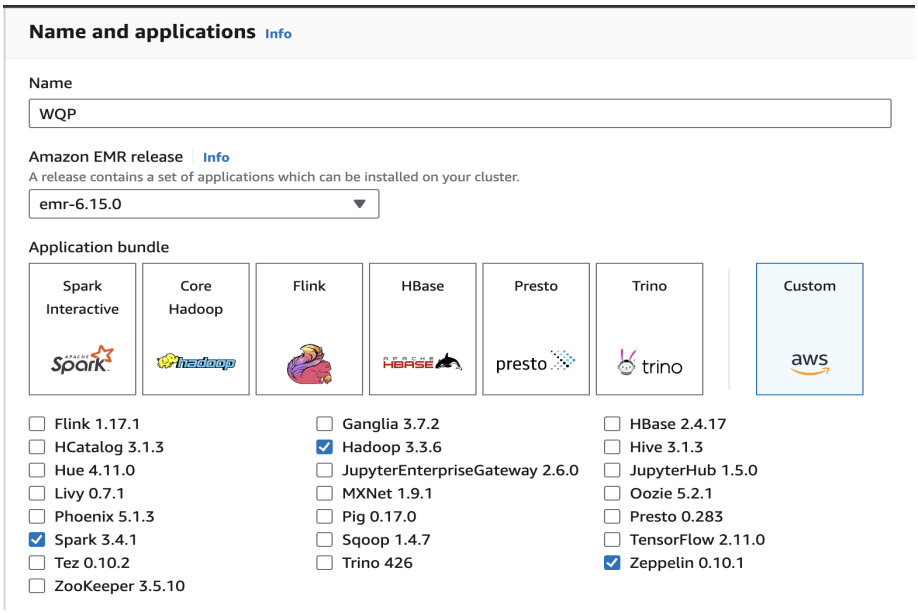
GitHub: https://github.com/sgudishetty9/cloudcomputing-proj2

DockerHub: https://hub.docker.com/repository/docker/sg2489/wqprepo/general

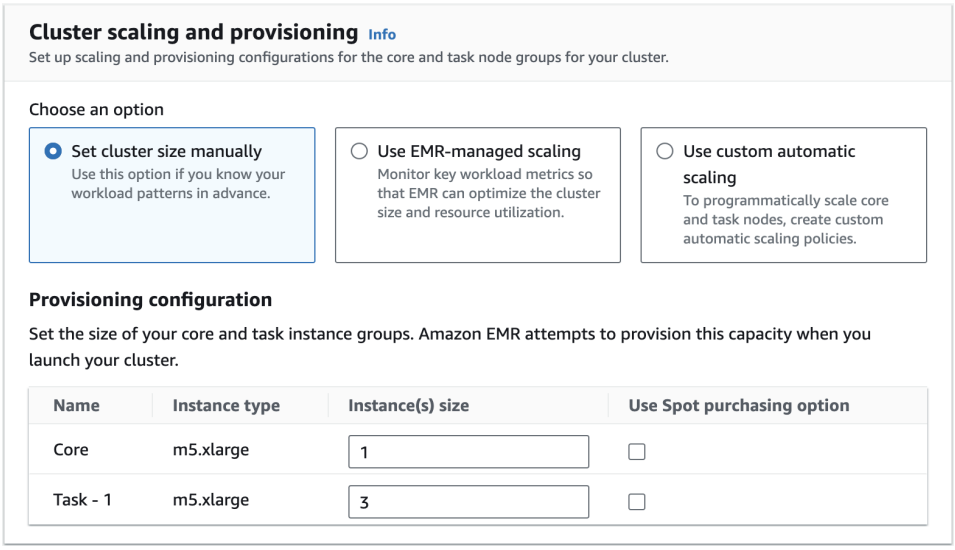
Procedure:

Cluster Creation

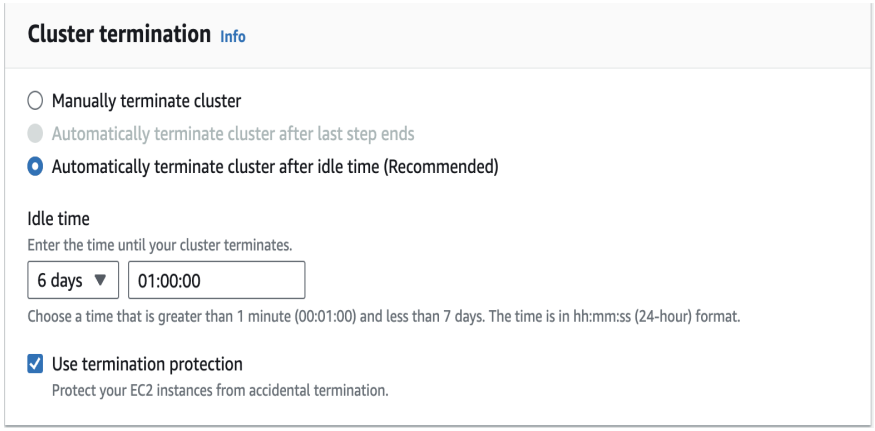
Navigate to Amazon EMR in AWS > Create the Cluster, and select the following apps:



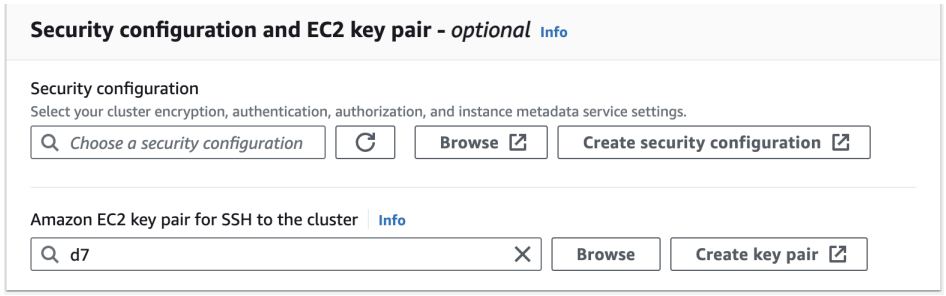
Had to train model using 4 EC2 instances, so configured 'Provisioning Configuration as below:



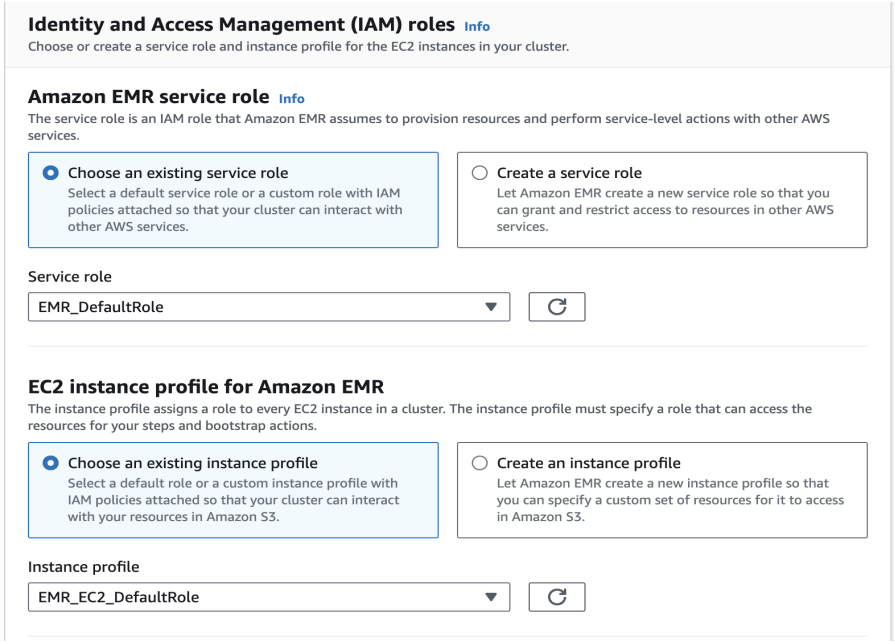
Went with Recommended Cluster Termination settings:



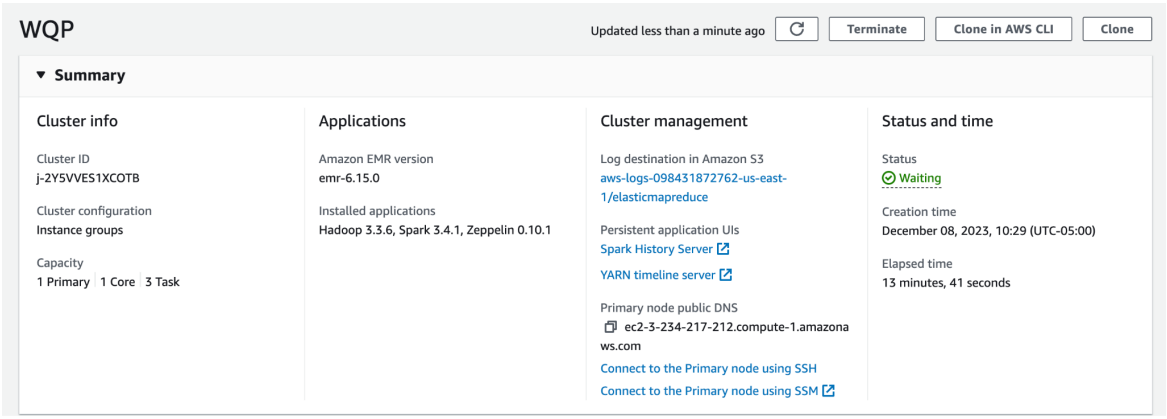
Create a key pair for SSH and SCP to the Instance/Cluster:



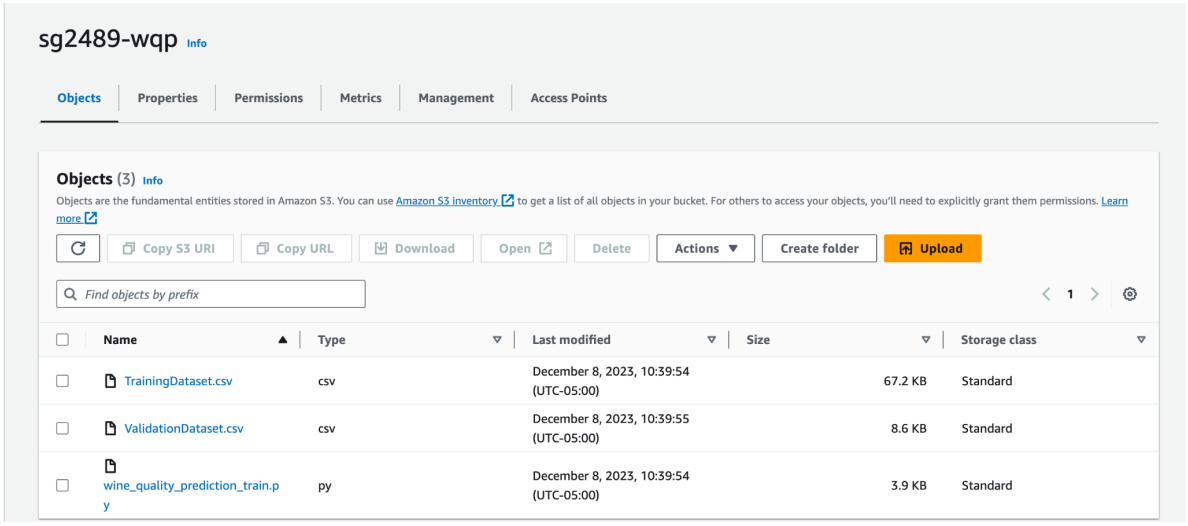
Select default IAM roles:



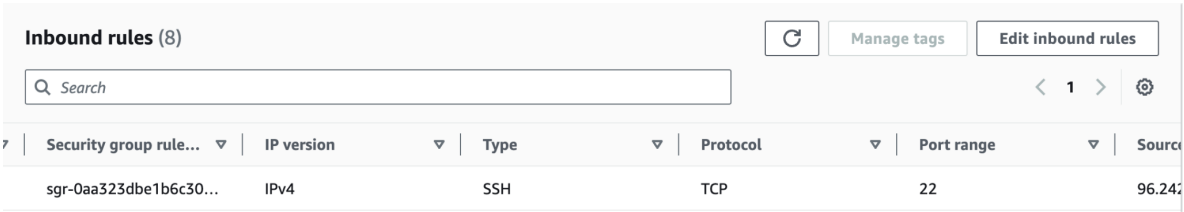
Rest, go with default configuration, cluster/s will be working upon turning to ‘Waiting’ status.



Create an S3 Bucket and drop training and testing files into it.



Ensure master instance of the cluster has permission to SSH, if not edit inbound rule after navigating to the security group of the instance.



Done SSH into my master instance using below command:
ssh -i "<my-key-pair.pem>" ec2-user@<Public DNS>

```
   ,     #_
   ~\_  ####_        Amazon Linux 2
  ~~  \_#####\
  ~~     \###|       AL2 End of Life is 2025-06-30.
  ~~       \#/ ___
   ~~       V~' '->
    ~~~         /    A newer version of Amazon Linux is available!
      ~~._.   _/
         _/ _/       Amazon Linux 2023, GA and supported until 2028-03-15.
       _/m/'           https://aws.amazon.com/linux/amazon-linux-2023/

19 package(s) needed for security, out of 20 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
```

Navigated to root user using:
sudo su

Now, trained the model with training file using the below command:
spark-submit s3://<bucket-name>/wine_quality_prediction_train.py

After successful training, results of trained model are:

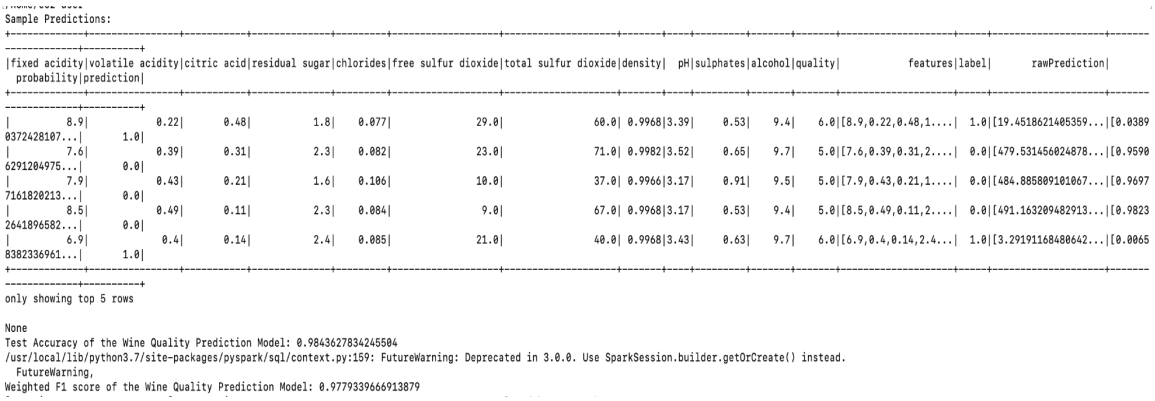
```
Test Accuracy of Initial Model: 0.99375
/usr/lib/spark/python/lib/pyspark.zip/pyspark/sql/context
Weighted f1 score of Initial Model: 0.9933730158730157
Best Model: PipelineModel_55a5c603be56
Test Accuracy of Best Model: 0.96875
Weighted f1 score of Best Model: 0.9547916666666667
```

Running Single Machine App in EC2 without Docker

Fetch trained model from S3 to master instance into a new folder (trained_model) using below command:
aws s3 sync aws s3 sync s3://<my-bucket-name>/trained_model.model ./trained_model

Done SCP for <testdataset> and Single Machine Prediction code file using below command:
scp -i <keypair.pem> <file> ec2-user@<Public DNS>:/home/ec2-user

Installed Pyspark with: pip install pyspark
for running single machine prediction application with external test dataset, using below command:
python wqp_single_machine.py



```
Sample Predictions:
+-------------+----------------+-----------+----------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------
------------+----------+
|fixed acidity|volatile acidity|citric acid|residual sugar|chlorides|free sulfur dioxide|total sulfur dioxide|density|  pH|sulphates|alcohol|quality|            features|label|       rawPrediction|
  probability|prediction|
+-------------+----------------+-----------+----------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------
------------+----------+
|         8.9|            0.22|       0.48|             1.8|    0.077|                29.0|                 60.0| 0.9968|3.39|     0.53|    9.4|    6.0|[8.9,0.22,0.48,1....|  1.0|[19.4518621405359...|[0.0389
0372428107...|       1.0|
|         7.6|            0.39|       0.31|             2.3|    0.082|                23.0|                 71.0| 0.9982|3.52|     0.65|    9.7|    5.0|[7.6,0.39,0.31,2....|  0.0|[479.531456024878...|[0.9590
6291204975...|       0.0|
|         7.9|            0.43|       0.21|             1.6|    0.106|                10.0|                 37.0| 0.9966|3.17|     0.91|    9.5|    5.0|[7.9,0.43,0.21,1....|  0.0|[484.885809101067...|[0.9697
7161820213...|       0.0|
|         8.5|            0.49|       0.11|             2.3|    0.084|                 9.0|                 67.0| 0.9968|3.17|     0.53|    9.4|    5.0|[8.5,0.49,0.11,2....|  0.0|[491.163209482913...|[0.9823
2641896582...|       0.0|
|         6.9|             0.4|       0.14|             2.4|    0.085|                21.0|                 40.0| 0.9968|3.43|     0.63|    9.7|    6.0|[6.9,0.4,0.14,2.4...|  1.0|[3.29191168480642...|[0.0065
8382336961...|       1.0|
+-------------+----------------+-----------+----------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------
------------+----------+
only showing top 5 rows

None
Test Accuracy of the Wine Quality Prediction Model: 0.9843627834245504
/usr/local/lib/python3.7/site-packages/pyspark/sql/context.py:159: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
  FutureWarning,
Weighted F1 score of the Wine Quality Prediction Model: 0.9779339666913879
```

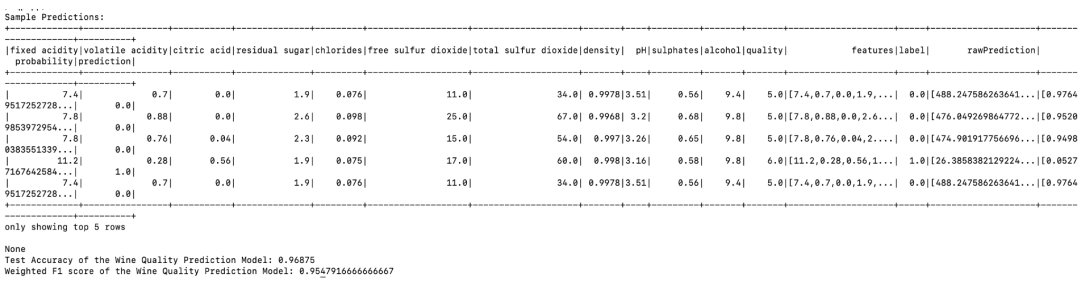
Running Single Machine App in Local with Docker:

Fetch trained model from master instance into local (in a new folder wine_quality_predictor) using below command:
scp -i <keypair.pem> -r ec2-user@<Public DNS>:/home/ec2-user/trained_model ./wine_quality_predictor

Install and sign up into Docker Desktop, login to Docker from terminal with command: ‘docker login’

cd to the source folder and build the image of the app: ‘docker build -t <image_name:version> .’

run the app: ‘docker run <image_name:version>’



```
Sample Predictions:
+-------------+----------------+-----------+----------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------
------------+----------+
|fixed acidity|volatile acidity|citric acid|residual sugar|chlorides|free sulfur dioxide|total sulfur dioxide|density|  pH|sulphates|alcohol|quality|            features|label|       rawPrediction|
  probability|prediction|
+-------------+----------------+-----------+----------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------
------------+----------+
|         7.4|             0.7|        0.0|             1.9|    0.076|                11.0|                 34.0| 0.9978|3.51|     0.56|    9.4|    5.0|[7.4,0.7,0.0,1.9,...|  0.0|[488.247586263641...|[0.9764
9517252728...|       0.0|
|         7.8|            0.88|        0.0|             2.6|    0.098|                25.0|                 67.0| 0.9968| 3.2|     0.68|    9.8|    5.0|[7.8,0.88,0.0,2.6...|  0.0|[476.049269864772...|[0.9520
9853972954...|       0.0|
|         7.8|            0.76|       0.04|             2.3|    0.092|                15.0|                 54.0|  0.997|3.26|     0.65|    9.8|    5.0|[7.8,0.76,0.04,2....|  0.0|[474.901917756696...|[0.9498
0383551339...|       0.0|
|        11.2|            0.28|       0.56|             1.9|    0.075|                17.0|                 60.0|  0.998|3.16|     0.58|    9.8|    6.0|[11.2,0.28,0.56,1...|  1.0|[26.3858382129224...|[0.0527
7167642584...|       1.0|
|         7.4|             0.7|        0.0|             1.9|    0.076|                11.0|                 34.0| 0.9978|3.51|     0.56|    9.4|    5.0|[7.4,0.7,0.0,1.9,...|  0.0|[488.247586263641...|[0.9764
9517252728...|       0.0|
+-------------+----------------+-----------+----------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------
------------+----------+
only showing top 5 rows

None
Test Accuracy of the Wine Quality Prediction Model: 0.96875
Weighted F1 score of the Wine Quality Prediction Model: 0.9547916666666667
```

run the app with external dataset: docker run -v /Users/<username>/<path_to_dataset_directory>:/<path_on_host> <image_name:version> <external_dataset>

```
Sample Predictions:
+-------------+----------+-----------+--------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------+
|fixed acidity|volatile acidity|citric acid|residual sugar|chlorides|free sulfur dioxide|total sulfur dioxide|density|  pH|sulphates|alcohol|quality|            features|label|       rawPrediction|
  probability|prediction|
+-------------+----------+-----------+--------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------+
|          8.9|           0.22|       0.48|           1.8|    0.077|                29.0|                60.0| 0.9968|3.39|     0.53|    9.4|    6.0|[8.9,0.22,0.48,1....|  1.0|[19.4518621405359...|[0.0389
0372428107...|       1.0|
|          7.6|           0.39|       0.31|           2.3|    0.082|                23.0|                71.0| 0.9982|3.52|     0.65|    9.7|    5.0|[7.6,0.39,0.31,2....|  0.0|[479.531456024878...|[0.9590
6291204975...|       0.0|
|          7.9|           0.43|       0.21|           1.6|    0.106|                10.0|                37.0| 0.9966|3.17|     0.91|    9.5|    5.0|[7.9,0.43,0.21,1....|  0.0|[484.885809101067...|[0.9697
7161820213...|       0.0|
|          8.5|           0.49|       0.11|           2.3|    0.084|                 9.0|                67.0| 0.9968|3.17|     0.53|    9.4|    5.0|[8.5,0.49,0.11,2....|  0.0|[491.163209482913...|[0.9823
2641896582...|       0.0|
|          6.9|            0.4|       0.14|           2.4|    0.085|                21.0|                40.0| 0.9968|3.43|     0.63|    9.7|    6.0|[6.9,0.4,0.14,2.4...|  1.0|[3.29191168480642...|[0.0065
8382336961...|       1.0|
+-------------+----------+-----------+--------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------+
only showing top 5 rows

None
Test Accuracy of the Wine Quality Prediction Model: 0.9843627834245504
Weighted F1 score of the Wine Quality Prediction Model: 0.9779339666913879
```

Create a new repo in DockerHub.

Push the image into repo with the below commands:
docker tag <img:ver> <user_id>/<repo>:<tag>
docker push <user_id>/<repo>:<tag>

Running Single Machine App in EC2 with Docker:

Create an EC2 instance with default configurations

Install docker with command: sudo yum install docker

Start docker: sudo service docker start

Pull the image into EC2: sudo docker pull <user_id>/<repo>:<tag>

Run the app: sudo docker run <user_id>/<repo>:<tag>


```
Sample Predictions:
+-------------+----------+-----------+--------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------+
|fixed acidity|volatile acidity|citric acid|residual sugar|chlorides|free sulfur dioxide|total sulfur dioxide|density|  pH|sulphates|alcohol|quality|            features|label|       rawPrediction|
  probability|prediction|
+-------------+----------+-----------+--------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------+
|          8.9|           0.22|       0.48|           1.8|    0.077|                29.0|                60.0| 0.9968|3.39|     0.53|    9.4|    6.0|[8.9,0.22,0.48,1....|  1.0|[19.4518621405359...|[0.0389
0372428107...|       1.0|
|          7.6|           0.39|       0.31|           2.3|    0.082|                23.0|                71.0| 0.9982|3.52|     0.65|    9.7|    5.0|[7.6,0.39,0.31,2....|  0.0|[479.531456024878...|[0.9590
6291204975...|       0.0|
|          7.9|           0.43|       0.21|           1.6|    0.106|                10.0|                37.0| 0.9966|3.17|     0.91|    9.5|    5.0|[7.9,0.43,0.21,1....|  0.0|[484.885809101067...|[0.9697
7161820213...|       0.0|
|          8.5|           0.49|       0.11|           2.3|    0.084|                 9.0|                67.0| 0.9968|3.17|     0.53|    9.4|    5.0|[8.5,0.49,0.11,2....|  0.0|[491.163209482913...|[0.9823
2641896582...|       0.0|
|          6.9|            0.4|       0.14|           2.4|    0.085|                21.0|                40.0| 0.9968|3.43|     0.63|    9.7|    6.0|[6.9,0.4,0.14,2.4...|  1.0|[3.29191168480642...|[0.0065
8382336961...|       1.0|
+-------------+----------+-----------+--------------+---------+--------------------+---------------------+-------+----+---------+-------+-------+--------------------+-----+--------------------+-------+
only showing top 5 rows

None
Test Accuracy of the Wine Quality Prediction Model: 0.9843627834245504
Weighted F1 score of the Wine Quality Prediction Model: 0.9779339666913879
```