Cloud Computing

Programming Assignment 2

Name – Soumilee Ghosh

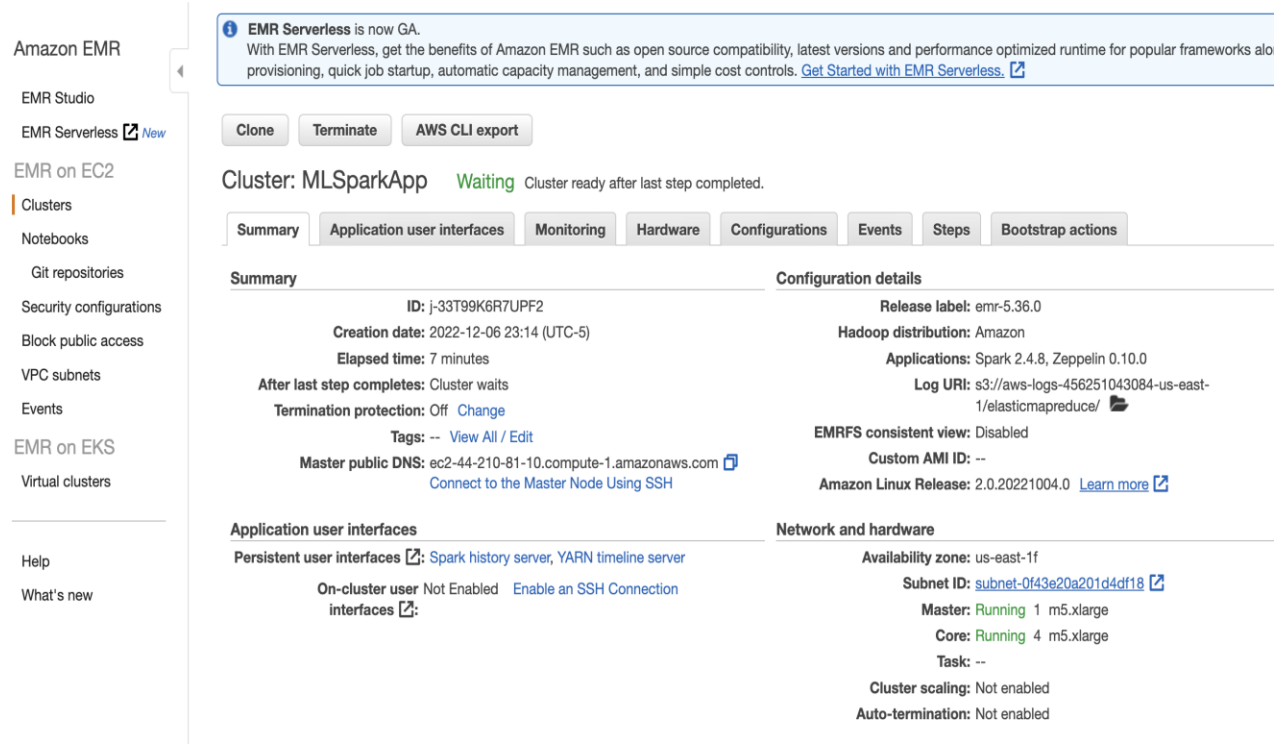
Ucid - sg342

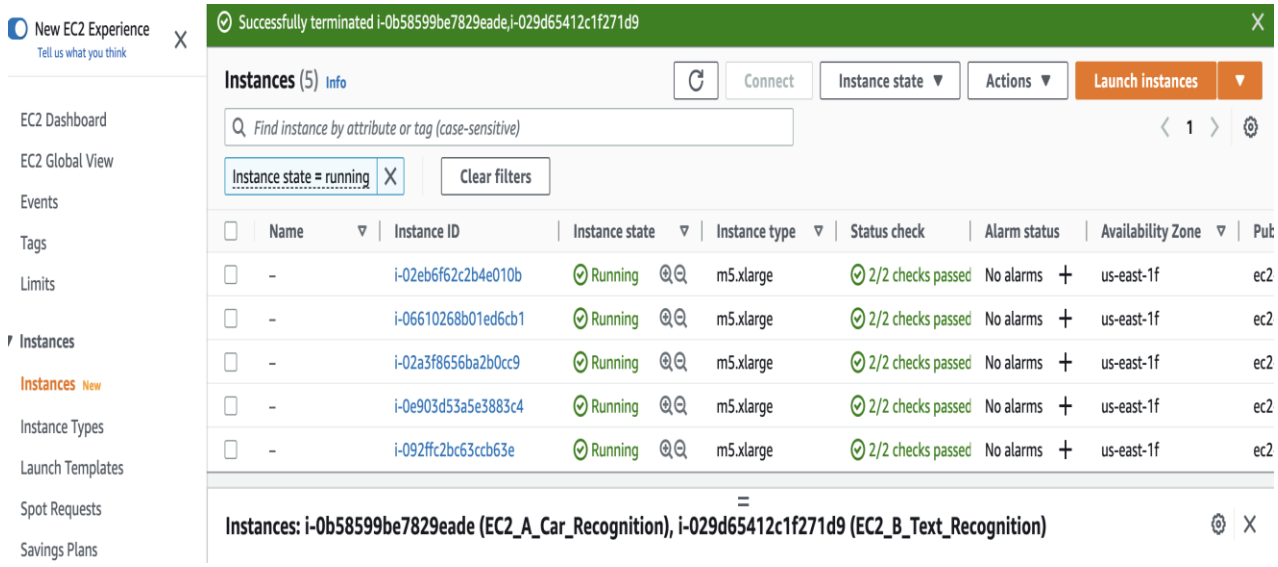
Goal -

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.

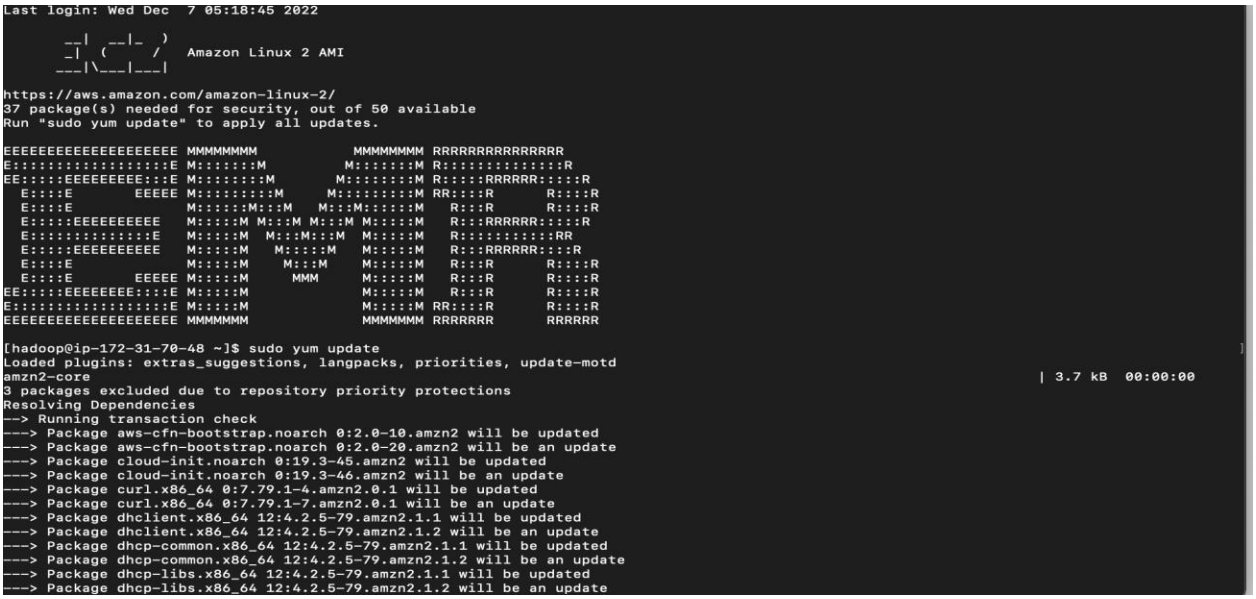
Steps Followed -

1. Create an EMR Cluster in the AWS dashboard under the analytics section click EMR now Click Create Cluster.
2. In the General Configuration for Cluster Name type desired cluster name. Under Software configuration` in the application column click the button which shows `Spark: Spark 2.4.8 and Zeppelin 0.10.0
3. Under Hardware Configuration click select 5 instances under the column Number of instances and disable auto termination. We are selecting 5 here so that there can be 1 master instance and 4 slave instances under it.
4. Under Security and access click the EC2 key pair already created else create a new one for this. Create .pem key for windows or .ppk for mac.
5. After all this enter create cluster. Go to EC2 dashboard and you will find 5 new instances created.

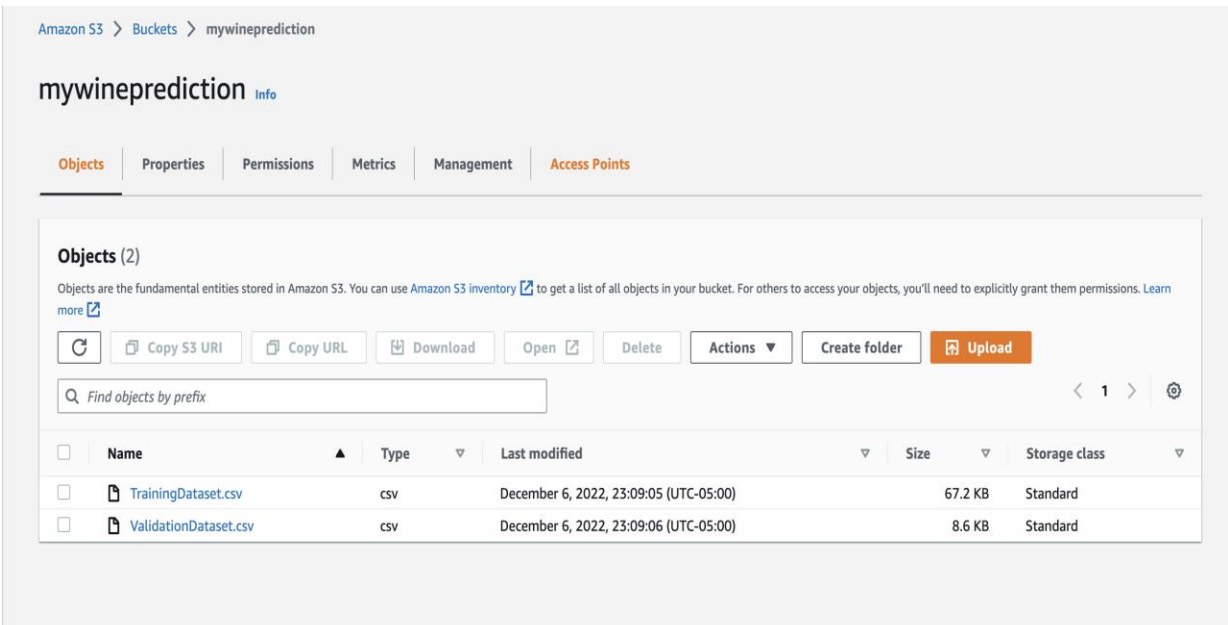




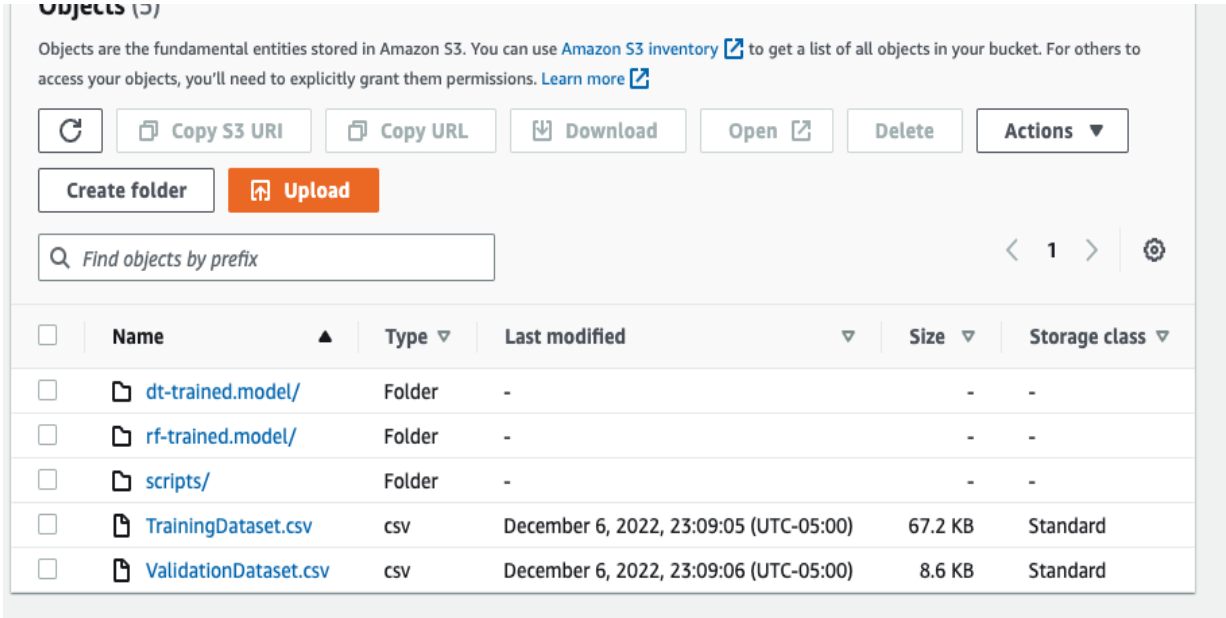
6. Go to the first instance and select security and go to inbound rules and select add rule. Select SSH and MyIP and add the rule.
7. Go to EMR dashboard and select your cluster and click on connect to master node using SSH. Follow the steps according to your OS and you will get it connected and running.



8. Create S3 bucket for storing the dataset and model output



9. After connecting to EMR using SSH run the job by executing the following command in EMR terminal. My model expects three parameters: 1. Dataset location 2. Output location to save the data 3. Output location to write fitness file. Using decision tree classifier and random forest classifier for training the data spark-submit model_training.py s3://mywineprediction/ValidationDataset.csv s3://mywineprediction/
10. Output of the models will be stored in S3 bucket (these will be later used in prediction)



11. Model Training in EMR -

```
Starting Spark Connection
22/12/07 06:00:34 INFO SparkContext: Running Spark version 2.4.8-amzn-2
22/12/07 06:00:34 INFO SparkContext: Submitted application: WineQuality-Training
22/12/07 06:00:34 INFO SecurityManager: Changing view acls to: root
22/12/07 06:00:34 INFO SecurityManager: Changing modify acls to: root
22/12/07 06:00:34 INFO SecurityManager: Changing view acls groups to:
22/12/07 06:00:34 INFO SecurityManager: Changing modify acls groups to:
22/12/07 06:00:34 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(root);
groups with view permissions: Set(); users  with modify permissions: Set(root); groups with modify permissions: Set()
22/12/07 06:00:34 INFO Utils: Successfully started service 'sparkDriver' on port 43661.
22/12/07 06:00:34 INFO SparkEnv: Registering MapOutputTracker
22/12/07 06:00:34 INFO SparkEnv: Registering BlockManagerMaster
22/12/07 06:00:34 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/12/07 06:00:34 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/12/07 06:00:34 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-3767bac0-155b-4544-9673-59c7173628f6
22/12/07 06:00:34 INFO MemoryStore: MemoryStore started with capacity 912.3 MB
22/12/07 06:00:34 INFO SparkEnv: Registering OutputCommitCoordinator
22/12/07 06:00:35 INFO Utils: Successfully started service 'SparkUI' on port 4040.
22/12/07 06:00:35 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://ip-172-31-70-48.ec2.internal:4040
22/12/07 06:00:35 INFO Executor: Starting executor ID driver on host localhost
22/12/07 06:00:35 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 39163.
22/12/07 06:00:35 INFO NettyBlockTransferService: Server created on ip-172-31-70-48.ec2.internal:39163
22/12/07 06:00:35 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
22/12/07 06:00:35 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-70-48.ec2.internal, 39163, None)
22/12/07 06:00:35 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-70-48.ec2.internal:39163 with 912.3 MB RAM, BlockManag
erId(driver, ip-172-31-70-48.ec2.internal, 39163, None)
22/12/07 06:00:35 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-172-31-70-48.ec2.internal, 39163, None)
22/12/07 06:00:35 INFO BlockManager: external shuffle service port = 7337
22/12/07 06:00:35 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-70-48.ec2.internal, 39163, None)
22/12/07 06:00:36 INFO SingleEventLogFileWriter: Logging events to hdfs:/var/log/spark/apps/local-1670392835422.inprogress
Loading Training data from s3://mywineprediction/TrainingDataset.csv ..
===================Decision Tree Classifier model===================
Decision tree trained model Location s3://mywineprediction/dt-trained.model ..
Evaluate the trained model...
Accuracy = 0.9827990617670055
Test Error = 0.01720093823299451
Decision Tree F1-Score = 0.9747156604496178
===================Random Forest model===================
Random forest trained model Location : s3://mywineprediction/rf-trained.model ..
Accuracy = 0.9921813917122753
Test Error = 0.007818608287724738
Random Forest F1-Score = 0.9903956890501021
[hadoop@ip-172-31-70-48 wine-prediction]$
```

Trained Model -

The training model I have used is decision tree classifier and random forest classifier. For decision tree the results I have got are -

Accuracy: 0.982799

Test Error: 0.017

Decision Tree F-1 score: 0.974715

For Random Forest classifier I have got -

Accuracy: 0.992181

Test Error: 0.00781

Random Forest F-1 score: 0.99039

So, in conclusion Random Forest is a better model as it is giving an accuracy of 99.2% with 0.7% error rate and F1 score of 99.0

Prediction Model -

12. Create EC2 Instance for Prediction Ubuntu instance is launched as follows: Go to EC2 dashboard and click on "Launch instances”. Select Ubuntu machine images. In Choose an Instance type select "t3.medium" and click on "Review and Launch”. Click on "Launch" Create a new key pair or choose an existing one and click on "Launch".
13. Setup Spark environment -

Install python 3.7 as 3.10 is not supported with spark version 2.4.8 -

Sudo apt –y update

Sudo add-apt-repository –y ppa:deadsnakes/ppa

Sudo apt-get update

Sudo apt-get install python 3.7

Make python 3.7 default python location

Sudo update–alternatives --install usr/bin/python3 python3 /usr/bin/python3.10 1

Sudo update–alternatives --install usr/bin/python3 python3 /usr/bin/python3.7 2

Install pip -

Sudo apt install python3-pip

sudo apt-get install python3.7-distutils

Install java 1.8 -

Sudo apt-get install openjdk-8-jdk

Install py4j -

Pip install py4j

Install Numpy -

Pip install Numpy

Install pandas -

Pip install pandas

Install spark and hadoop -

wget https://archive.apache.org/dist/spark/spark-2.4.8/spark-2.4.8-bin-hadoop2.7.tgz

sudo tar -zxvf spark-2.4.8-bin-hadoop2.7.tgz

sudo pip install findspark

echo "Set environmental variables for Spark.."

mv spark-2.4.8-bin-hadoop2.7 /home/ubuntu/

Set the path after creating the above setup -

export SPARK_HOME=/home/ubuntu/spark-2.4.8-bin-hadoop2.7

export PATH=$SPARK_HOME/bin:$PATH

export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH

export PYSPARK_PYTHON=python3

export PATH=$PATH:$JAVA_HOME/jre/bin

Getting files from S3-

Install AWS Client and AWS Configure-

apt install unzip

curl "https://awscli.amazonaws.com/awscli-exe-linux-x86_64.zip" -o "awscliv2.zip"

unzip awscliv2.zip

sudo ./aws/install

aws configure

Get files from S3 using Sync-

aws s3 sync s3://mywineprediction/ datamodel/

Command to Run Prediction model (Trained model is inside datamodel)-

spark-submit model_prediction.py /home/ubuntu/datamodel/ValidationDataset.csv /home/ubuntu/datamodel/ /home/ubuntu/datamodel/ .

Prediction Model without Docker in EC2 -

```
                 load Java programming language agent, see java.lang.instrument
    -splash:<imagepath>
                 show splash screen with specified image
See http://www.oracle.com/technetwork/java/javase/documentation/index.html for more details.
ubuntu@ip-172-31-13-224:~$ java -version
openjdk version "1.8.0_352"
OpenJDK Runtime Environment (build 1.8.0_352-8u352-ga-1~22.04-b08)
OpenJDK 64-Bit Server VM (build 25.352-b08, mixed mode)
ubuntu@ip-172-31-13-224:~$ cd datamodel/scripts/
ubuntu@ip-172-31-13-224:~/datamodel/scripts$ spark-submit model_prediction.py /home/ubuntu/datamodel/ValidationDataset.csv /home/ubuntu/datamodel/ /home/ubuntu/datamodel/ .
22/12/08 03:31:21 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
22/12/08 03:31:22 INFO SparkContext: Running Spark version 2.4.8
22/12/08 03:31:22 INFO SparkContext: Submitted application: WineQuality-Prediction
22/12/08 03:31:22 INFO SecurityManager: Changing view acls to: ubuntu
22/12/08 03:31:22 INFO SecurityManager: Changing modify acls to: ubuntu
22/12/08 03:31:22 INFO SecurityManager: Changing view acls groups to:
22/12/08 03:31:22 INFO SecurityManager: Changing modify acls groups to:
22/12/08 03:31:22 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(ubuntu); groups with view permissions: Set(); u
sers  with modify permissions: Set(ubuntu); groups with modify permissions: Set()
22/12/08 03:31:22 INFO Utils: Successfully started service 'sparkDriver' on port 46571.
22/12/08 03:31:22 INFO SparkEnv: Registering MapOutputTracker
22/12/08 03:31:22 INFO SparkEnv: Registering BlockManagerMaster
22/12/08 03:31:22 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/12/08 03:31:22 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/12/08 03:31:22 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-8eb6ec81-81c9-46f3-a68c-ae30867d9579
22/12/08 03:31:22 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
22/12/08 03:31:22 INFO SparkEnv: Registering OutputCommitCoordinator
22/12/08 03:31:22 INFO Utils: Successfully started service 'SparkUI' on port 4040.
22/12/08 03:31:23 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://ip-172-31-13-224.ec2.internal:4040
22/12/08 03:31:23 INFO Executor: Starting executor ID driver on host localhost
22/12/08 03:31:23 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 37159.
22/12/08 03:31:23 INFO NettyBlockTransferService: Server created on ip-172-31-13-224.ec2.internal:37159
22/12/08 03:31:23 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
22/12/08 03:31:23 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-13-224.ec2.internal, 37159, None)
22/12/08 03:31:23 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-13-224.ec2.internal:37159 with 366.3 MB RAM, BlockManagerId(driver, ip-172-31-13-224.ec2.inter
nal, 37159, None)
22/12/08 03:31:23 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-172-31-13-224.ec2.internal, 37159, None)
22/12/08 03:31:23 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-13-224.ec2.internal, 37159, None)
Loading data from /home/ubuntu/datamodel/ValidationDataset.csv..
==================Random Forest Prediction model==================
Trained model Location: /home/ubuntu/datamodel/rf-predicted.model ..
Evaluating the trained model...
Accuracy = 0.975
Test Error = 0.025000000000000022
F1-Score = 0.9635416666666666
```

14. Setup Docker -

Install Docker in EC2 -

sudo apt-get update

sudo apt install -y docker.io

sudo systemctl enable docker.service

sudo systemctl status docker.service

sudo usermod -a -G docker ubuntu

15. Create Docker image -

sudo docker build -t wine-prediction-app .

```
Dockerfile  datamodel  model_prediction.py
ubuntu@ip-172-31-9-96:~/dockercreate$ sudo docker build -t wine-prediction-app .
Sending build context to Docker daemon  164.4kB
Step 1/8 : From datamechanics/spark:2.4.5-hadoop-3.1.0-java-8-scala-2.11-python-3.7-dm18
2.4.5-hadoop-3.1.0-java-8-scala-2.11-python-3.7-dm18: Pulling from datamechanics/spark
214ca5fb9032: Pull complete
ebf31789c5c1: Pull complete
ab322dde1f12: Pull complete
5678190c8569: Pull complete
ed79cdacc966: Pull complete
39be23d30f30: Pull complete
251422b76562: Pull complete
da6b2a15c1a7: Pull complete
282cb8c457e0: Pull complete
d942add2765b: Pull complete
56be2a5b0c8d: Pull complete
8c7da9c3facd: Pull complete
4f4fb700ef54: Pull complete
dec63732166b: Pull complete
9c43b6d246e2: Pull complete
1256a346a938: Pull complete
52f9c1a36504: Pull complete
c9434d1fc9d0: Pull complete
6e64472d4652: Pull complete
3dc809aab2ab: Pull complete
723e605e3a72: Pull complete
be525bf9ee3c: Pull complete
fe0cdc369444: Pull complete
b3f06cbb3e0d: Pull complete
7673d24f358e: Pull complete
55af2afb5b30: Pull complete
Digest: sha256:52ea90c3832b7b340d78192de8d5d2237567e79506119914260 6b61681d3101d
Status: Downloaded newer image for datamechanics/spark:2.4.5-hadoop-3.1.0-java-8-scala-2.11-python-3.7-dm18
 ---> 70f0b6e1f4bd
Step 2/8 : ENV PYSPARK_MAJOR_PYTHON_VERSION=3
 ---> Running in 2a68f6c60e6e
Removing intermediate container 2a68f6c60e6e
 ---> e07b04491e8d
Step 3/8 : RUN conda install -y numpy
 ---> Running in 8fb8d382012f
Collecting package metadata (current_repodata.json): ...working... done
```

```
  conda                               4.12.0-py37h06a4308_0 --> 22.11.1-py37h06a4308_3
  libgcc-ng                               9.1.0-hdf63c60_0 --> 11.2.0-h1234567_1
  libstdcxx-ng                            8.2.0-hdf63c60_1 --> 11.2.0-h1234567_1


Downloading and Extracting Packages
ruamel.yaml-0.17.21  | 177 KB    | ########## | 100%
zipp-3.8.0           | 15 KB     | ########## | 100%
libgomp-11.2.0       | 474 KB    | ########## | 100%
toolz-0.12.0         | 104 KB    | ########## | 100%
bottleneck-1.3.5     | 115 KB    | ########## | 100%
pytz-2022.1          | 196 KB    | ########## | 100%
ruamel.yaml.clib-0.2 | 140 KB    | ########## | 100%
_openmp_mutex-5.1    | 21 KB     | ########## | 100%
numexpr-2.8.4        | 133 KB    | ########## | 100%
conda-22.11.1        | 932 KB    | ########## | 100%
python-dateutil-2.8. | 233 KB    | ########## | 100%
pyparsing-3.0.9      | 150 KB    | ########## | 100%
typing_extensions-4. | 45 KB     | ########## | 100%
pluggy-1.0.0         | 29 KB     | ########## | 100%
pandas-1.3.5         | 9.3 MB    | ########## | 100%
libgcc-ng-11.2.0     | 5.3 MB    | ########## | 100%
libstdcxx-ng-11.2.0  | 4.7 MB    | ########## | 100%
flit-core-3.6.0      | 42 KB     | ########## | 100%
importlib_metadata-4 | 12 KB     | ########## | 100%
packaging-21.3       | 36 KB     | ########## | 100%
importlib-metadata-4 | 40 KB     | ########## | 100%
Preparing transaction: ...working... done
Verifying transaction: ...working... done
Executing transaction: ...working... done
Removing intermediate container ca4d1b8104a0
 ---> 8d71d872a840
Step 5/8 : WORKDIR /opt/wine-prediction-app
 ---> Running in cdea14a3209d
Removing intermediate container cdea14a3209d
 ---> 4a4db976c5de
Step 6/8 : COPY model_prediction.py .
 ---> 13b66a120c31
Step 7/8 : ADD datamodel/ValidationDataset.csv .
 ---> dcdd4044bf4d
Step 8/8 : ADD datamodel ./datamodel/
 ---> 90865ba5fe73
Successfully built 90865ba5fe73
Successfully tagged wine-prediction-app:latest
ubuntu@ip-172-31-9-96:~/dockercreate$
```

```
ubuntu@ip-172-31-67-164:~/dockercreate$ sudo docker image ls
REPOSITORY           TAG                                                IMAGE ID       CREATED             SIZE
wine-prediction-app  latest                                             4ddbc162bd66   About a minute ago  3.66GB
datamechanics/spark  2.4.7-hadoop-3.1.0-java-8-scala-2.12-python-3.7-dm18  ec9c7ace56e0   6 months ago        2.38GB
ubuntu@ip-172-31-67-164:~/dockercreate$
```

16. Push the prediction application into Docker Hub -

sudo docker login -u soumilee

sudo docker tag wine-prediction-app:latest soumilee/wine-prediction:latest

sudo docker push soumilee/wine-prediction:latest

```
ubuntu@ip-172-31-67-164:~/dockercreate$ sudo docker push soumilee/wine-prediction:latest
The push refers to repository [docker.io/soumilee/wine-prediction]
952840939b5b: Pushed
c653036eb841: Pushed
8a179767cad4: Pushed
691932afadbe: Pushed
a35b0f837957: Pushed
96a52a841073: Pushed
8b7cde469e8b: Pushed
762ba40c6ff8: Mounted from datamechanics/spark
52c9a3fd2fdb: Mounted from datamechanics/spark
ddef20fcb230: Mounted from datamechanics/spark
2574d33df3ee: Mounted from datamechanics/spark
70e6ed32ee66: Mounted from datamechanics/spark
3c4fe403647b: Mounted from datamechanics/spark
d8361e1d392d: Mounted from datamechanics/spark
2c8a66d8e359: Mounted from datamechanics/spark
42ee61ba57ce: Mounted from datamechanics/spark
acc113369534: Mounted from datamechanics/spark
a98d2da3e2b2: Mounted from datamechanics/spark
1a49c5b442b0: Mounted from datamechanics/spark
2833f61fe10b: Mounted from datamechanics/spark
163acb2ef19f: Mounted from datamechanics/spark
5f70bf18a086: Mounted from datamechanics/spark
a3472b551ed8: Mounted from datamechanics/spark
d4454921b358: Mounted from datamechanics/spark
493629289764: Mounted from datamechanics/spark
df5eb8e7ce9e: Mounted from datamechanics/spark
1eb0ae09a239: Mounted from datamechanics/spark
d9ed100561cb: Mounted from datamechanics/spark
b5cd7ef483a5: Mounted from datamechanics/spark
d8717a08c273: Mounted from datamechanics/spark
9579832344fa: Mounted from datamechanics/spark
094a290fbb48: Mounted from datamechanics/spark
6d1bd5a53aa6: Mounted from datamechanics/spark
fd95118eade9: Mounted from datamechanics/spark
latest: digest: sha256:afddba9d84c730caa86dcb61aacd251fed095088c84d87f3571cd7dade3f3e33 size: 7454
ubuntu@ip-172-31-67-164:~/dockercreate$
```

Prediction using Docker -

sudo docker run wine-prediction-app driver model_prediction.py datamodel/ValidationDataset.csv datamodel/ datamodel/

```
ubuntu@ip-172-31-67-164:~/dockercreate$ sudo docker run wine-prediction-app driver model_prediction.py datamodel/ValidationDataset.csv datamodel/ datamodel/
Unsetting extraneous env vars (UTC): 02:29:11
Finished unsetting extraneous env vars (UTC): 02:29:11
++ id -u
+ myuid=185
++ id -g
+ mygid=0
+ set +e
++ getent passwd 185
+ uidentry=
+ set -e
+ '[' -z '' ']'
+ '[' -w /etc/passwd ']'
+ echo '185:x:185:0:anonymous uid:/opt/spark:/bin/false'
+ SPARK_K8S_CMD=driver
+ case "$SPARK_K8S_CMD" in
+ shift 1
+ SPARK_CLASSPATH=':/opt/spark/jars/*'
+ grep SPARK_JAVA_OPT_
+ sed 's/[^=]*=\(.*\)/\1/g'
+ sort -t_ -k4 -n
+ env
+ readarray -t SPARK_EXECUTOR_JAVA_OPTS
+ '[' -n '' ']'
+ '[' -n '' ']'
+ PYSPARK_ARGS=
+ '[' -n '' ']'
+ R_ARGS=
+ '[' -n '' ']'
+ '[' 3 == 2 ']'
+ '[' 3 == 3 ']'
++ python3 -V
+ pyv3='Python 3.7.13'
+ export PYTHON_VERSION=3.7.13
+ PYTHON_VERSION=3.7.13
+ export PYSPARK_PYTHON=python3
+ PYSPARK_PYTHON=python3
+ export PYSPARK_DRIVER_PYTHON=python3
+ PYSPARK_DRIVER_PYTHON=python3
+ case "$SPARK_K8S_CMD" in
+ CMD=("$SPARK_HOME/bin/spark-submit" --conf "spark.driver.bindAddress=$SPARK_DRIVER_BIND_ADDRESS" --deploy-mode client "$@")
+ exec /usr/bin/tini -s -- /opt/spark/bin/spark-submit --conf spark.driver.bindAddress= --deploy-mode client model_prediction.py datamodel/ValidationDataset.csv datamodel/ datamodel/
22/12/09 02:29:13 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
22/12/09 02:29:14 INFO SparkContext: Running Spark version 2.4.7
22/12/09 02:29:14 INFO SparkContext: Submitted application: WineQuality-Prediction
22/12/09 02:29:14 INFO SecurityManager: Changing view acls to: 185
22/12/09 02:29:14 INFO SecurityManager: Changing modify acls to: 185
22/12/09 02:29:14 INFO SecurityManager: Changing view acls groups to:
22/12/09 02:29:14 INFO SecurityManager: Changing modify acls groups to:
22/12/09 02:29:14 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(185); groups with view permissions: Set(); users  with modify permissions: Set(185); groups with modify permissions: Set()
22/12/09 02:29:15 INFO Utils: Successfully started service 'sparkDriver' on port 40645.
22/12/09 02:29:15 INFO SparkEnv: Registering MapOutputTracker
22/12/09 02:29:15 INFO SparkEnv: Registering BlockManagerMaster
22/12/09 02:29:15 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/12/09 02:29:15 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/12/09 02:29:15 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-b0168b10-bf06-43b7-9ff1-617902fe36a6
22/12/09 02:29:15 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
22/12/09 02:29:15 INFO SparkEnv: Registering OutputCommitCoordinator
22/12/09 02:29:15 INFO Utils: Successfully started service 'SparkUI' on port 4040.
22/12/09 02:29:15 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://237b5d2fc4ed:4040
22/12/09 02:29:15 INFO Executor: Starting executor ID driver on host localhost
22/12/09 02:29:15 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 39977.
22/12/09 02:29:15 INFO NettyBlockTransferService: Server created on 237b5d2fc4ed:39977
22/12/09 02:29:15 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
22/12/09 02:29:15 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 237b5d2fc4ed, 39977, None)
22/12/09 02:29:15 INFO BlockManagerMasterEndpoint: Registering block manager 237b5d2fc4ed:39977 with 366.3 MB RAM, BlockManagerId(driver, 237b5d2fc4ed, 39977, None)
22/12/09 02:29:15 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 237b5d2fc4ed, 39977, None)
22/12/09 02:29:15 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 237b5d2fc4ed, 39977, None)
Loading data from datamodel/ValidationDataset.csv..
=================Random Forest Prediction model=================
Trained model location: datamodel/rf-predicted.model ..
Evaluating the trained model...
Accuracy = 0.975
Test Error = 0.02500000000000022
F1-Score = 0.9635416666666666
ubuntu@ip-172-31-67-164:~/dockercreate$
```

As it can be seen with Docker using Random Forest model the prediction accuracy is 0.975 the error is 0.025 and the F1 score is 0.9635

Docker hub link - https://hub.docker.com/r/soumilee/wine-prediction

GitHub link - https://github.com/Soumilee/wine-prediction