**Kairiz CYber Technologies**

***Submitted By: Abdullah Amjad***

***Cell No: 03309772003***

***Gmail: abdullahamjad17301@gmail.com***

***Instructor: Sir Afaq and Mam Noor***

***CEO: Kainat Rizwan***

***Date: 17 July 2024.***

**Task: 3**

1. **Libraries:**

* **Spark-submit:**

Basically, this command runs the spark applications. It allows you to submit a job to a Spark cluster.

* **org.apache.spark:spark-sql-kafka-0-10\_2.12:3.1.2:**

This library is too much important because this library will enable spark to read from and write it to Kafka. It specifies the packages option of the spark-submit command.

1. **Streaming File (spark\_streaming\_kafka.py):**

Basically, Spark streaming Kafka is the python script I can say that contains the code to process data. Its work is to connect Kafka then reads the data and then process that data using the spark streaming.

1. **Producer File:**

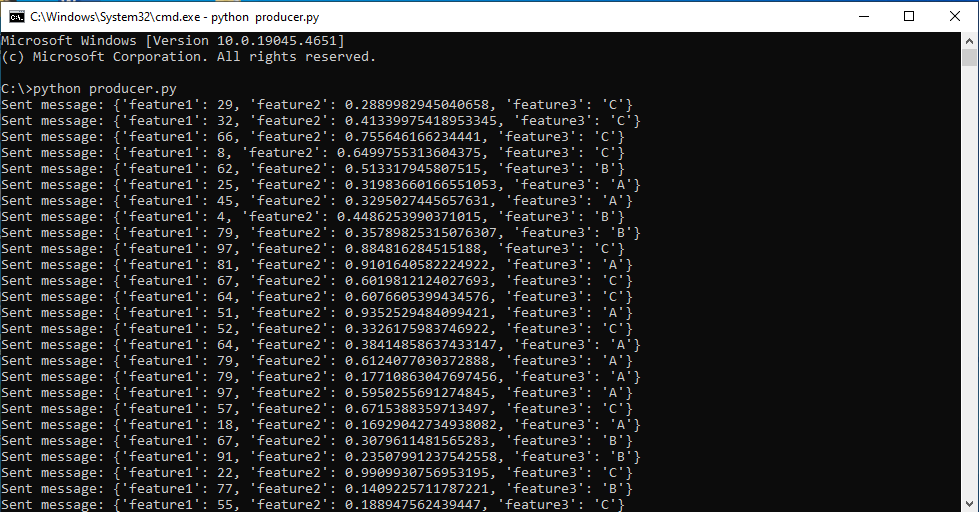
Every file plays an its rule same as this file is also important because this file is responsible for sending data to Kafka. Usually, the producer file is a separate script that puts data into Kafka topics for Spark to consume. It looks like a bridge.

1. **Features:**

* **Kafka**: It is used for sending and receiving data streams in **real-time.**
* **Spark Streaming**: It looks like a tool within **Spark** for processing data in real-time as it arrives.
* **Processing Data**: The script (Producer.py, spark\_streaming\_kafka.py) uses **Spark** to perform tasks like filtering, transforming, and analyzing the data from **Kafka**.

1. **Overall Working:**
2. **Producer:**

Sends data to Kafka.

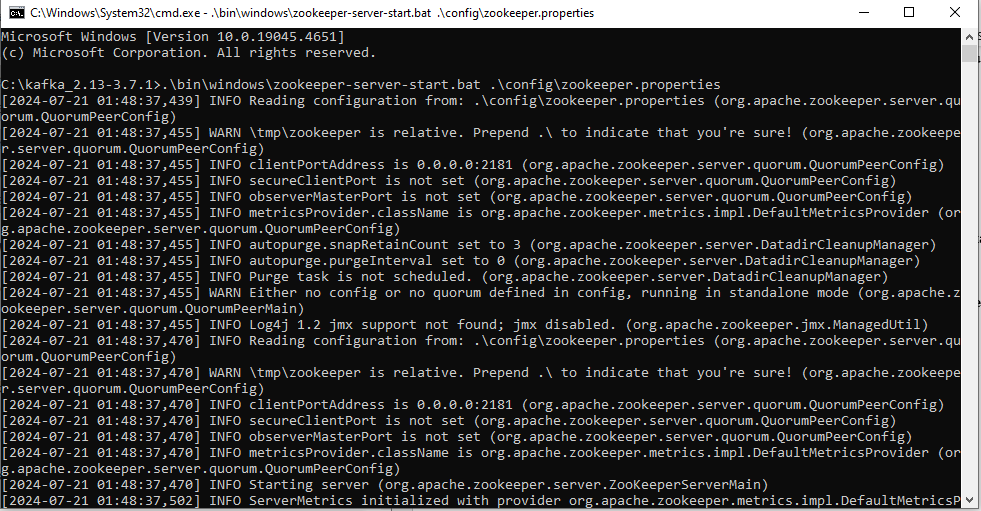


1. **Kafka:**

Stores and manages data streams. These 2 commands are run in Kafka folder

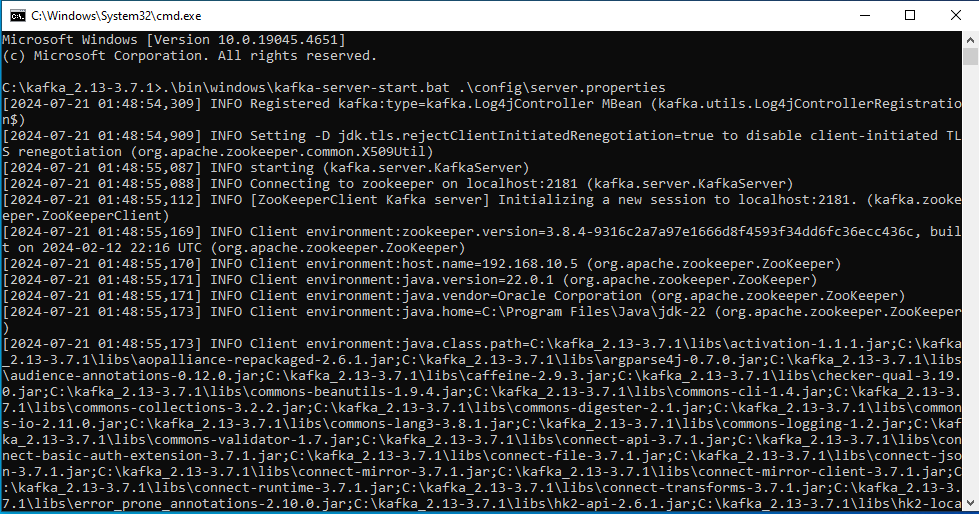
1. This command is used to start the zookeeper in the Kafka folder.

**.\bin\windows\zookeeper-server-start.bat .\config\zookeeper.properties**



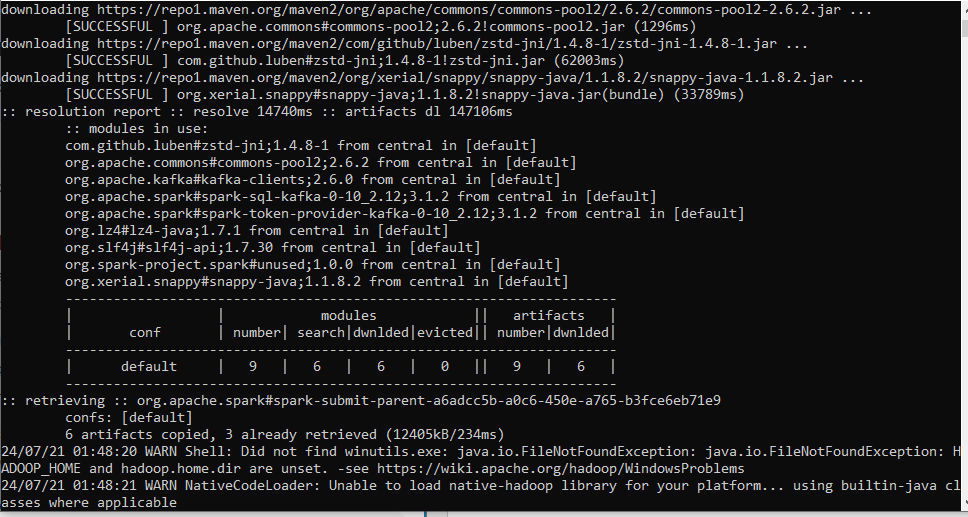
1. This command is used to start the Kafka server.

**.\bin\windows\kafka-server-start.bat .\config\server.properties**



1. **Spark Streaming:**

It will read data from Kafka because the producer script sends data to Kafka and spark streaming processes it according to the logic in spark\_streaming\_kafka.py script.



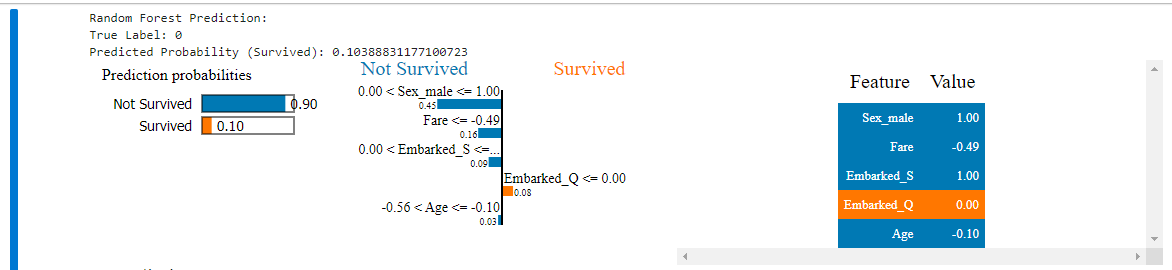
1. **Predictions:**

Basically, I had broken my model into two .pkl format models. One is rf\_model.pkl and another is svm\_model.pkl and then on these models I am using the predictions by using LIME.

First, I will load the models where my data is going to stored after the spark streaming then after that I will do some process on that because I have to balance all columns then I had used the LIME for explaining the Predictions.

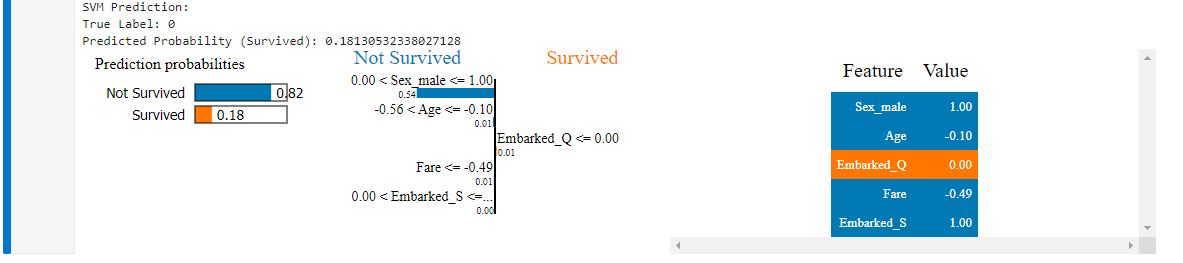
This is my first model predictions.

Random Forest Predictions:



This is my second model predictions.

SVM Predictions:



**Helping Material:**

<https://youtu.be/heXd6JA2TQc?si=5Rf5kDO3pYnOVCrh>

<https://youtu.be/lFox22RJE7s?si=P2XS9JaGRMuKx2Hx>

<https://youtu.be/heXd6JA2TQc?si=qFlwmpsfC22F5A9X>

<https://youtu.be/lFox22RJE7s?si=4_oufmfGZI1E2KaG>

<https://youtu.be/lFox22RJE7s?si=Dvy7fiy883eBm4Go>

<https://youtu.be/fATVVQfFyU0?si=sazfWdfU3JPitRpt>

<https://youtu.be/25WeAQyoHXc?si=6vN8e89hawEIG0e0>

<https://stackoverflow.com/questions/70374571/connecting-pyspark-with-kafka> (Forums)

<https://www.openlogic.com/blog/kafka-and-spark-streaming> (Forums)

Thanks.