AWS Spark Wine Quality Prediction Application

Pyspark application PREDWINE. In order to forecast wine quality using publically available data, the goal of this project is to train a machine learning model in parallel on EC2 instances. Once trained, the model will be used to make wine quality predictions.   
In order to make deployments easier, the project also uses Docker to create a container for a trained machine learning model.  
  
  
There are two primary Python source files in this project:   
1. Using an EMR spark cluster, `qualityprediction.py` trains the model in parallel by reading the training dataset from S3. After the model has been trained, test data sent over S3 can be used to execute it. This application saves the trained model in an S3 bucket.  
2. The program `qualitytestdataprediction.py` loads the trained model and runs it on the specified testdata file.

**Link for GitHub:**

[**https://github.com/Avanthika2812/Cloud-project**](https://github.com/Avanthika2812/Cloud-project)

**Link for Docker:**

<https://hub.docker.com/repository/docker/am3737/qulpred/general>

**Docker file:** Docker file to create docker image and run container for simplified deployment.

Instruction to use:

1. How to create Spark cluster in AWS

User can create spark cluster using EMR console provided by AWS. Please follow steps to create one with 4 ec2 instances (use more instances depending upon your load).

1. Create Key-Pair for EMR cluster using navigation ```EC2-> Network & Security -> Key-pairs```.

Use .pem as format. This will download {name of key pair}>.pem file. Keep it safe you will need that to do SSH to EC2 instances.

2. Navigate to Amazon EMR console using link https://console.aws.amazon.com/elasticmapreduce/home?region=us-east-1. Then, navigate to clusters-> create cluster.

3. Now fill in respective sections:

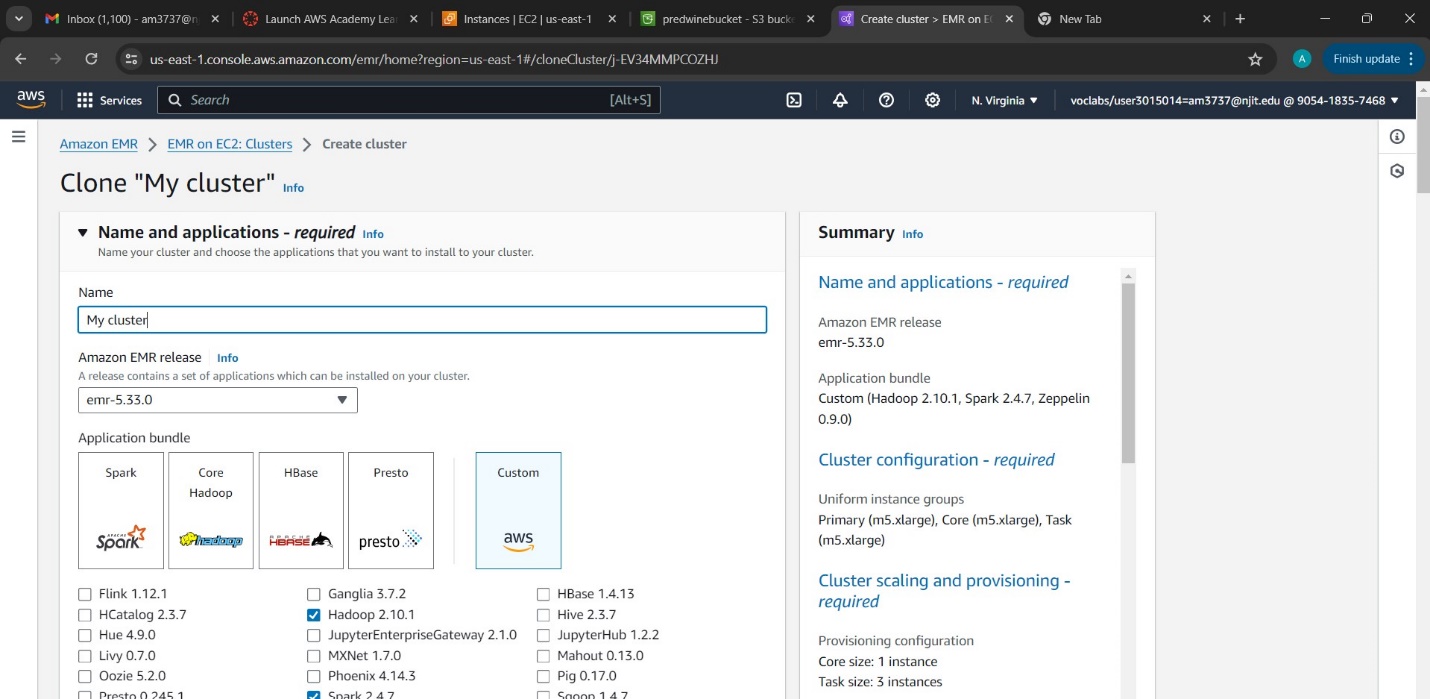
General Configuratin -> My cluster

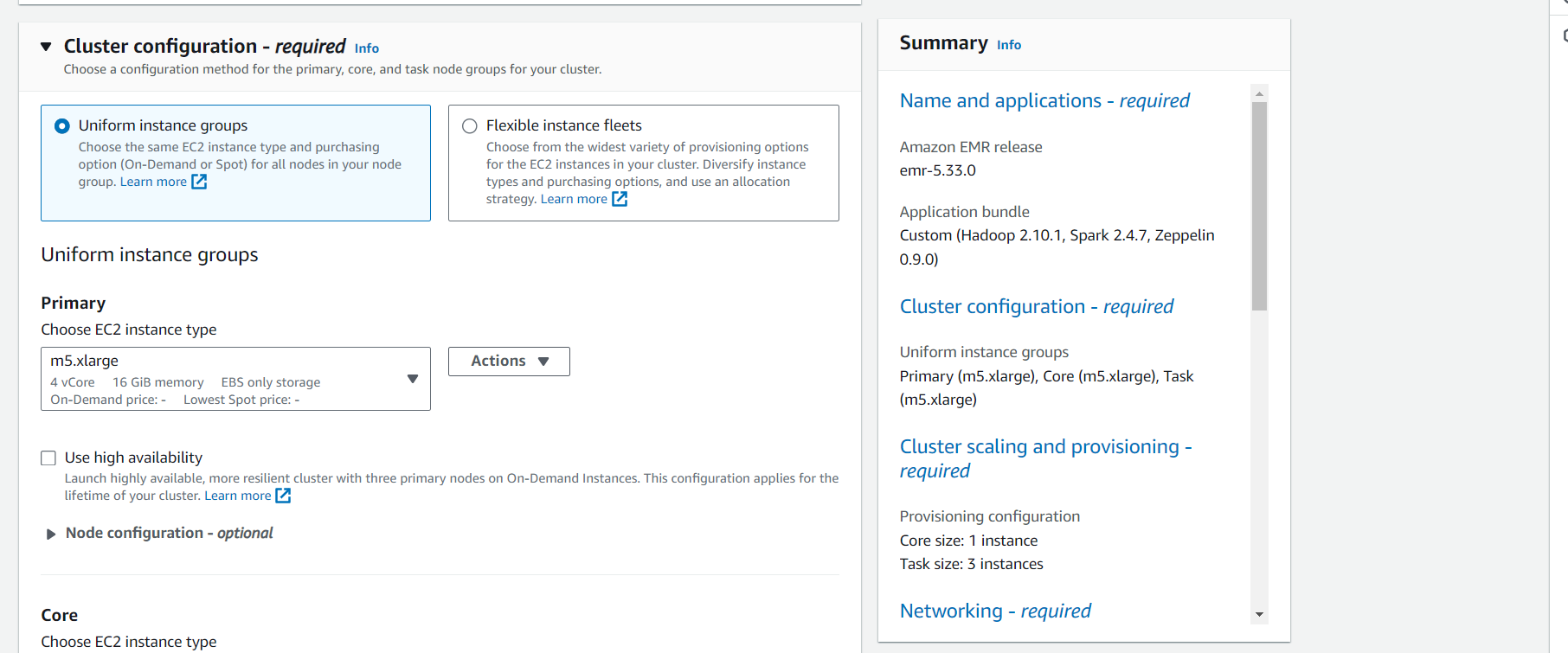
Software Configuration-> EMR 5.33 , do select 'Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.9.0' option menu.

Hardware Configuration -> Make instance count as 4

Security Access -> Ava.pem

Rest of parameters can be left default.





4. Cluster status should be 'Waiting' on successful cluster creation.

2. How to train ML model in Spark cluster with 4 ec2 instances in parallel

1. Now when cluster is ready to accept jobs, to submit one you can either use step button to add steps or submit manually.

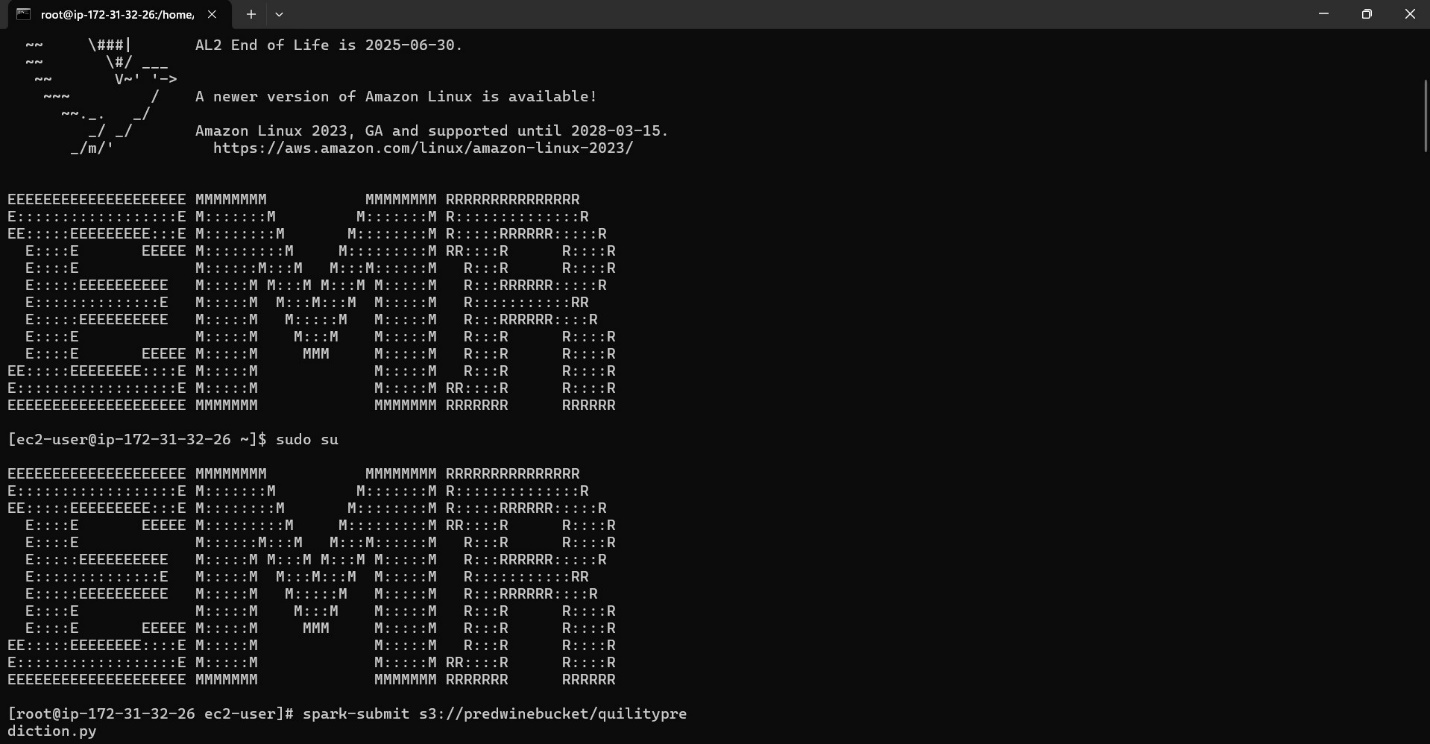
To submit manually, Perform SSH to Master of cluster using below command:

ssh -i "Ava.pem" ec2-user@ec2-3-91-246-174.compute-1.amazonaws.com

and it is successfully logged in.

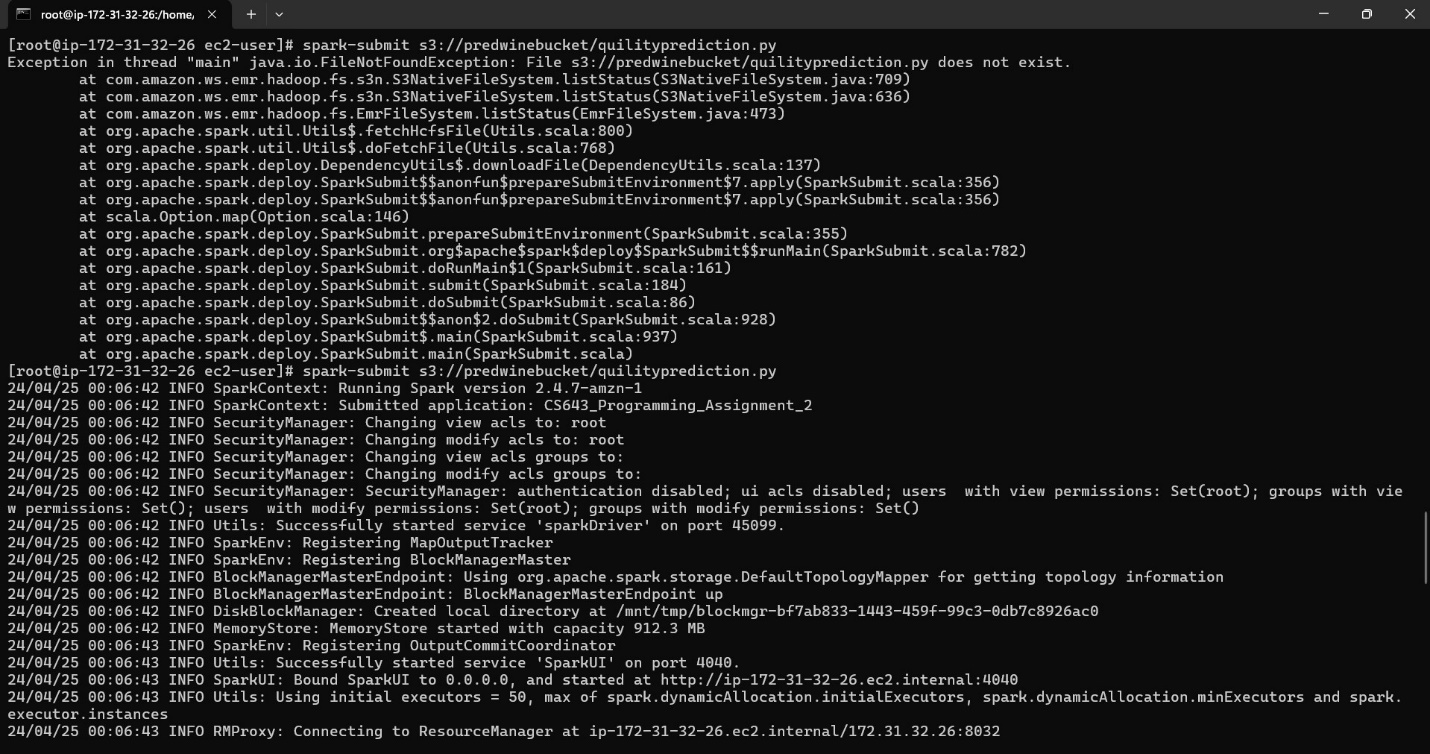
2. On successful login to master , change to root user by running command:

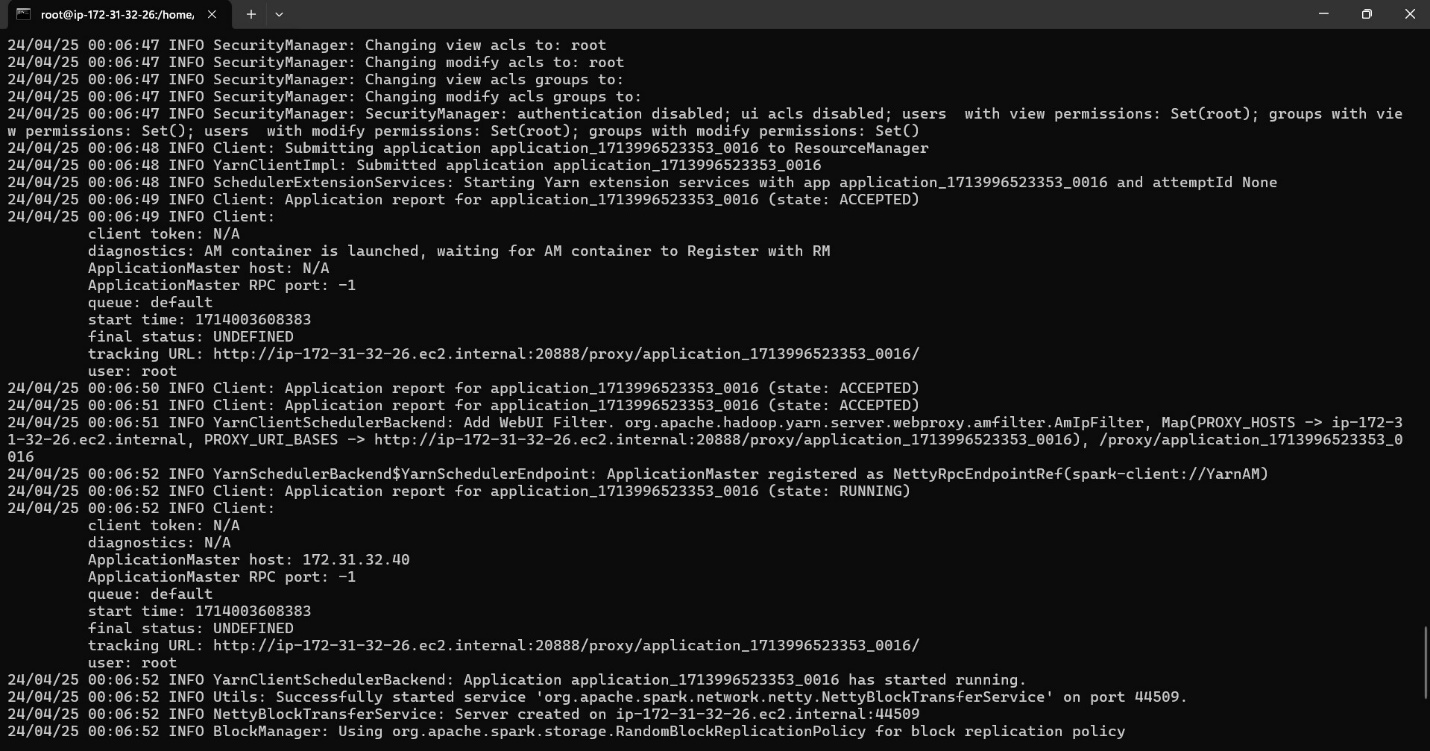
sudo su

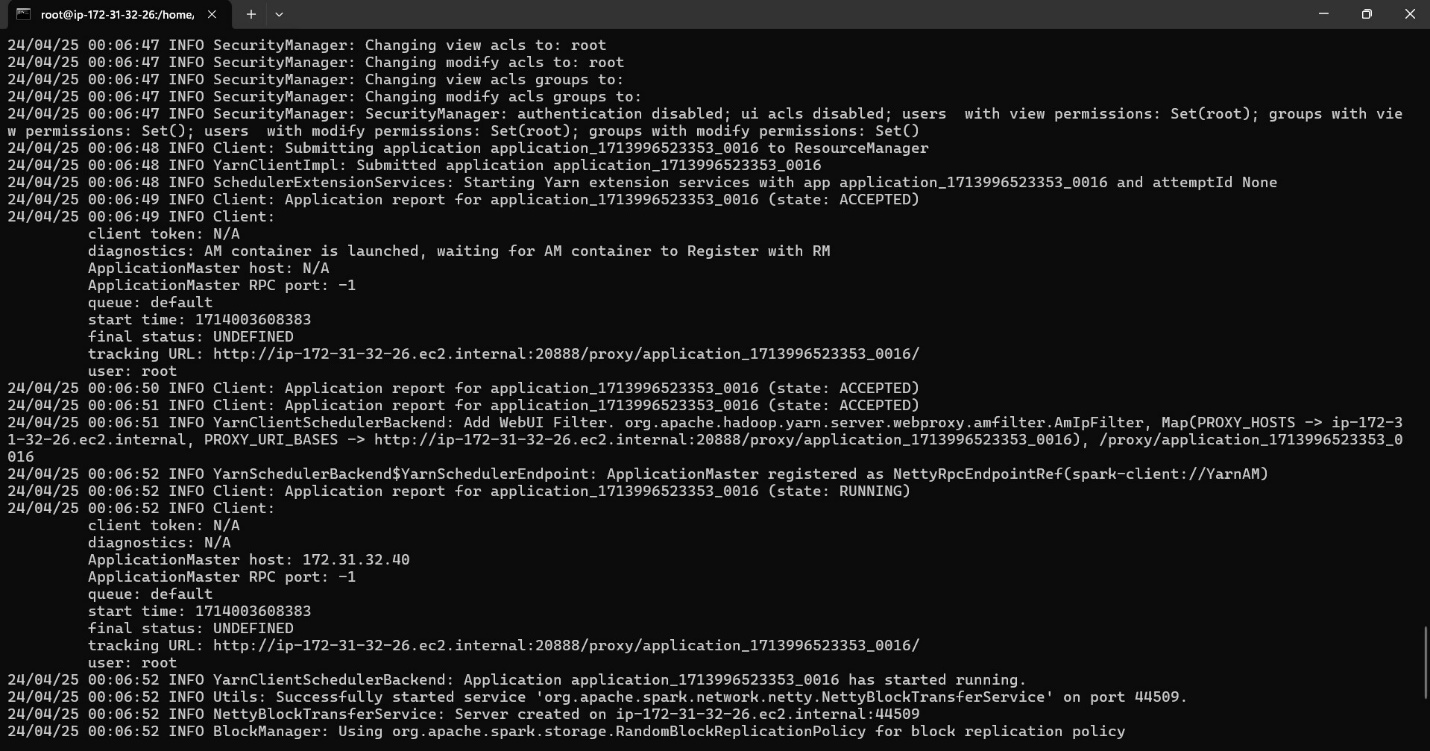


3. Submit job using following command:

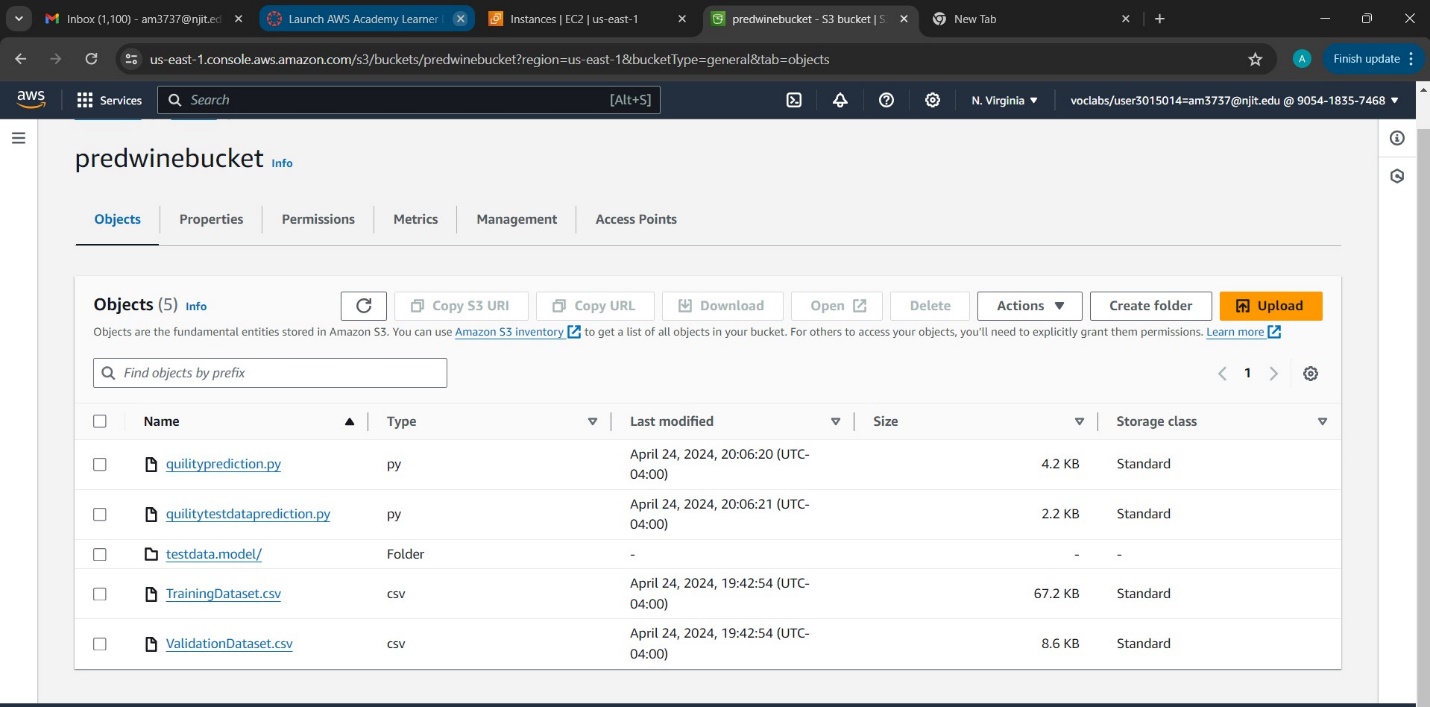
spark-submit s3://am3737qulpred/winequilityprediction.py



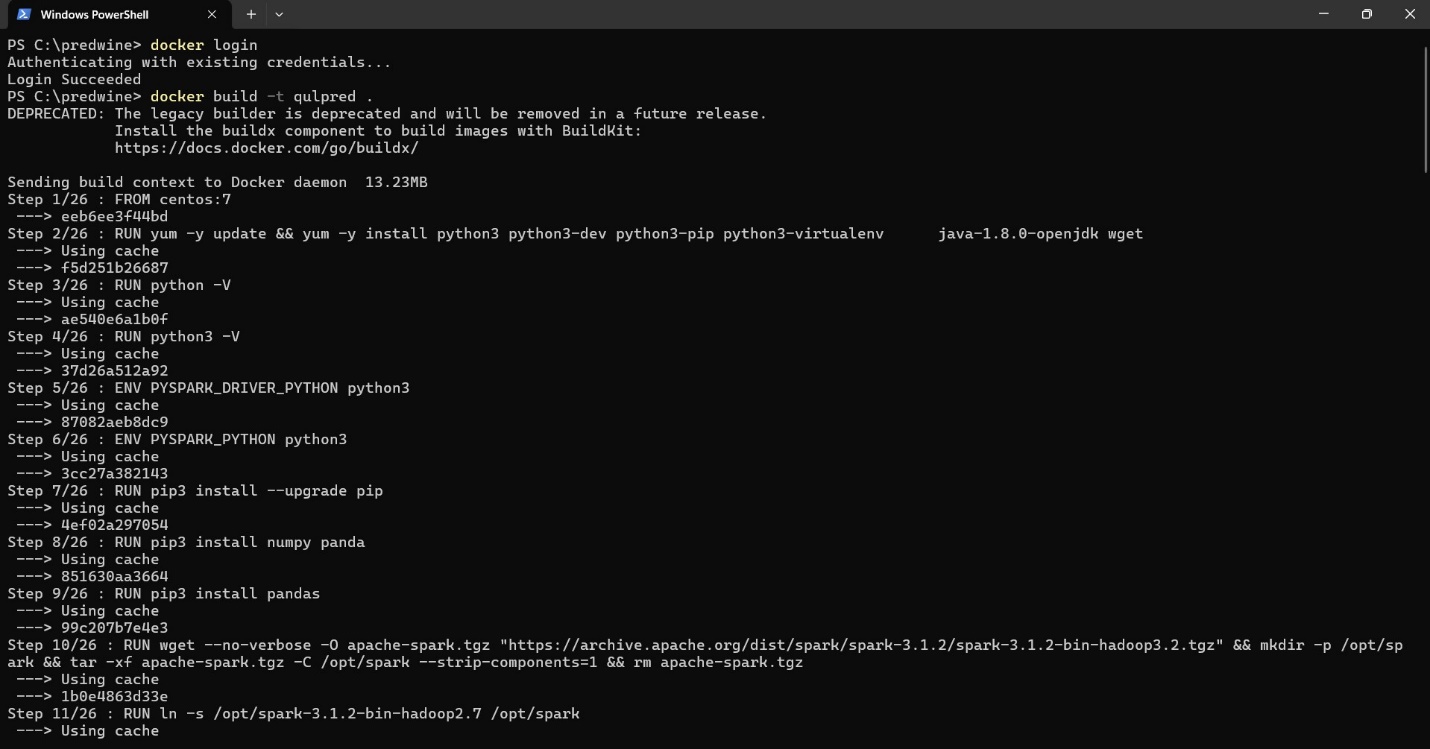




4. You can trace status of this job in EMR UI application logs. Once status is succeded a test.model will be created in s3://predwinebucket



3. Locally running a learned machine learning model without Docker.   
1. Make a clone of this repository.  
2. To run this, confirm that your local Spark environment is set up. Click this link to put one up: https://spark.apache.org/docs/latest   
  
3. Open the predwine or src folder.   
4. Set up the Docker software.5. Use the command "docker login Pwd" to access the docker in the power shell.   
6. Once everything is set up, run the following command:   
build -t qulpred docker image.



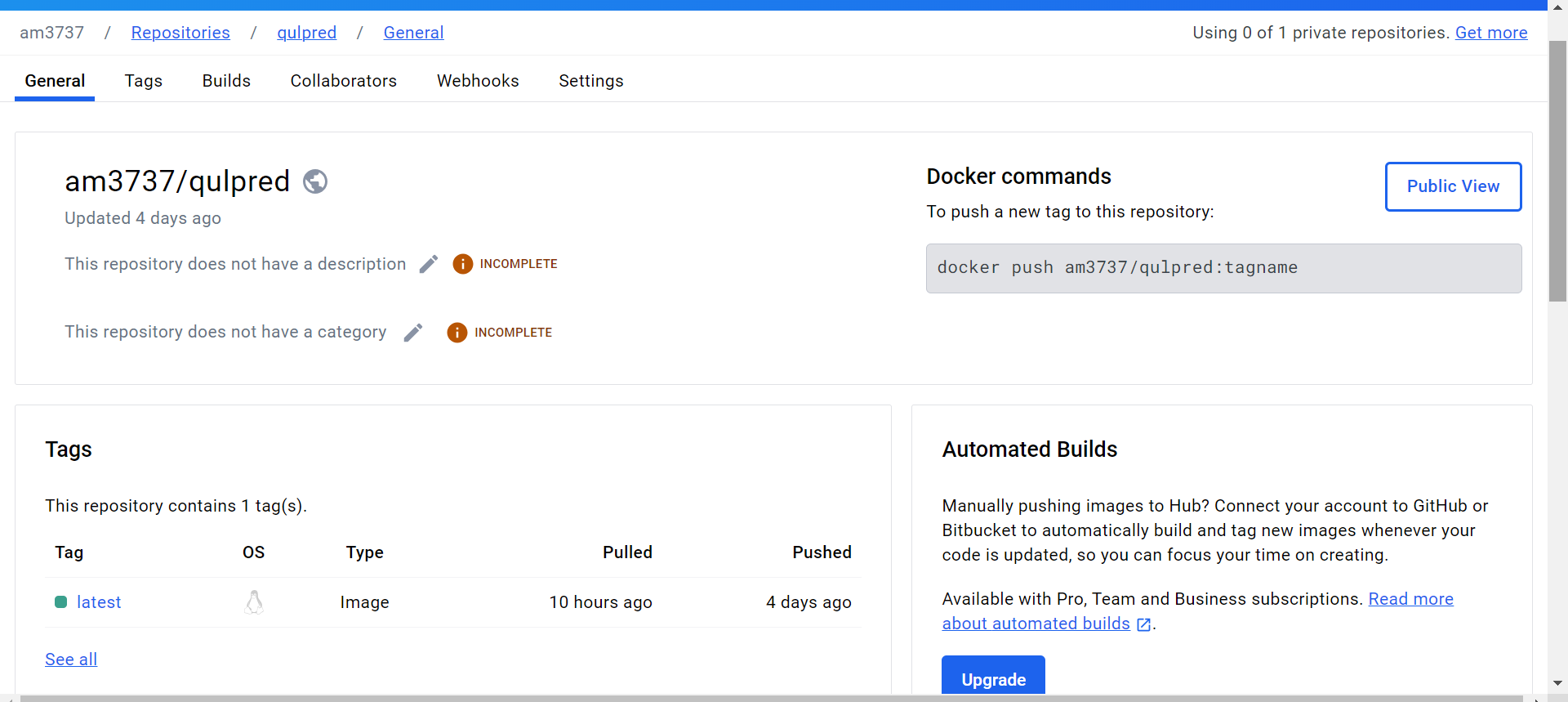
4. Use Docker to run the ML model

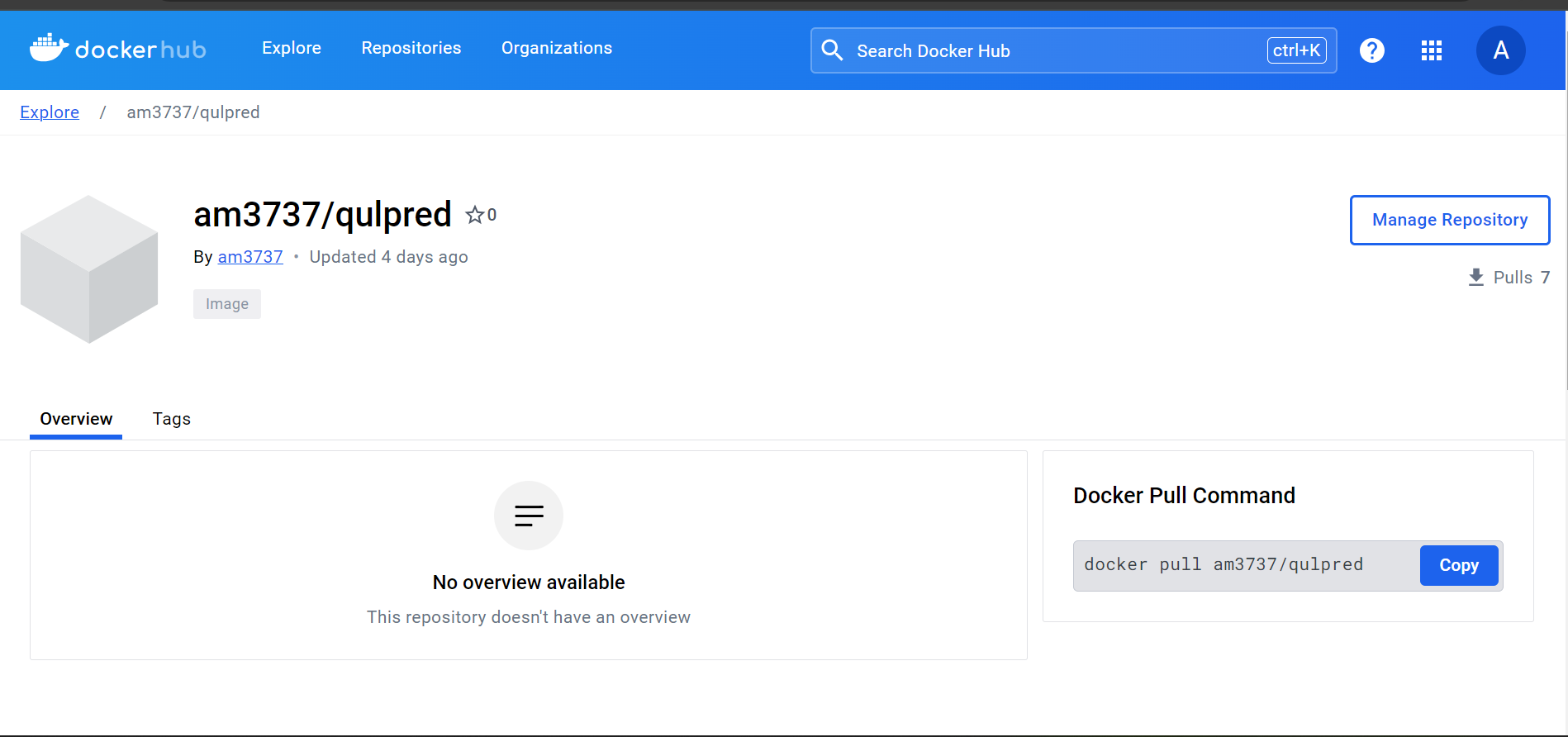
1. Install Docker where you wish to use this container.

2. Adding and removing objects from the Docker Hub repository:

PUSH: docker push am3737/qulpred docker tag qulpred am3737/qulpred

PULL: docker pull qulpred/am3737

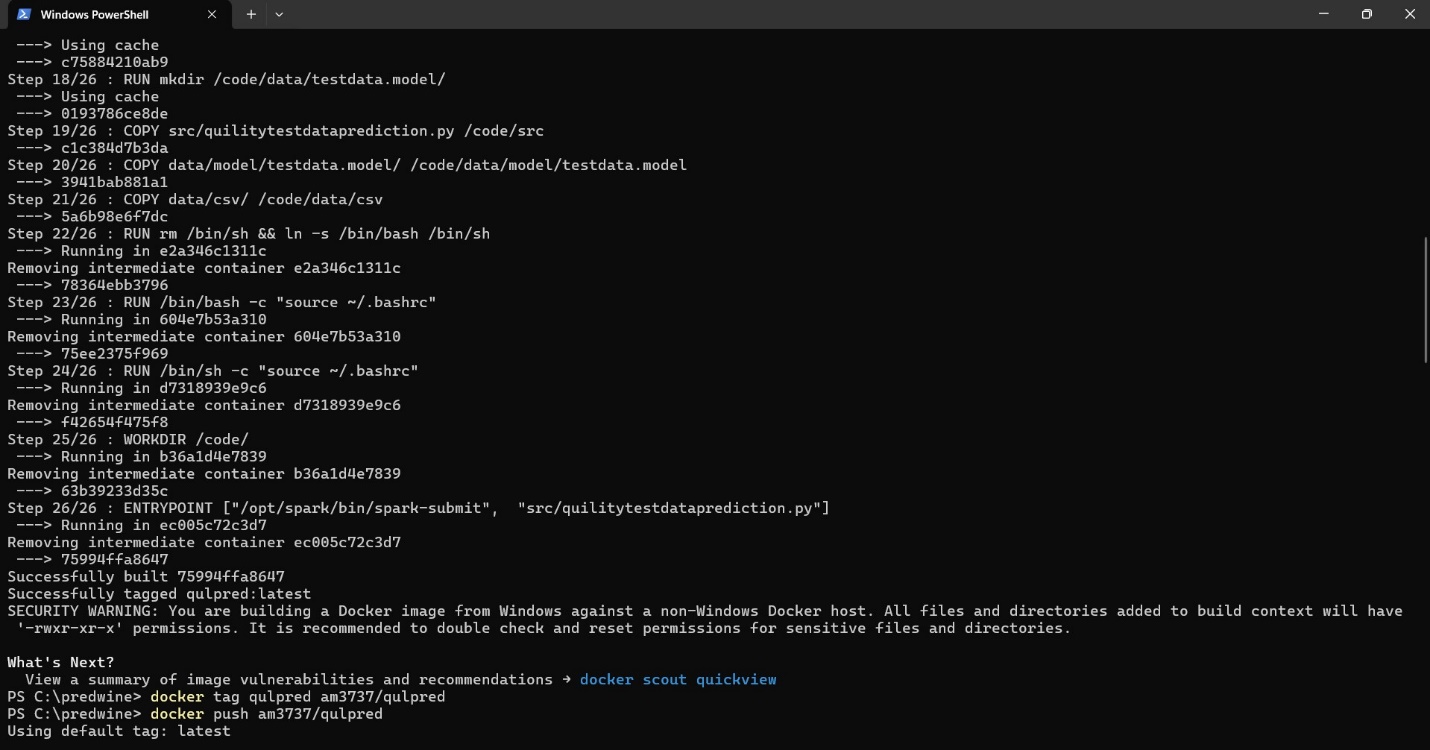


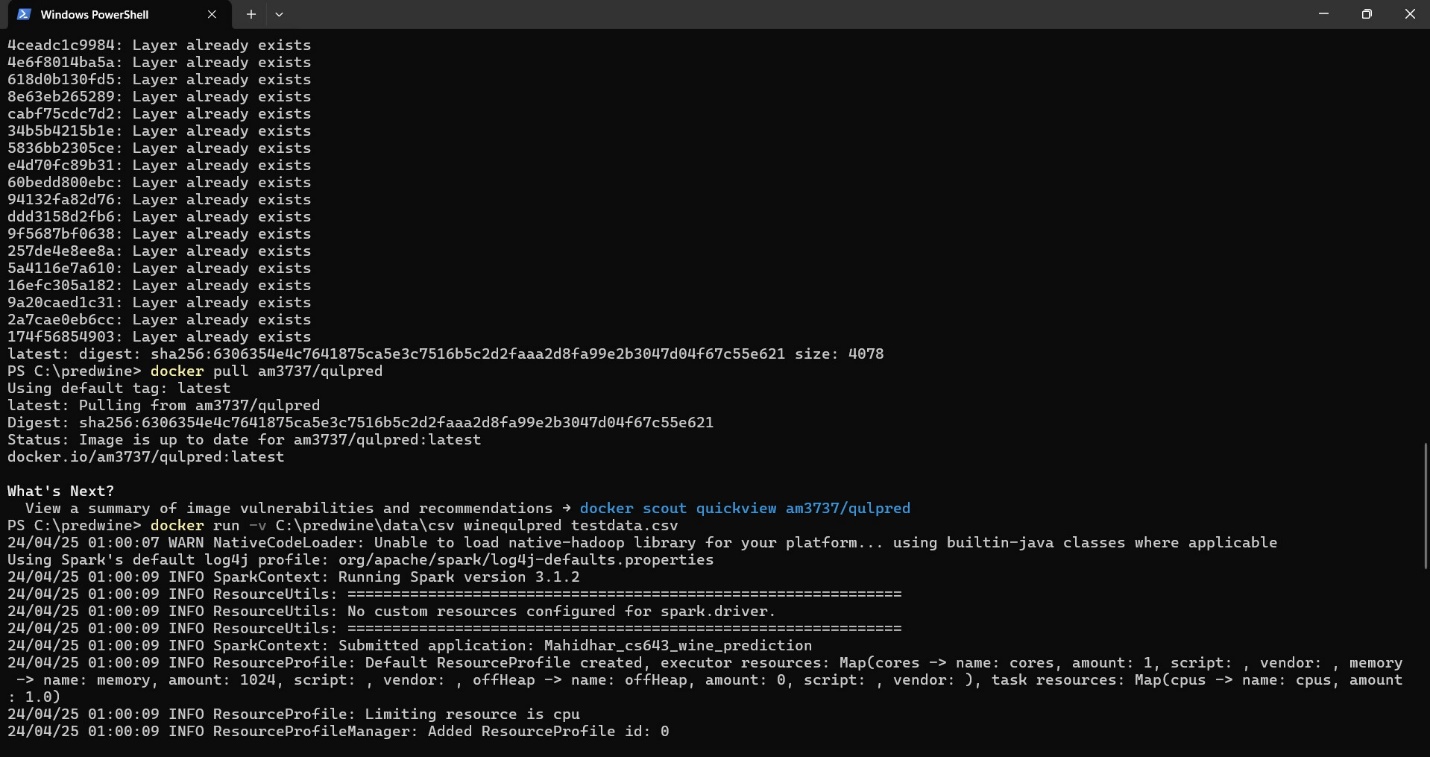


3. Place your testdata file in a folder (lets call it directory dirA) , which you will mount with docker container.

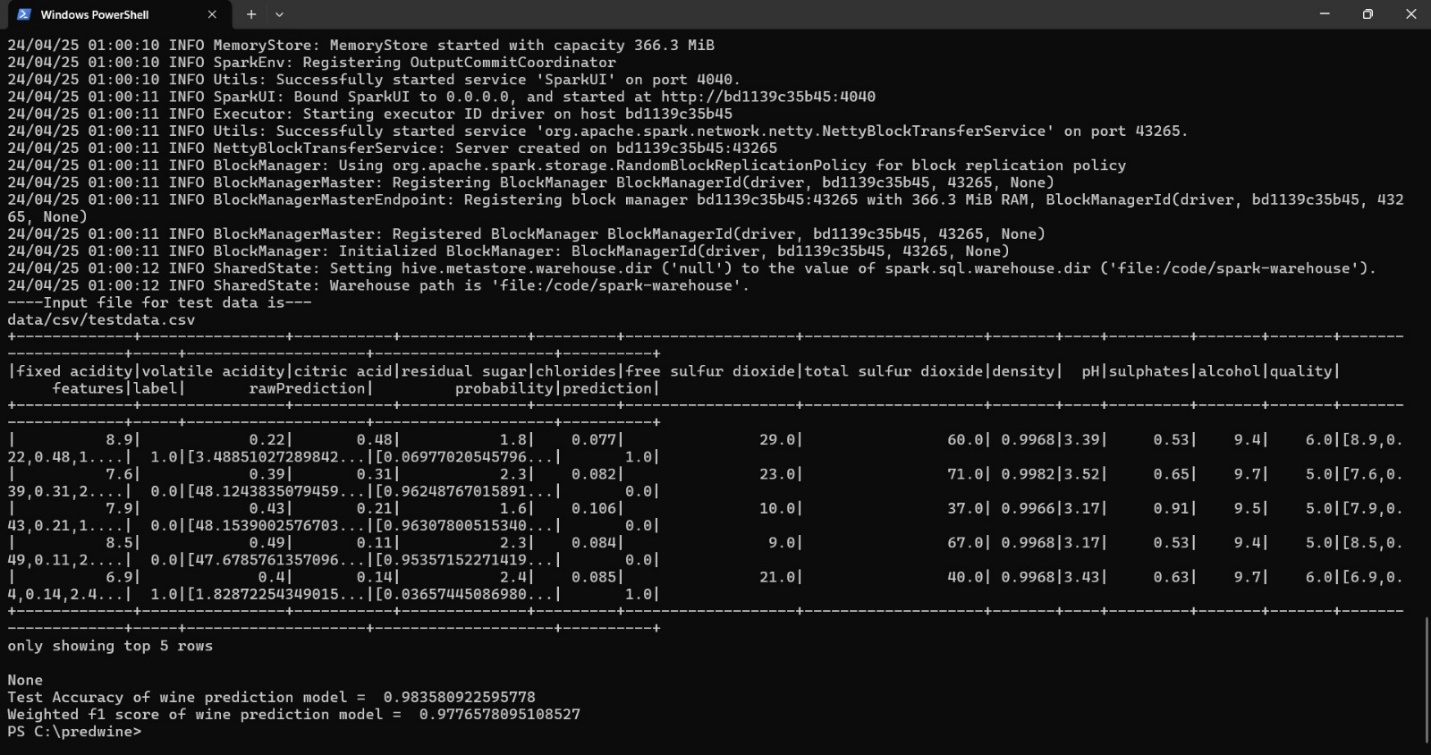
4. docker run -v {directory path for data dirA}:/predwine/data/csv/winequlpred {testdata file name}

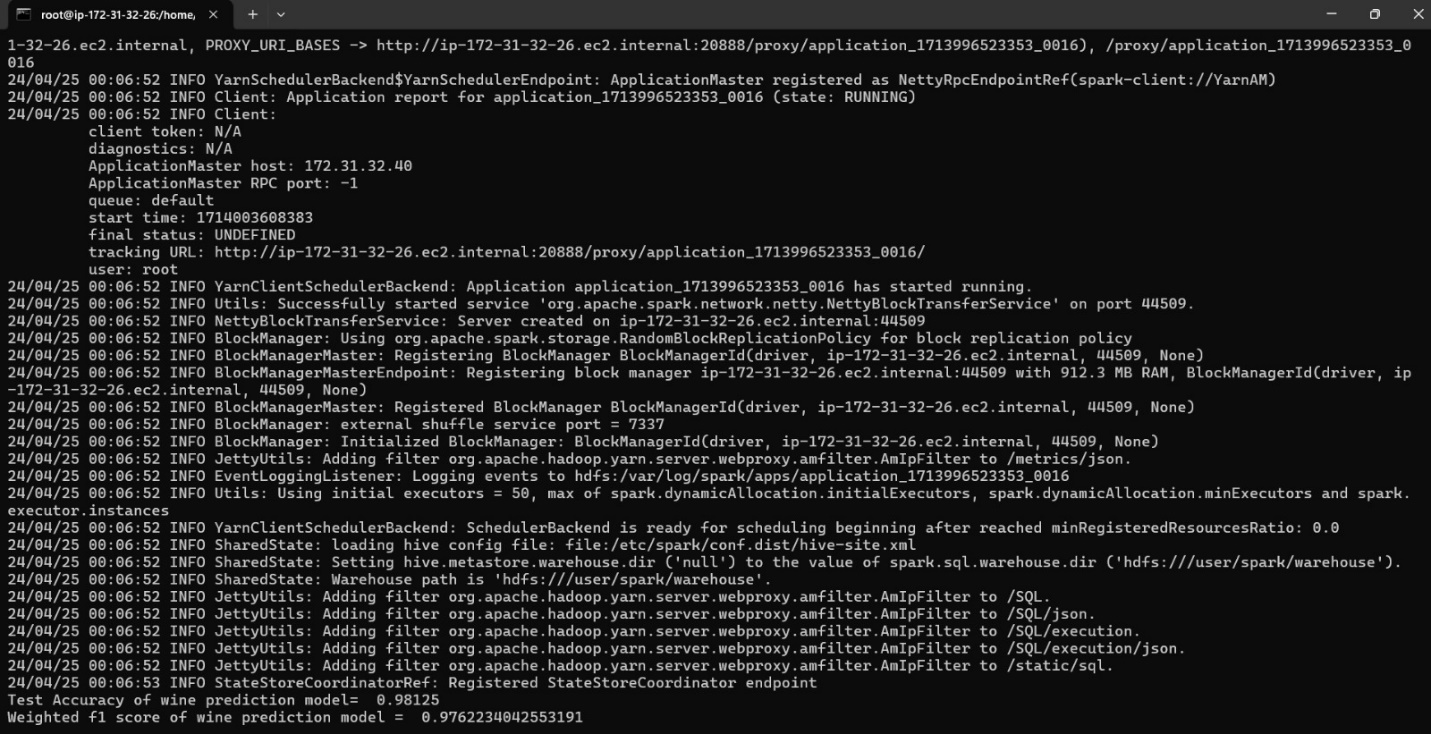
Sample command





docker run -v /Users/<username>/<path-to-folder>/predwine/data/csv:/code C:\predwine\data\csv winequlpred testdata.csv





\testdata.csv

**3. Locally running a learned machine learning model without Docker**.

Make a clone of this repository.  
To execute this, make sure your local Spark environment is set up. Click this link to put one up: https://spark.apache.org/docs/latest   
  
Open the folder predwine/src.   
Put the test data into the 'predwine/data/csv' directory.   
Use pip -m install pyspark to install Pyspark. or set up the conda package.   
When everything is ready, run the following command: Carry out the subsequent command :

cd predwine/src

spark-submit wine\_test\_data\_prediction.py <filename>