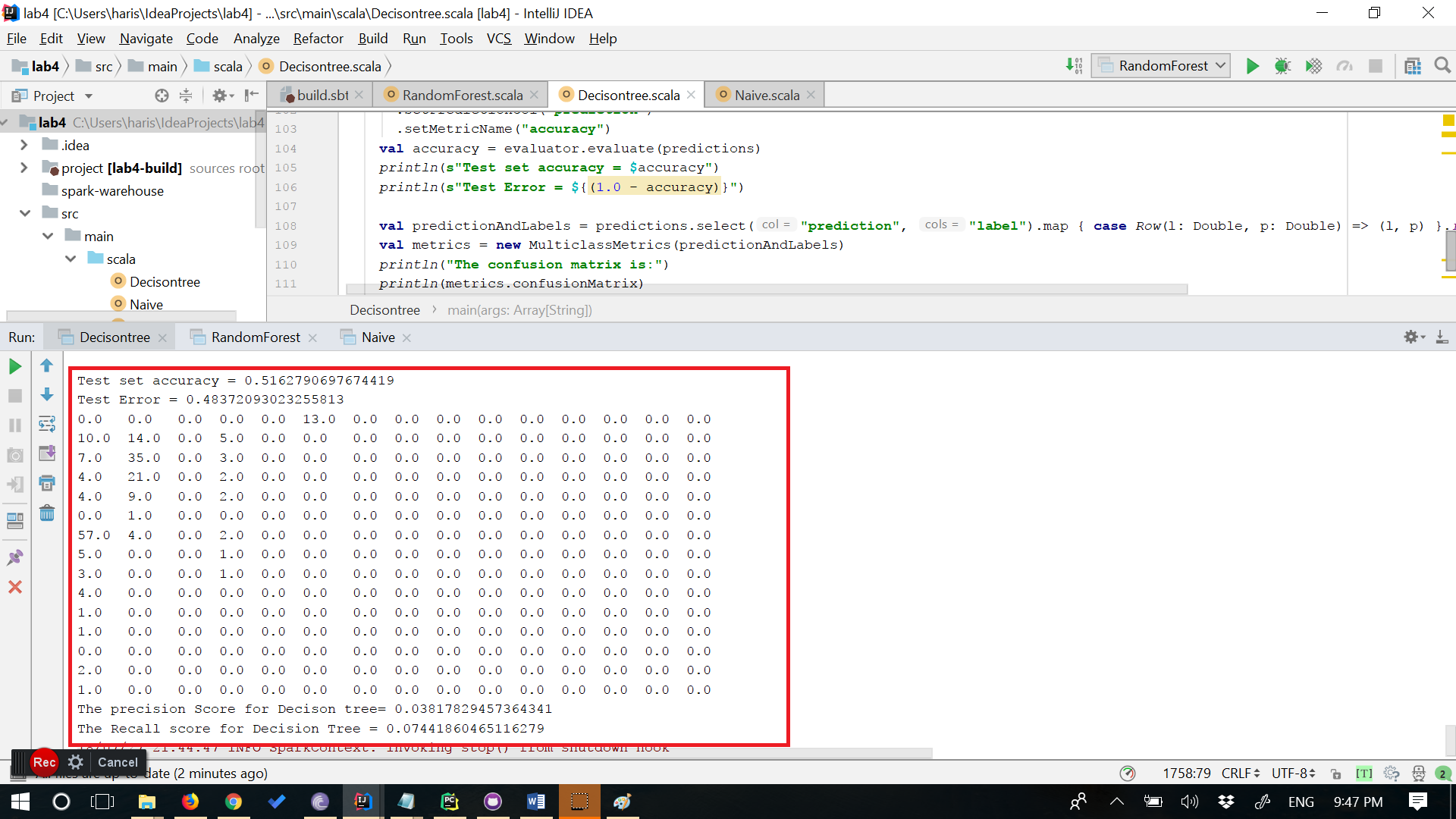


**2. Decision Tree**

The Input data is split into 70%-30%

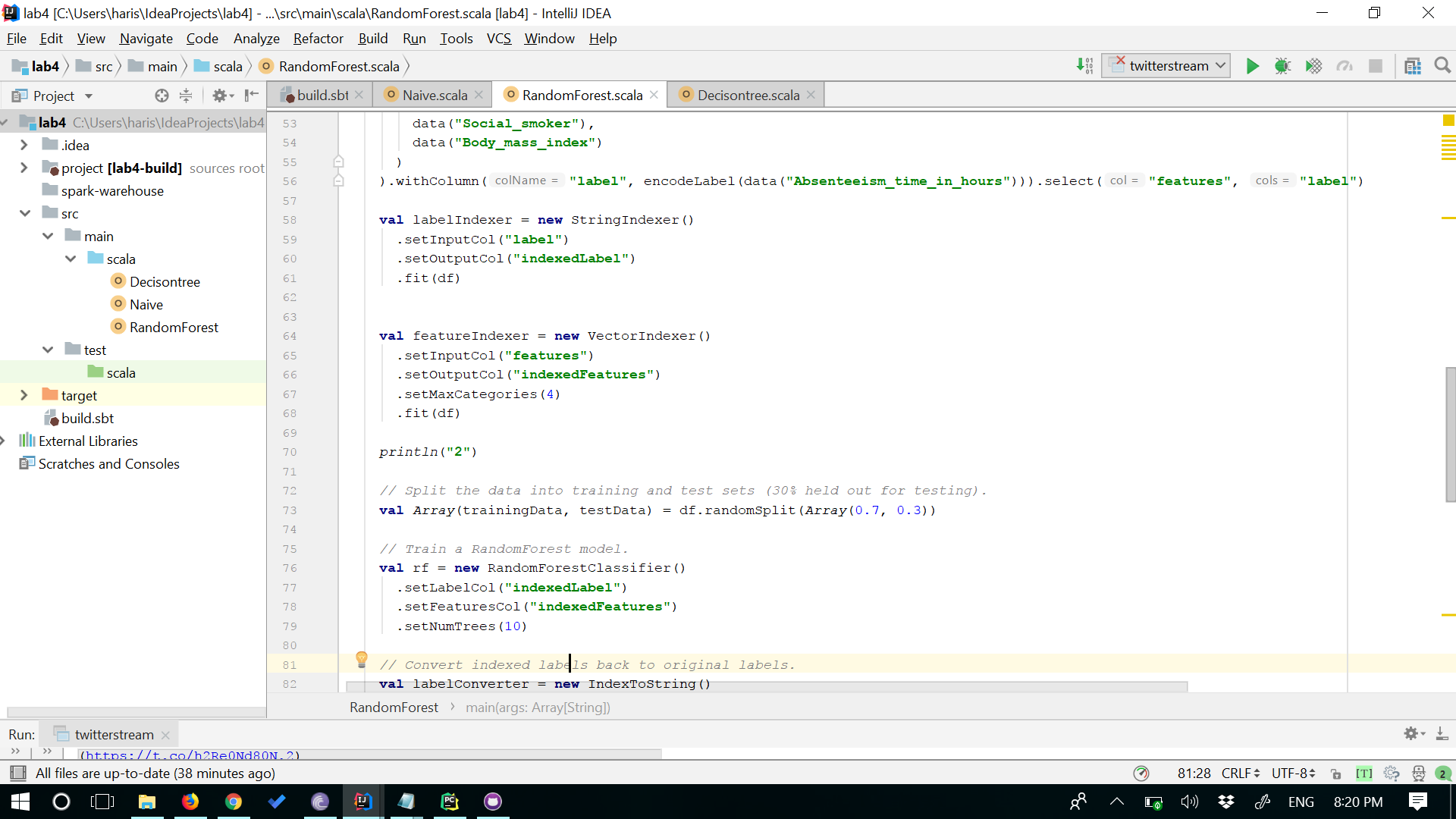


**output for Decision tree Algorithm**

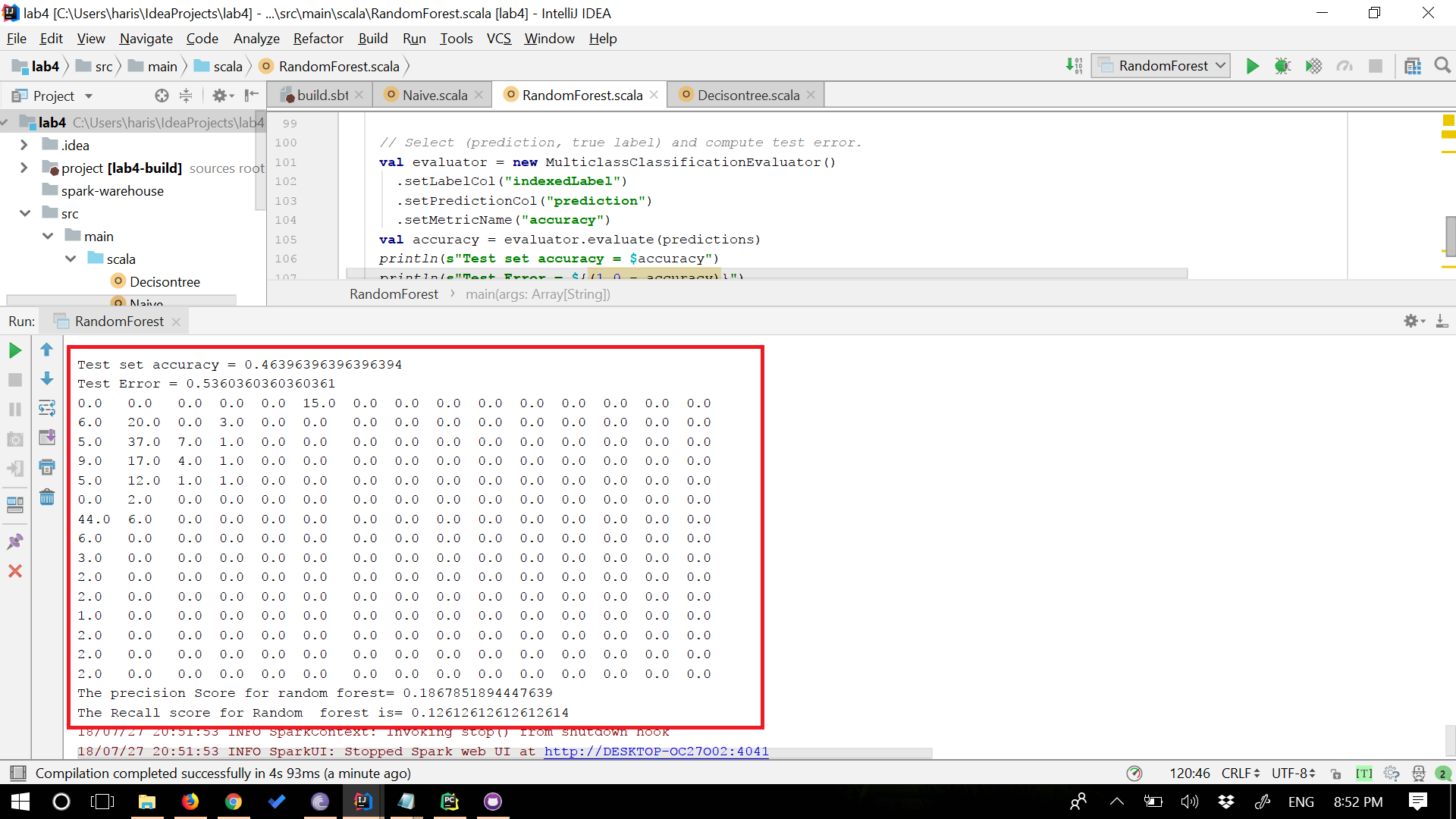


**3. Random forest**

The Input data is split into 70%-30%



**Output for random Forest algorithm**



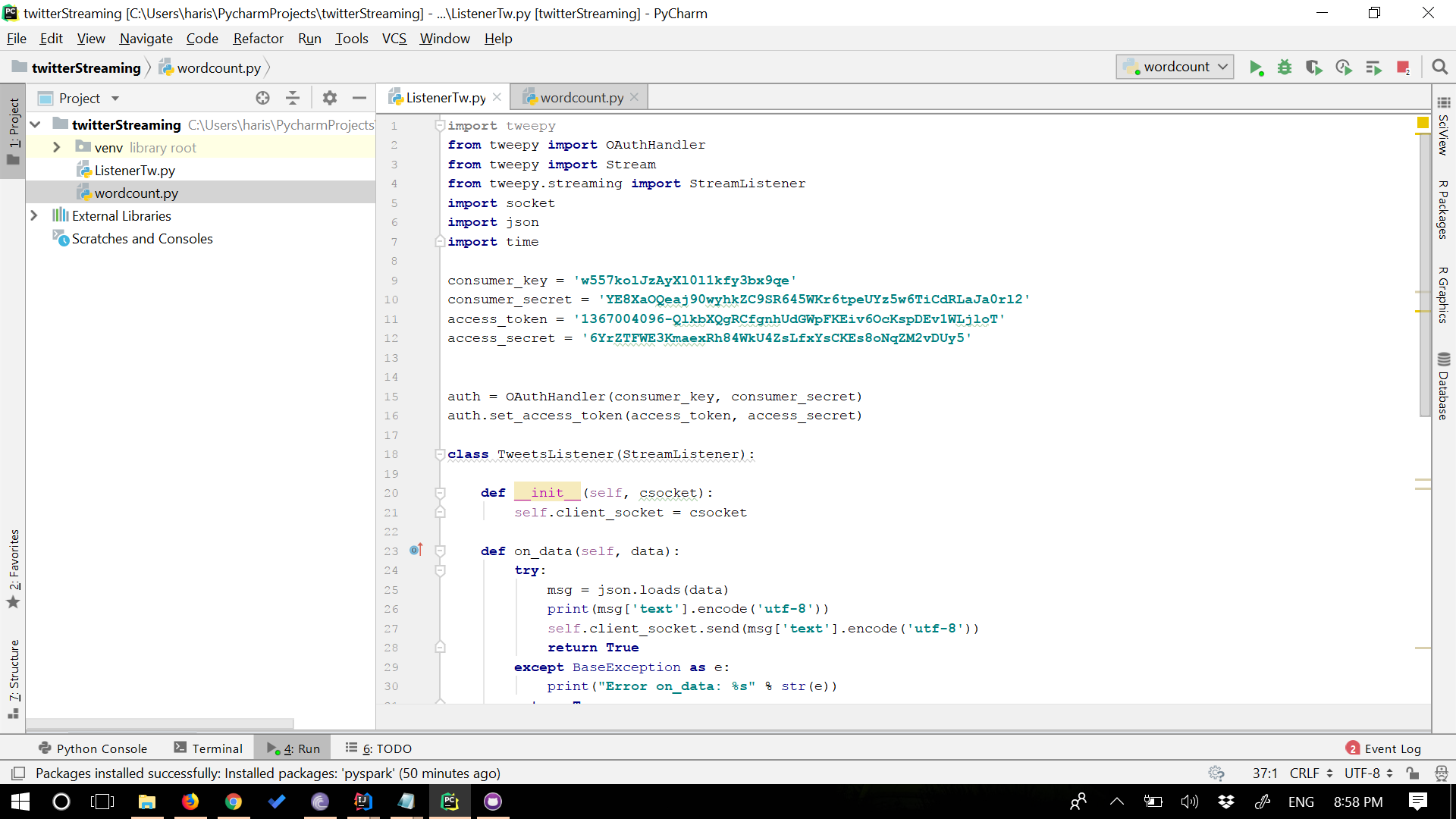
**Conclusion:**

According to the Results, Random Forest has the highest accuracy of 46% compared to decison tree and naive bayes. The confusion matrix and precision and recall values are shown for all the algorithms.

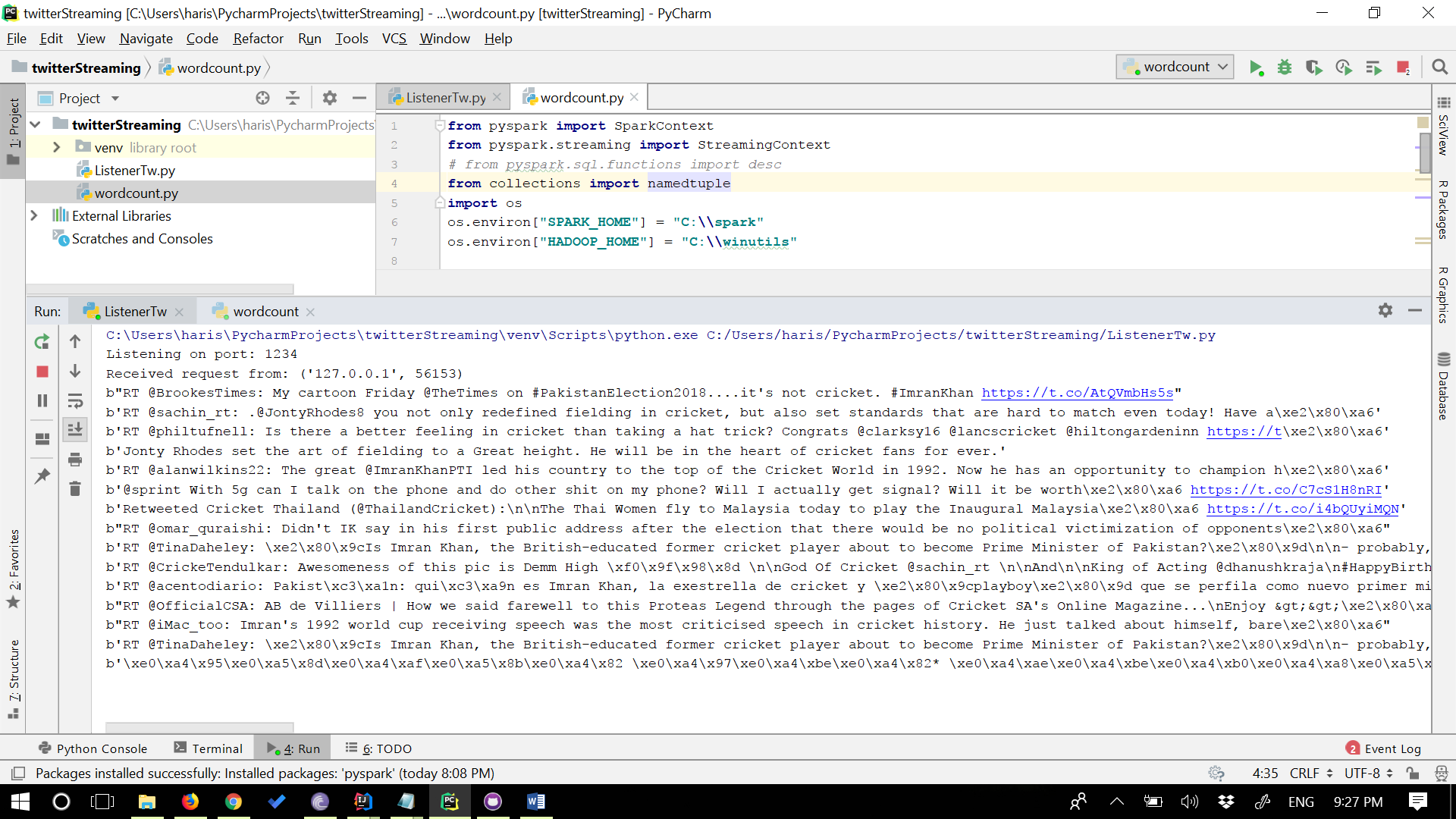
**Task 2: Spark Streaming**

In this task, streaming is performed on the Twitter data which is fetched using access and consumer keys.

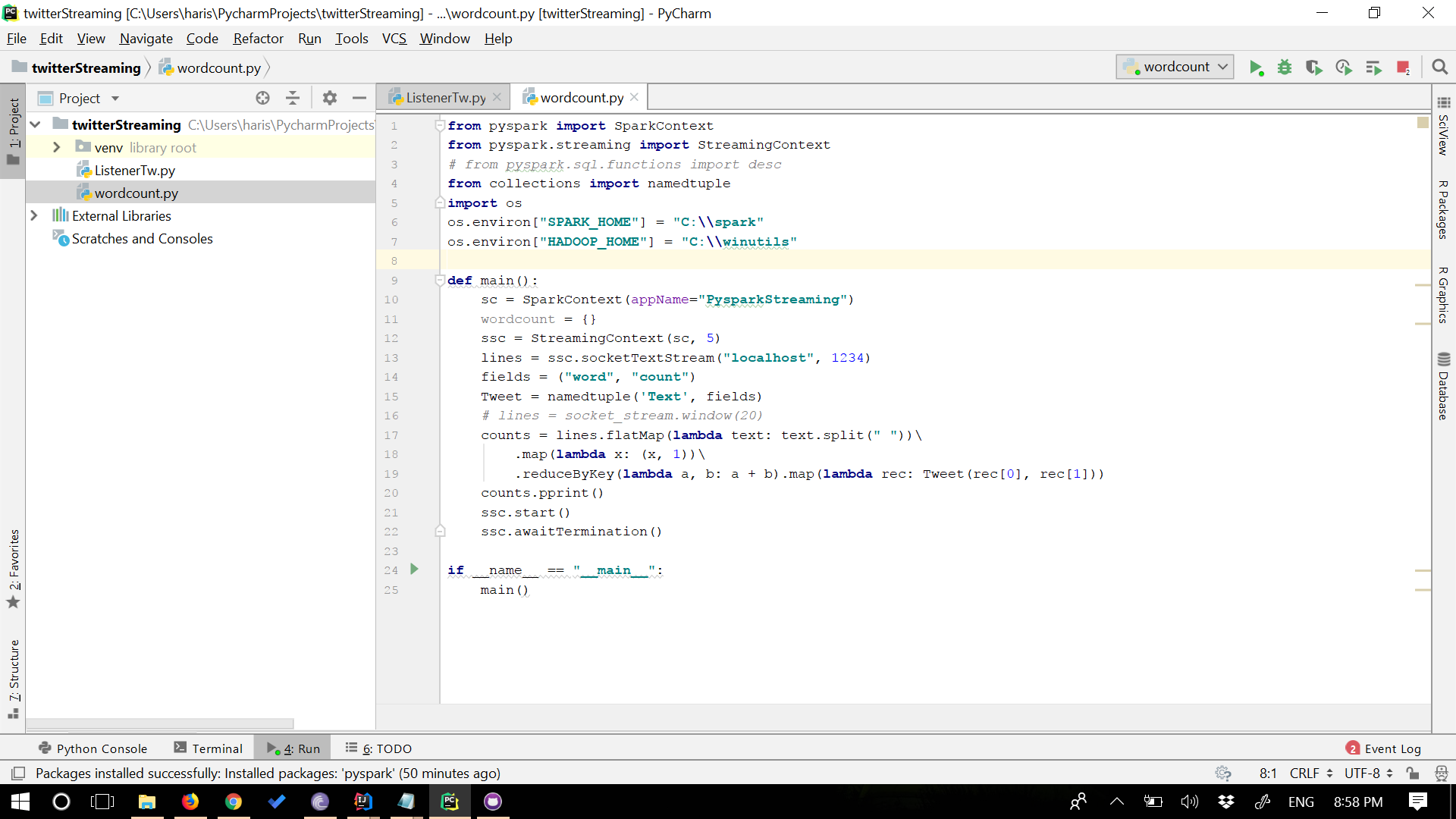
**1.Tweets collection:**



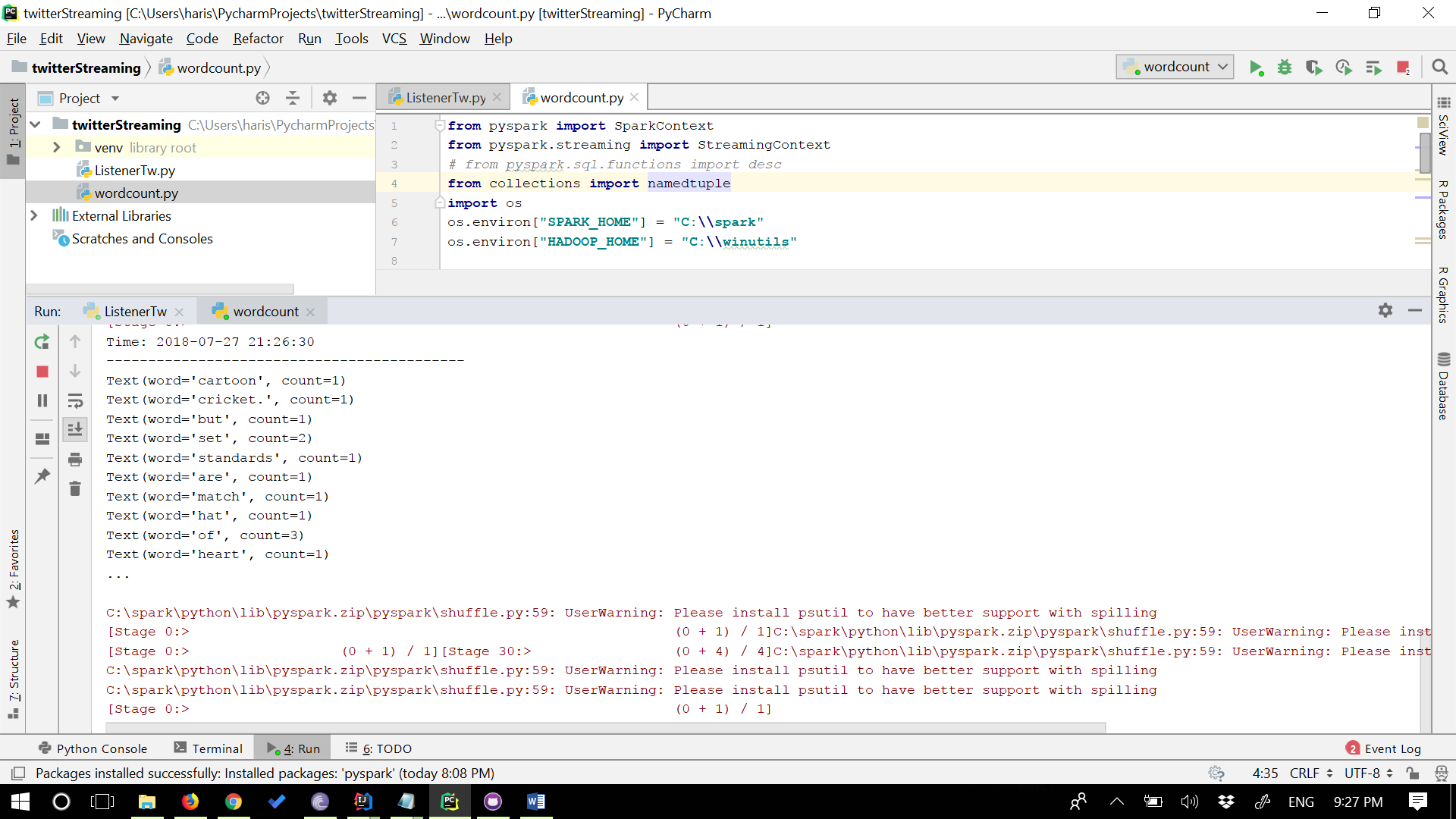
**Output for tweets collection:**



**2. Streaming and performing word count on twitter data:**



**output for Wordcount of Twitter data:**



References:

1. <https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work>
2. <https://spark.apache.org/docs/1.5.2/streaming-programming-guide.html>
3. <https://spark.apache.org/docs/2.1.0/mllib-naive-bayes.html>
4. <https://spark.apache.org/docs/2.2.0/mllib-decision-tree.html>

5.<https://github.com/dennyglee/databricks/blob/master/notebooks/Users/denny%40databricks.com/blog%20books/Scalable%20Decision%20Trees%20with%20MLlib.scala>

1. <https://www.linkedin.com/pulse/apache-spark-streaming-twitter-python-laurent-weichberger/>