# Programming Assignment 2

**ALAA SALIH as422**

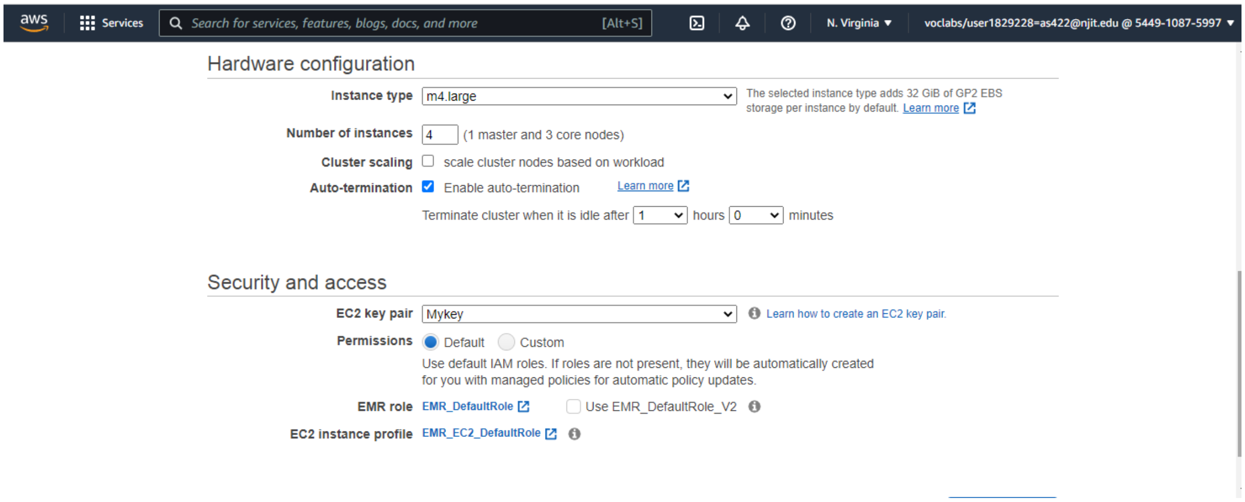
**CS643**

https://github.com/Alaa422/Programming-Assignment-2

**Goal:** The purpose of this individual assignment is to learn how to develop parallel machine learning (ML) applications in Amazon AWS cloud platform. Specifically, you will learn: (1) how to use Apache Spark to train an ML model in parallel on multiple EC2 instances; (2) how to use Spark's MLlib to develop and use an ML model in the cloud; (3) How to use  Docker to create a container for your ML model to simplify model deployment.

- **Setting cloud Environment For Model Creation and Training**

1. Login into AWS Account
2. Search for EMR in service ,then Open EMR
3. Click Create Cluster
4. Required details of configurations: General Configuration for Cluster Name type desired cluster name. Under Software configuration in the application column click the button which shows Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.8.2. Under Hardware Configuration click m4.large rather than m5.xlarge.
5. Select 4 instances under the column Number of instances. (1 master and 3 core nodes). Then Under Security and access choose your key that you created in Create an Amazon EC2 Key Pair (mykey)for SSH.


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
N. Virginia ▾
voclabs/user1829228=as422@njit.edu @ 5449-1087-5997 ▾
Learn more
Learn more
Learn how to create an EC2 key pair.

Hardware configuration

Instance type: m4.large
The selected instance type adds 32 GiB of GP2 EBS storage per instance by default.

Number of instances: 4 (1 master and 3 core nodes)

Cluster scaling: □ scale cluster nodes based on workload

Auto-termination: ☑ Enable auto-termination

Terminate cluster when it is idle after 1 hours 0 minutes

Security and access

EC2 key pair: Mykey

Permissions: ● Default ○ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role: EMR_DefaultRole □ Use EMR_DefaultRole_V2

EC2 instance profile: EMR_EC2_DefaultRole

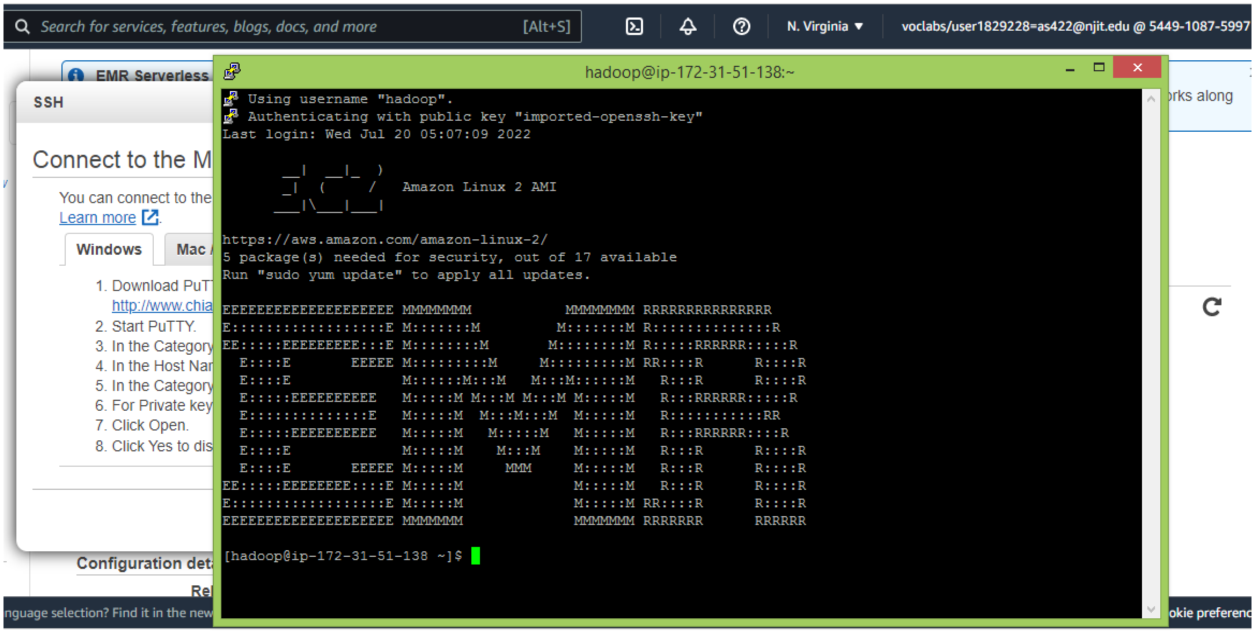
6. Cluster created. The status should change from Starting to Running to Waiting during the cluster creation process.

7. When the status progresses to Waiting, your cluster is up, running, and ready to accept work.

8. We must connect to the master node of EMR cluster, I am using SSH with Putty from my desktop on windows.

   - Download PuTTY.exe to your computer from: http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
   - Start PuTTY.
   - In the Category list, click Session.
   - In the Host Name field, type your host-name (see the picture)
   - In the Category list, expand Connection > SSH, and then click Auth.
   - For Private key file for authentication, click Browse and select the private key file (Mykey.ppk) used to launch the cluster.
   - Click Open.
   - Click Yes to dismiss the security alert.

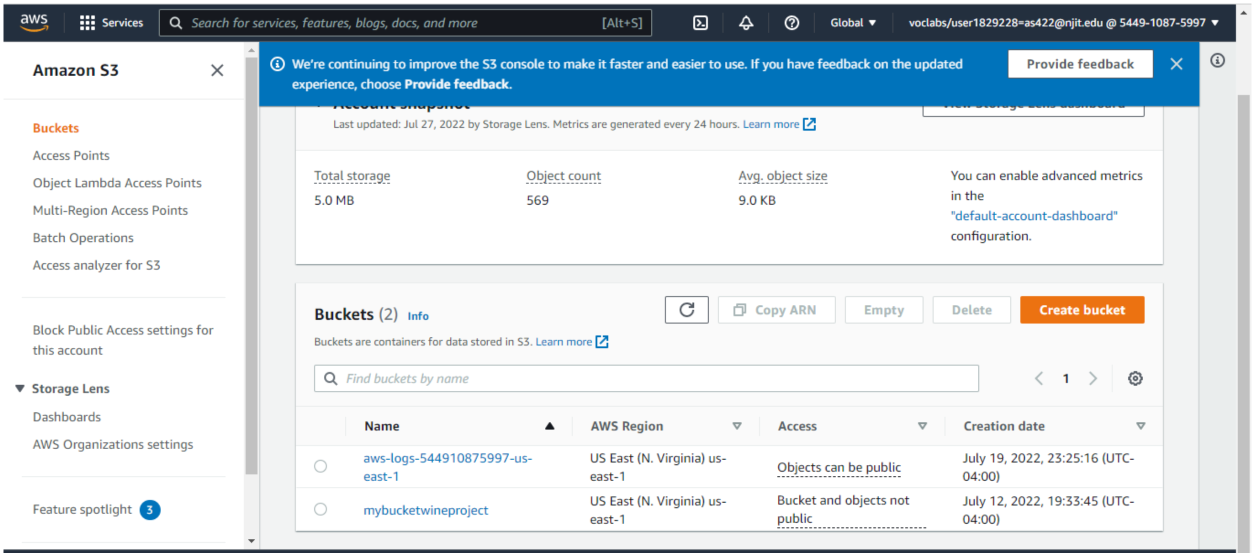
9. The screenshot shows the connection to the master node.



```
Using username "hadoop".
Authenticating with public key "imported-openssh-key"
Last login: Wed Jul 20 05:07:09 2022

       __|  __|_  )
       _|  (     /   Amazon Linux 2 AMI
      ___|\___|___|

https://aws.amazon.com/amazon-linux-2/
5 package(s) needed for security, out of 17 available
Run "sudo yum update" to apply all updates.

EEEEEEEEEEEEEEEEEEEE MMMMMMMM           MMMMMMMM RRRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE::::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-51-138 ~]$
```

- **Creation of S3 bucket**

To upload the training and validation datasets and model output.
Command :  aws s3 cp s3://bucket_name/file_name
My bucketname is= mybucketwineproject


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global
voclabs/user1829228=as422@njit.edu @ 5449-1087-5997
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
Provide feedback
Copy ARN
Empty
Delete
Create bucket
Find buckets by name
1

We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Last updated: Jul 27, 2022 by Storage Lens. Metrics are generated every 24 hours. Learn more

| Total storage | Object count | Avg. object size | |
|---|---|---|---|
| 5.0 MB | 569 | 9.0 KB | You can enable advanced metrics in the "default-account-dashboard" configuration. |

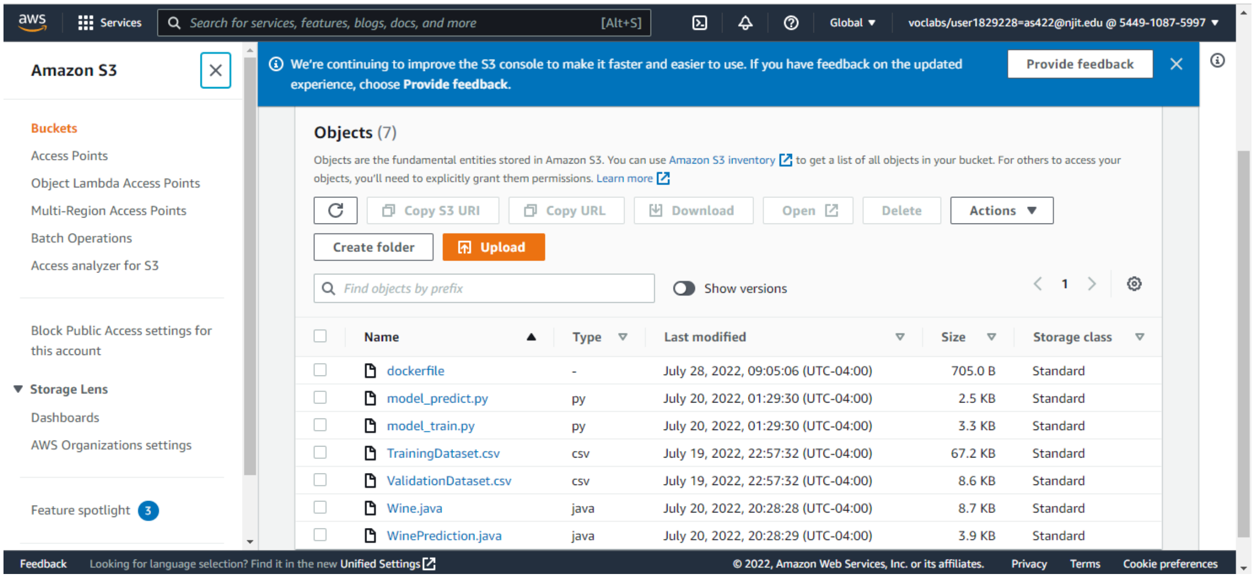
Buckets (2) Info

Buckets are containers for data stored in S3. Learn more

| | Name | AWS Region | Access | Creation date |
|---|---|---|---|---|
| ○ | aws-logs-544910875997-us-east-1 | US East (N. Virginia) us-east-1 | Objects can be public | July 19, 2022, 23:25:16 (UTC-04:00) |
| ○ | mybucketwineproject | US East (N. Virginia) us-east-1 | Bucket and objects not public | July 12, 2022, 19:33:45 (UTC-04:00) |

Now, upload the files:

Into the bucket, click on upload button


aws
Services
Search for services, features, blogs, docs, and more
[Alt+S]
Global
voclabs/user1829228=as422@njit.edu @ 5449-1087-5997
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
Provide feedback
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Show versions
1
Feedback
Looking for language selection? Find it in the new Unified Settings
© 2022, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

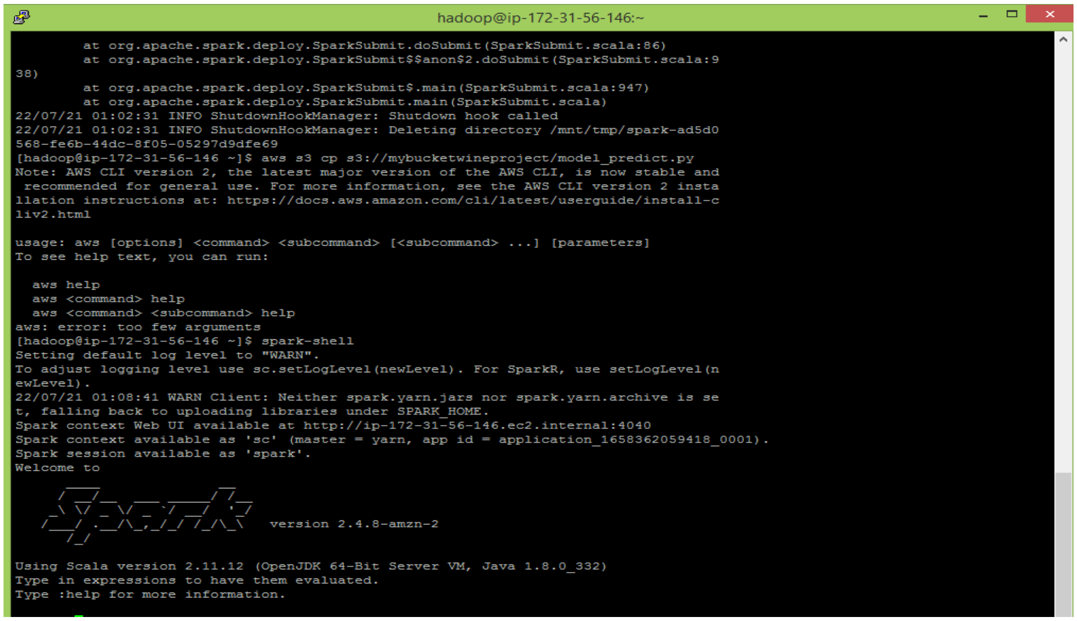
We're continuing to improve the S3 console to make it faster and easier to use. If you have feedback on the updated experience, choose **Provide feedback**.

Objects (7)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

| □ | Name | Type | Last modified | Size | Storage class |
|---|---|---|---|---|---|
| □ | dockerfile | - | July 28, 2022, 09:05:06 (UTC-04:00) | 705.0 B | Standard |
| □ | model_predict.py | py | July 20, 2022, 01:29:30 (UTC-04:00) | 2.5 KB | Standard |
| □ | model_train.py | py | July 20, 2022, 01:29:30 (UTC-04:00) | 3.3 KB | Standard |
| □ | TrainingDataset.csv | csv | July 19, 2022, 22:57:32 (UTC-04:00) | 67.2 KB | Standard |
| □ | ValidationDataset.csv | csv | July 19, 2022, 22:57:32 (UTC-04:00) | 8.6 KB | Standard |
| □ | Wine.java | java | July 20, 2022, 20:28:28 (UTC-04:00) | 8.7 KB | Standard |
| □ | WinePrediction.java | java | July 20, 2022, 20:28:29 (UTC-04:00) | 3.9 KB | Standard |

## Application Prediction:



```
        at org.apache.spark.deploy.SparkSubmit.doSubmit(SparkSubmit.scala:86)
        at org.apache.spark.deploy.SparkSubmit$$anon$2.doSubmit(SparkSubmit.scala:9
38)
        at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:947)
        at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)
22/07/21 01:02:31 INFO ShutdownHookManager: Shutdown hook called
22/07/21 01:02:31 INFO ShutdownHookManager: Deleting directory /mnt/tmp/spark-ad5d0
568-fe6b-44dc-8f05-05297d9dfe69
[hadoop@ip-172-31-56-146 ~]$ aws s3 cp s3://mybucketwineproject/model_predict.py
Note: AWS CLI version 2, the latest major version of the AWS CLI, is now stable and
 recommended for general use. For more information, see the AWS CLI version 2 insta
llation instructions at: https://docs.aws.amazon.com/cli/latest/userguide/install-c
liv2.html

usage: aws [options] <command> <subcommand> [<subcommand> ...] [parameters]
To see help text, you can run:

  aws help
  aws <command> help
  aws <command> <subcommand> help
aws: error: too few arguments
[hadoop@ip-172-31-56-146 ~]$ spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(n
ewLevel).
22/07/21 01:08:41 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is se
t, falling back to uploading libraries under SPARK_HOME.
Spark context Web UI available at http://ip-172-31-56-146.ec2.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_1658362059418_0001).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.8-amzn-2
      /_/

Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_332)
Type in expressions to have them evaluated.
Type :help for more information.
```

- **Execute the model training**:

Copy from S3 to the master node the JAR file using the following command:

aws s3 cp s3:// mybucketwineproject/WineQualityPrediction.jar .

Then

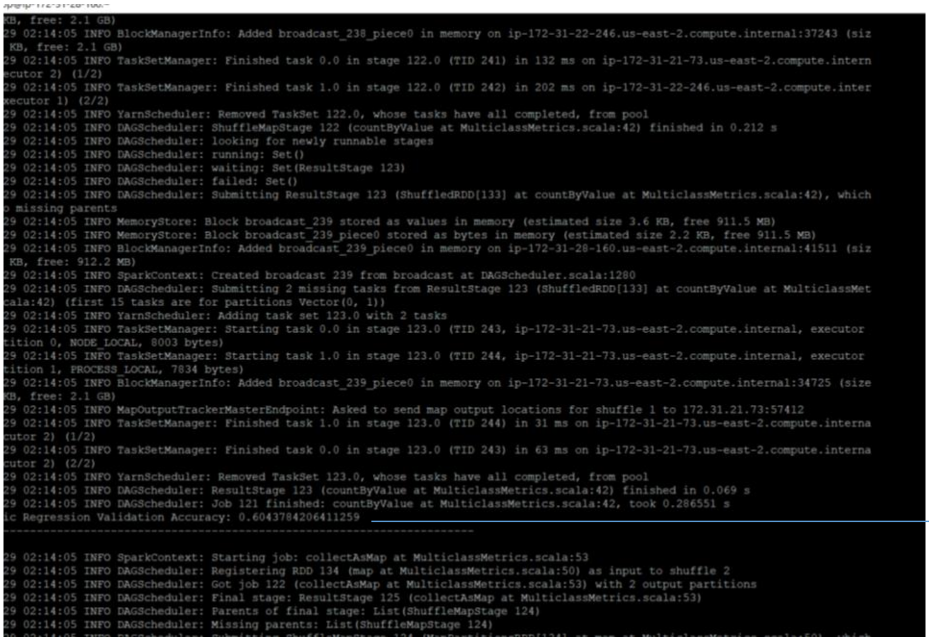
- spark-submit --master yarn  --class com.test.spark.Wine WineQualityPrediction.jar

- **The Results:**

The accuracy of model using Logistic Regression Validation Accuracy:0.60

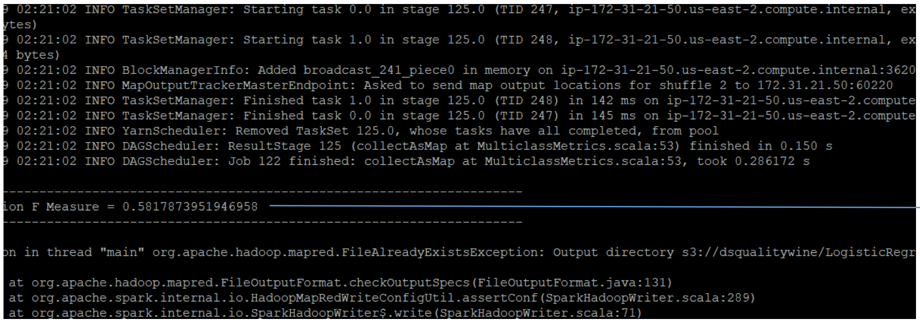
And Validation F1 score Measure is 0.58

Shown below



```
p@ip-172-31-28-160:~
KB, free: 2.1 GB)
29 02:14:05 INFO BlockManagerInfo: Added broadcast_238_piece0 in memory on ip-172-31-22-246.us-east-2.compute.internal:37243 (siz
 KB, free: 2.1 GB)
29 02:14:05 INFO TaskSetManager: Finished task 0.0 in stage 122.0 (TID 241) in 132 ms on ip-172-31-21-73.us-east-2.compute.intern
ecutor 2) (1/2)
29 02:14:05 INFO TaskSetManager: Finished task 1.0 in stage 122.0 (TID 242) in 202 ms on ip-172-31-22-246.us-east-2.compute.inter
xecutor 1) (2/2)
29 02:14:05 INFO YarnScheduler: Removed TaskSet 122.0, whose tasks have all completed, from pool
29 02:14:05 INFO DAGScheduler: ShuffleMapStage 122 (countByValue at MulticlassMetrics.scala:42) finished in 0.212 s
29 02:14:05 INFO DAGScheduler: looking for newly runnable stages
29 02:14:05 INFO DAGScheduler: running: Set()
29 02:14:05 INFO DAGScheduler: waiting: Set(ResultStage 123)
29 02:14:05 INFO DAGScheduler: failed: Set()
29 02:14:05 INFO DAGScheduler: Submitting ResultStage 123 (ShuffledRDD[133] at countByValue at MulticlassMetrics.scala:42), which
o missing parents
29 02:14:05 INFO MemoryStore: Block broadcast_239 stored as values in memory (estimated size 3.6 KB, free 911.5 MB)
29 02:14:05 INFO MemoryStore: Block broadcast_239_piece0 stored as bytes in memory (estimated size 2.2 KB, free 911.5 MB)
29 02:14:05 INFO BlockManagerInfo: Added broadcast_239_piece0 in memory on ip-172-31-28-160.us-east-2.compute.internal:41511 (siz
 KB, free: 912.2 MB)
29 02:14:05 INFO SparkContext: Created broadcast 239 from broadcast at DAGScheduler.scala:1280
29 02:14:05 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 123 (ShuffledRDD[133] at countByValue at MulticlassMet
cala:42) (first 15 tasks are for partitions Vector(0, 1))
29 02:14:05 INFO YarnScheduler: Adding task set 123.0 with 2 tasks
29 02:14:05 INFO TaskSetManager: Starting task 0.0 in stage 123.0 (TID 243, ip-172-31-21-73.us-east-2.compute.internal, executor
tition 0, NODE_LOCAL, 8003 bytes)
29 02:14:05 INFO TaskSetManager: Starting task 1.0 in stage 123.0 (TID 244, ip-172-31-21-73.us-east-2.compute.internal, executor
tition 1, PROCESS_LOCAL, 7834 bytes)
29 02:14:05 INFO BlockManagerInfo: Added broadcast_239_piece0 in memory on ip-172-31-21-73.us-east-2.compute.internal:34725 (size
KB, free: 2.1 GB)
29 02:14:05 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 1 to 172.31.21.73:57412
29 02:14:05 INFO TaskSetManager: Finished task 1.0 in stage 123.0 (TID 244) in 31 ms on ip-172-31-21-73.us-east-2.compute.interna
cutor 2) (1/2)
29 02:14:05 INFO TaskSetManager: Finished task 0.0 in stage 123.0 (TID 243) in 63 ms on ip-172-31-21-73.us-east-2.compute.interna
cutor 2) (2/2)
29 02:14:05 INFO YarnScheduler: Removed TaskSet 123.0, whose tasks have all completed, from pool
29 02:14:05 INFO DAGScheduler: ResultStage 123 (countByValue at MulticlassMetrics.scala:42) finished in 0.069 s
29 02:14:05 INFO DAGScheduler: Job 121 finished: countByValue at MulticlassMetrics.scala:42, took 0.286551 s
ic Regression Validation Accuracy: 0.6043784206411259
----------------------------------------------------------------------

29 02:14:05 INFO SparkContext: Starting job: collectAsMap at MulticlassMetrics.scala:53
29 02:14:05 INFO DAGScheduler: Registering RDD 134 (map at MulticlassMetrics.scala:50) as input to shuffle 2
29 02:14:05 INFO DAGScheduler: Got job 122 (collectAsMap at MulticlassMetrics.scala:53) with 2 output partitions
29 02:14:05 INFO DAGScheduler: Final stage: ResultStage 125 (collectAsMap at MulticlassMetrics.scala:53)
29 02:14:05 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 124)
29 02:14:05 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 124)
```

Accuracy 0.604 =60.4%



```
9 02:21:02 INFO TaskSetManager: Starting task 0.0 in stage 125.0 (TID 247, ip-172-31-21-50.us-east-2.compute.internal, ex
ytes)
9 02:21:02 INFO TaskSetManager: Starting task 1.0 in stage 125.0 (TID 248, ip-172-31-21-50.us-east-2.compute.internal, ex
4 bytes)
9 02:21:02 INFO BlockManagerInfo: Added broadcast_241_piece0 in memory on ip-172-31-21-50.us-east-2.compute.internal:3620
9 02:21:02 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 2 to 172.31.21.50:60220
9 02:21:02 INFO TaskSetManager: Finished task 1.0 in stage 125.0 (TID 248) in 142 ms on ip-172-31-21-50.us-east-2.compute
9 02:21:02 INFO TaskSetManager: Finished task 0.0 in stage 125.0 (TID 247) in 145 ms on ip-172-31-21-50.us-east-2.compute
9 02:21:02 INFO YarnScheduler: Removed TaskSet 125.0, whose tasks have all completed, from pool
9 02:21:02 INFO DAGScheduler: ResultStage 125 (collectAsMap at MulticlassMetrics.scala:53) finished in 0.150 s
9 02:21:02 INFO DAGScheduler: Job 122 finished: collectAsMap at MulticlassMetrics.scala:53, took 0.286172 s

----------------------------------------------------------------------
ion F Measure = 0.5817873951946958
----------------------------------------------------------------------

on in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory s3://dsqualitywine/LogisticRegr

 at org.apache.hadoop.mapred.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:131)
 at org.apache.spark.internal.io.HadoopMapRedWriteConfigUtil.assertConf(SparkHadoopWriter.scala:289)
 at org.apache.spark.internal.io.SparkHadoopWriter$.write(SparkHadoopWriter.scala:71)
```

F1 measure=0.5817

- The results of model trained in EMR is stored in my s3 bucket
- So in order to pass it to my prediction model in Ec2, we download/sync s3 bucket data using aws cli
    - **Run The Application Prediction Without Docker:**

Execute the model using :

spark-submit --master yarn  --class com.test.spark.Wine WineQualityPrediction.jar "s3://mybucketwineproject/ValidationDataset.csv"

"s3:// mybucketwineproject /LogisticRegressionModel/"

- **Docker Installation/With Docker: (Have problem with this)**

I coudnt scucced to work with Docker , its first time to work with this but as I learn online , I think we have to create a Docker container then Execute the following command: This should be executed from the same folder where you have the Dockerfile created

Docker build –t docker-ml-model -f dockerfile

Docker run docker-ml-model

## Repository Link

https://github.com/Alaa422/Programming-Assignment-2