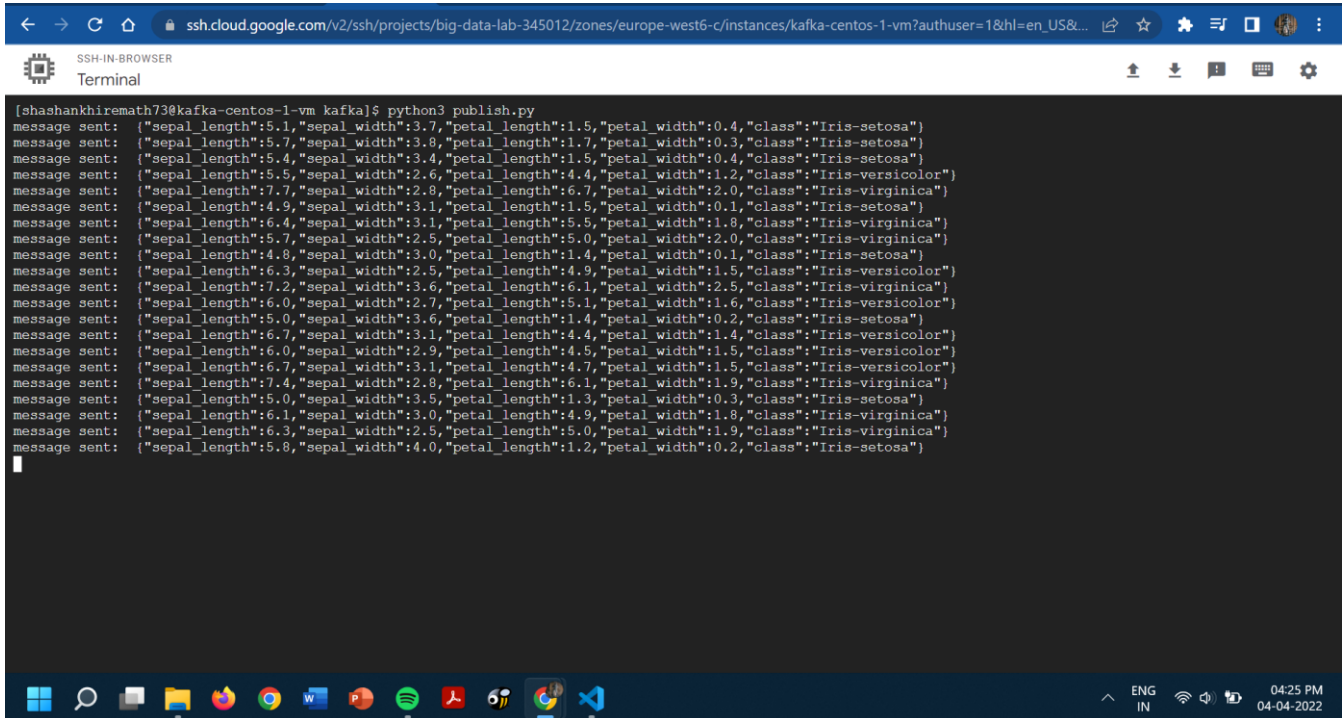


CS4830 Big Data Lab Assignment 5

Shashank H S

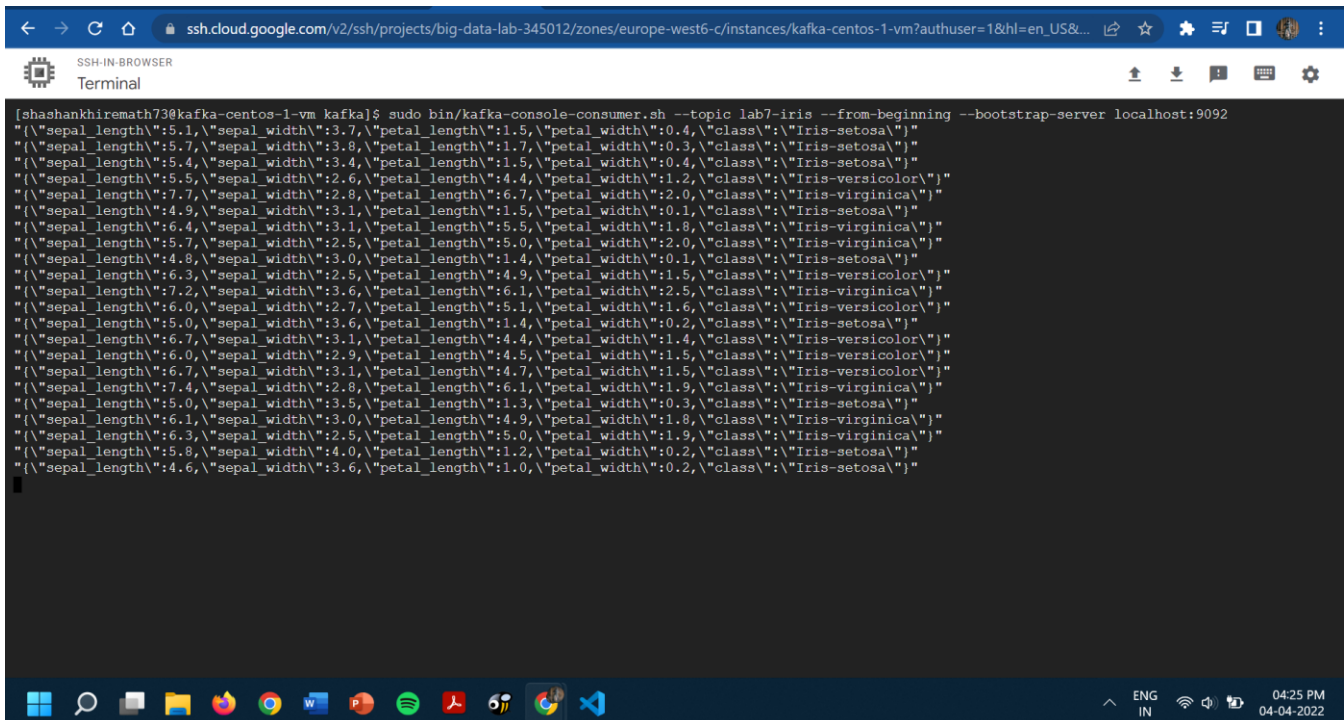
BE18B006

1. The screenshots show that publish.py is working as required.



```
[shashankhiremath73@kafka-centos-1-vm kafka]$ python3 publish.py
message sent: {"sepal_length":5.1,"sepal_width":3.7,"petal_length":1.5,"petal_width":0.4,"class":"Iris-setosa"}
message sent: {"sepal_length":5.7,"sepal_width":3.8,"petal_length":1.7,"petal_width":0.3,"class":"Iris-setosa"}
message sent: {"sepal_length":5.4,"sepal_width":3.4,"petal_length":1.5,"petal_width":0.4,"class":"Iris-setosa"}
message sent: {"sepal_length":5.5,"sepal_width":2.6,"petal_length":4.4,"petal_width":1.2,"class":"Iris-versicolor"}
message sent: {"sepal_length":7.7,"sepal_width":2.8,"petal_length":6.7,"petal_width":2.0,"class":"Iris-virginica"}
message sent: {"sepal_length":4.9,"sepal_width":3.1,"petal_length":1.5,"petal_width":0.1,"class":"Iris-setosa"}
message sent: {"sepal_length":6.4,"sepal_width":3.1,"petal_length":5.5,"petal_width":1.8,"class":"Iris-virginica"}
message sent: {"sepal_length":5.7,"sepal_width":2.5,"petal_length":5.0,"petal_width":2.0,"class":"Iris-virginica"}
message sent: {"sepal_length":4.8,"sepal_width":3.0,"petal_length":1.4,"petal_width":0.1,"class":"Iris-setosa"}
message sent: {"sepal_length":6.3,"sepal_width":2.5,"petal_length":4.9,"petal_width":1.5,"class":"Iris-versicolor"}
message sent: {"sepal_length":7.2,"sepal_width":3.6,"petal_length":6.1,"petal_width":2.5,"class":"Iris-virginica"}
message sent: {"sepal_length":6.0,"sepal_width":2.7,"petal_length":5.1,"petal_width":1.6,"class":"Iris-versicolor"}
message sent: {"sepal_length":5.0,"sepal_width":3.6,"petal_length":1.4,"petal_width":0.2,"class":"Iris-setosa"}
message sent: {"sepal_length":6.7,"sepal_width":3.1,"petal_length":4.4,"petal_width":1.4,"class":"Iris-versicolor"}
message sent: {"sepal_length":6.0,"sepal_width":2.9,"petal_length":4.5,"petal_width":1.5,"class":"Iris-versicolor"}
message sent: {"sepal_length":4.9,"sepal_width":3.1,"petal_length":4.7,"petal_width":1.5,"class":"Iris-versicolor"}
message sent: {"sepal_length":7.4,"sepal_width":2.8,"petal_length":6.1,"petal_width":1.9,"class":"Iris-virginica"}
message sent: {"sepal_length":5.0,"sepal_width":3.5,"petal_length":1.3,"petal_width":0.3,"class":"Iris-setosa"}
message sent: {"sepal_length":6.1,"sepal_width":3.0,"petal_length":4.9,"petal_width":1.8,"class":"Iris-virginica"}
message sent: {"sepal_length":6.3,"sepal_width":2.5,"petal_length":5.0,"petal_width":1.9,"class":"Iris-virginica"}
message sent: {"sepal_length":5.8,"sepal_width":4.0,"petal_length":1.2,"petal_width":0.2,"class":"Iris-setosa"}

```



```
[shashankhiremath73@kafka-centos-1-vm kafka]$ sudo bin/kafka-console-consumer.sh --topic lab7-iris --from-beginning --bootstrap-server localhost:9092
{"sepal_length":5.1,"sepal_width":3.7,"petal_length":1.5,"petal_width":0.4,"class":"Iris-setosa"}
{"sepal_length":5.7,"sepal_width":3.8,"petal_length":1.7,"petal_width":0.3,"class":"Iris-setosa"}
{"sepal_length":5.4,"sepal_width":3.4,"petal_length":1.5,"petal_width":0.4,"class":"Iris-setosa"}
{"sepal_length":5.5,"sepal_width":2.6,"petal_length":4.4,"petal_width":1.2,"class":"Iris-versicolor"}
{"sepal_length":7.7,"sepal_width":2.8,"petal_length":6.7,"petal_width":2.0,"class":"Iris-virginica"}
{"sepal_length":4.9,"sepal_width":3.1,"petal_length":1.5,"petal_width":0.1,"class":"Iris-setosa"}
{"sepal_length":6.4,"sepal_width":3.1,"petal_length":5.5,"petal_width":1.8,"class":"Iris-virginica"}
{"sepal_length":5.7,"sepal_width":2.5,"petal_length":5.0,"petal_width":2.0,"class":"Iris-virginica"}
{"sepal_length":4.8,"sepal_width":3.0,"petal_length":1.4,"petal_width":0.1,"class":"Iris-setosa"}
{"sepal_length":6.3,"sepal_width":2.5,"petal_length":4.9,"petal_width":1.5,"class":"Iris-versicolor"}
{"sepal_length":7.2,"sepal_width":3.6,"petal_length":6.1,"petal_width":2.5,"class":"Iris-virginica"}
{"sepal_length":6.0,"sepal_width":2.7,"petal_length":5.1,"petal_width":1.6,"class":"Iris-versicolor"}
{"sepal_length":5.0,"sepal_width":3.6,"petal_length":1.4,"petal_width":0.2,"class":"Iris-setosa"}
{"sepal_length":6.7,"sepal_width":3.1,"petal_length":4.4,"petal_width":1.4,"class":"Iris-versicolor"}
{"sepal_length":6.0,"sepal_width":2.9,"petal_length":4.5,"petal_width":1.5,"class":"Iris-versicolor"}
{"sepal_length":4.9,"sepal_width":3.1,"petal_length":4.7,"petal_width":1.5,"class":"Iris-versicolor"}
{"sepal_length":7.4,"sepal_width":2.8,"petal_length":6.1,"petal_width":1.9,"class":"Iris-virginica"}
{"sepal_length":5.0,"sepal_width":3.5,"petal_length":1.3,"petal_width":0.3,"class":"Iris-setosa"}
{"sepal_length":6.1,"sepal_width":3.0,"petal_length":4.9,"petal_width":1.8,"class":"Iris-virginica"}
{"sepal_length":6.3,"sepal_width":2.5,"petal_length":5.0,"petal_width":1.9,"class":"Iris-virginica"}
{"sepal_length":5.8,"sepal_width":4.0,"petal_length":1.2,"petal_width":0.2,"class":"Iris-setosa"}
{"sepal_length":4.6,"sepal_width":3.6,"petal_length":1.0,"petal_width":0.2,"class":"Iris-setosa"}

```

publish.py reads each row in iris.csv as a dataframe and dumps it using JSON. Then it encodes each row and publishes it to the topic irispred. This is decoded back in subscribe.py and converted from json to dataframe following the appropriate schema.

2. The random forest classifier trained on the iris dataset with 100 trees was saved in the bucket. The iris2.py file is in the zipped folder.

The image shows two screenshots from a Google Cloud Platform environment. The top screenshot is the 'Job details' page for a Dataproc job (ID: job-31f1f10b). The job status is 'Succeeded'. The output log shows a sequence of INFO messages from the Google Cloud Spark driver and executors, indicating the successful execution of a query and read session for the 'big-data-lab-345012.iris.irisdata' table. The bottom screenshot is a terminal window titled 'SSH-IN-BROWSER' showing the execution of Kafka commands. The user creates a topic 'lab7-kafka', then uses 'nano' to edit 'publish.py'. Finally, they use 'describe' to show the topic details: 'Topic: lab7-kafka', 'Partition: 0', 'Leader: 0', 'Replicas: 0', 'Isr: 0'.

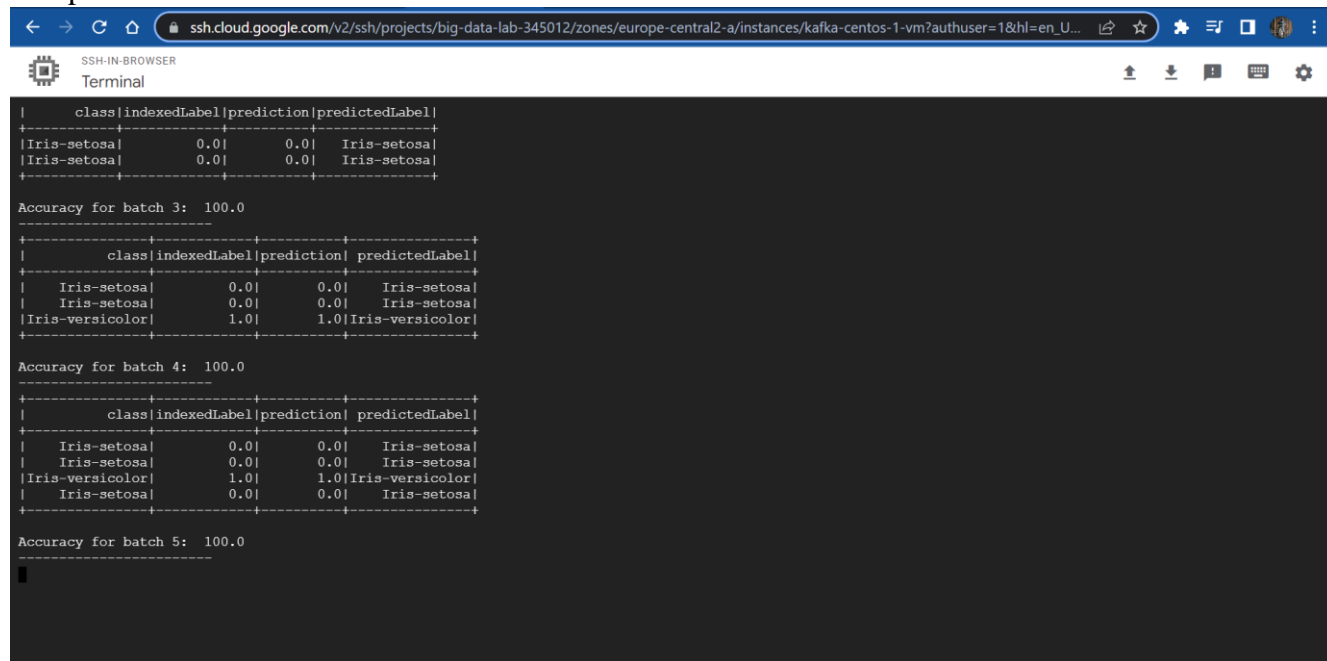
```
Job ID: job-31f1f10b
Job UUID: 4a14e11a-0f06-4287-b775-325b3433ef0d
Type: Dataproc Job
Status: Succeeded

Output
22/04/04 05:52:13 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table big-data-lab-345012.iris.irisdata, parameters se
22/04/04 05:52:13 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from big-data-lab-345012.iris.irisdata columns=[s
22/04/04 05:52:13 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'big-data-lab-345012.iris.irisda
22/04/04 05:52:30 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table big-data-lab-345012.iris.irisdata, parameters se
22/04/04 05:52:30 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from big-data-lab-345012.iris.irisdata columns=[s
22/04/04 05:52:30 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'big-data-lab-345012.iris.irisda
22/04/04 05:52:31 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Querying table big-data-lab-345012.iris.irisdata, parameters se
22/04/04 05:52:31 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Going to read from big-data-lab-345012.iris.irisdata columns=[s
22/04/04 05:52:31 INFO com.google.cloud.spark.bigquery.direct.DirectBigQueryRelation: Created read session for table 'big-data-lab-345012.iris.irisda
Evaluating on Test data: 1.0
Evaluator: MulticlassClassificationEvaluator_1c0a6c5f493
22/04/04 05:52:36 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFilesystem: Successfully repaired 'gs:/
22/04/04 05:52:40 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFilesystem: Successfully repaired 'gs:/
22/04/04 05:52:43 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFilesystem: Successfully repaired 'gs:/
22/04/04 05:52:46 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFilesystem: Successfully repaired 'gs:/
22/04/04 05:52:48 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFilesystem: Successfully repaired 'gs:/
22/04/04 05:52:51 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFilesystem: Successfully repaired 'gs:/
22/04/04 05:52:54 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFilesystem: Successfully repaired 'gs:/
22/04/04 05:52:56 INFO com.google.cloud.hadoop.repackaged.gcs.com.google.cloud.hadoop.gcsio.GoogleCloudStorageFilesystem: Successfully repaired 'gs:/
22/04/04 05:52:57 INFO org.sparkproject.jetty.server.AbstractConnector: Stopped Spark@38614ef6(HTTP/1.1, (http/1.1)){0.0.0.0:0}
```

```
Last login: Mon Apr 4 08:27:02 2022 from 35.235.240.114
[shashankhiremath73@kafka-centos-1-vm ~]$ cd /opt/kafka
[shashankhiremath73@kafka-centos-1-vm kafka]$ sudo bin/kafka-topics.sh --create --topic lab7-kafka --bootstrap-server localhost:9092
Created topic lab7-kafka.
[shashankhiremath73@kafka-centos-1-vm kafka]$ sudo nano publish.py
[shashankhiremath73@kafka-centos-1-vm kafka]$ sudo bin/kafka-topics.sh --describe --topic lab7-kafka --bootstrap-server localhost:9092
Topic: lab7-kafka    TopicId: 4XmghsOpQi6Y_wajtqP2ng PartitionCount: 1    ReplicationFactor: 1    Configs: segment.bytes=1073741824
        Topic: lab7-kafka    Partition: 0    Leader: 0    Replicas: 0    Isr: 0
[shashankhiremath73@kafka-centos-1-vm kafka]$
```

The topic was created on the kafka server and tested. Next, the subscribe.py file was created which gives real-time predictions for each row in the iris.csv file.

The pictures are attached below.



SSH-IN-BROWSER Terminal

```
| class|indexedLabel|prediction|predictedLabel|
+-----+-----+-----+-----+
|Iris-setosa|    0.0|    0.0|  Iris-setosa|
|Iris-setosa|    0.0|    0.0|  Iris-setosa|
+-----+-----+-----+-----+

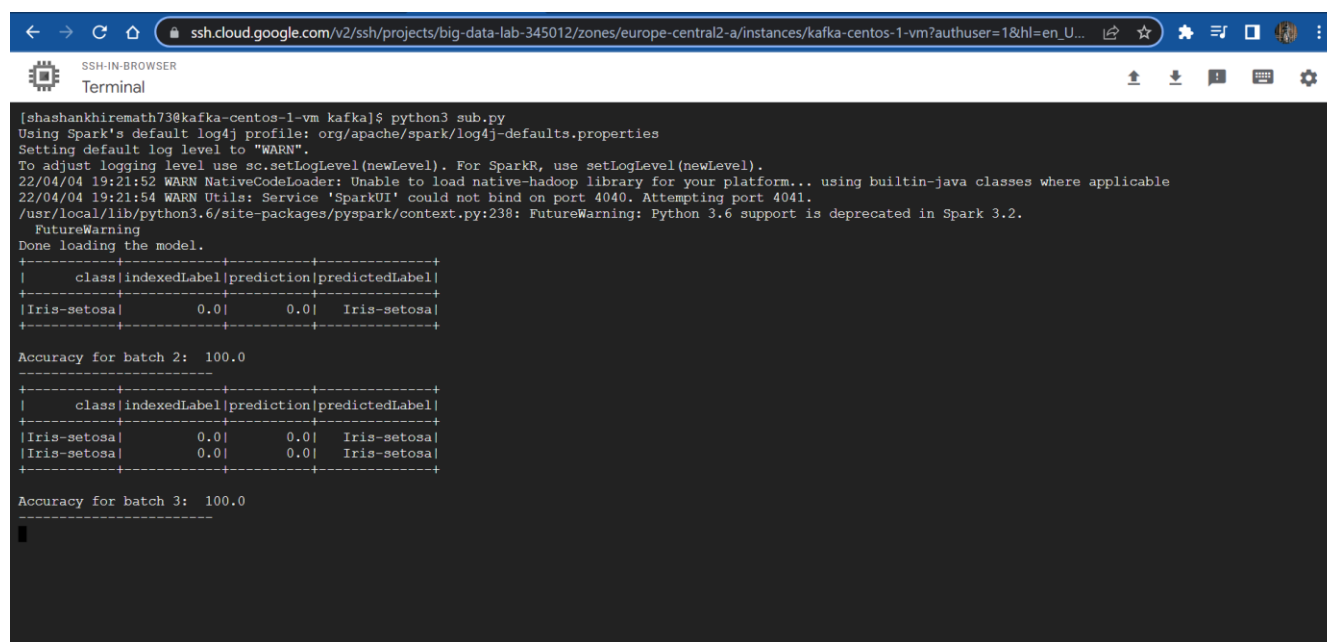
Accuracy for batch 3: 100.0

| class|indexedLabel|prediction| predictedLabel|
+-----+-----+-----+-----+
|  Iris-setosa|    0.0|    0.0|  Iris-setosa|
|  Iris-setosa|    0.0|    0.0|  Iris-setosa|
|Iris-versicolor|    1.0|    1.0|Iris-versicolor|
+-----+-----+-----+-----+

Accuracy for batch 4: 100.0

| class|indexedLabel|prediction| predictedLabel|
+-----+-----+-----+-----+
|  Iris-setosa|    0.0|    0.0|  Iris-setosa|
|  Iris-setosa|    0.0|    0.0|  Iris-setosa|
|Iris-versicolor|    1.0|    1.0|Iris-versicolor|
|  Iris-setosa|    0.0|    0.0|  Iris-setosa|
+-----+-----+-----+-----+

Accuracy for batch 5: 100.0
```



SSH-IN-BROWSER Terminal

```
[shashankhiremath73@kafka-centos-1-vm kafka]$ python3 sub.py
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/04/04 19:21:52 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/04/04 19:21:54 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
/usr/local/lib/python3.6/site-packages/pyspark/context.py:238: FutureWarning: Python 3.6 support is deprecated in Spark 3.2.
  FutureWarning
Done loading the model.

| class|indexedLabel|prediction|predictedLabel|
+-----+-----+-----+-----+
|Iris-setosa|    0.0|    0.0|  Iris-setosa|
+-----+-----+-----+-----+

Accuracy for batch 2: 100.0

| class|indexedLabel|prediction|predictedLabel|
+-----+-----+-----+-----+
|Iris-setosa|    0.0|    0.0|  Iris-setosa|
|Iris-setosa|    0.0|    0.0|  Iris-setosa|
+-----+-----+-----+-----+

Accuracy for batch 3: 100.0
```

The RF classifier with 100 trees has not misclassified any data points (after several batches, picture not shown).