CS643 - PROGRAMMING ASSIGNMENT 2
Wine Quality Prediction ML Model

This project involves creating a Python application that utilizes the PySpark interface. The application runs on an Amazon Web Services (AWS) Elastic MapReduce (EMR) cluster. Its main goal is to train a machine learning model in parallel on EC2 instances to predict wine quality using publicly accessible data. After training, the model is used to predict wine quality. Docker is used to produce a container image for the trained machine learning model, simplifying the deployment process.

The primary Python source files in this project are:

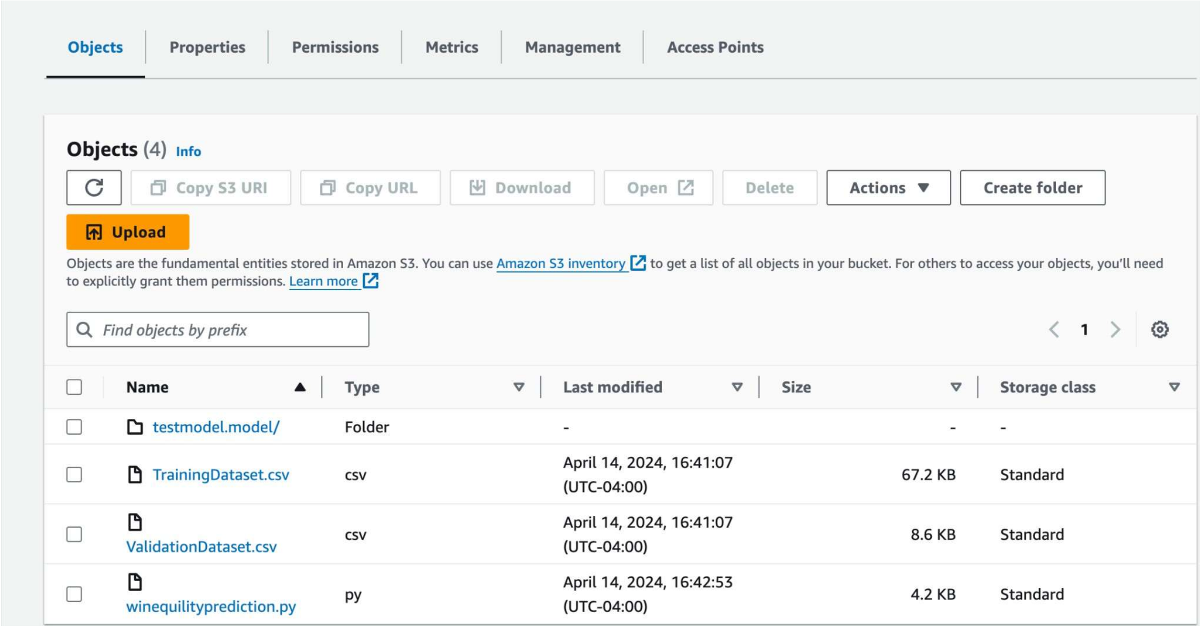
1. **winequilityprediction.py:** Reads the training dataset from S3 and trains the model in parallel on an EMR Spark cluster. Once trained, the model can be executed on provided test data via S3. The program stores the trained model in the S3 bucket.
2. **winequilitytestdataprediction.py:** Loads the trained model and executes it on a given test data file. This program prints the F1 score as a metric for the accuracy of the trained model.

GitHub: https://github.com/iamshashwat10/cloudComputing

Docker: https://hub.docker.com/r/iamshashwat10/pa2-docker

Steps to Launch an EMR Cluster on AWS and Train the ML Model without Docker:

- Create an S3 bucket and upload the following files: winequalityprediction.py, TrainingDataset.csv and ValidationDataset.csv in the bucket.



Objects (4) Info

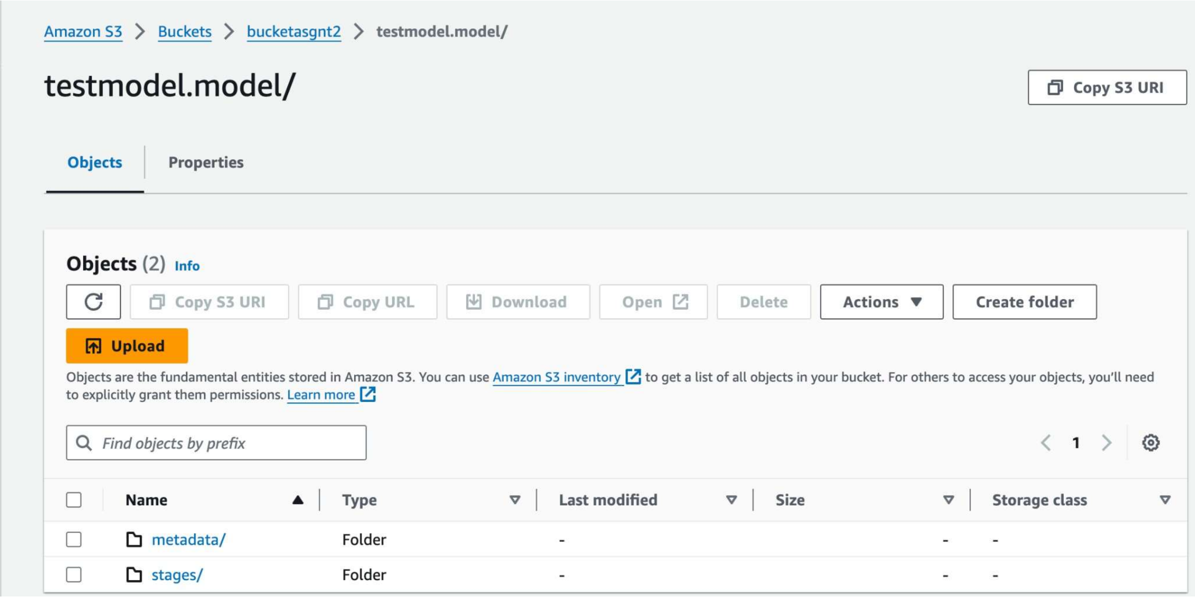
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type	Last modified	Size	Storage class
□	testmodel.model/	Folder	-	-	-
□	TrainingDataset.csv	csv	April 14, 2024, 16:41:07 (UTC-04:00)	67.2 KB	Standard
□	ValidationDataset.csv	csv	April 14, 2024, 16:41:07 (UTC-04:00)	8.6 KB	Standard
□	winequilityprediction.py	py	April 14, 2024, 16:42:53 (UTC-04:00)	4.2 KB	Standard

- Navigate to the EMR Service and configure the EMR cluster.
    a. Provide the details such as Cluster Configuration: (Spark, Hadoop).
    b. Select the EC2 instance type for the cluster nodes.
    c. Select the number of instances for the cluster.
    d. Provide the EC2 key pair that will be used to SSH connect with the instance.
- Connect to the EC2 instance using the SSH command and use the key pair defined above.
- After the connection is successful, submit the task for execution.

  Command: spark-submit s3://bucketasgnt2/winequilityprediction.py

- Once the task execution is complete the model will be generated in the S3 bucket defined above.



testmodel.model/

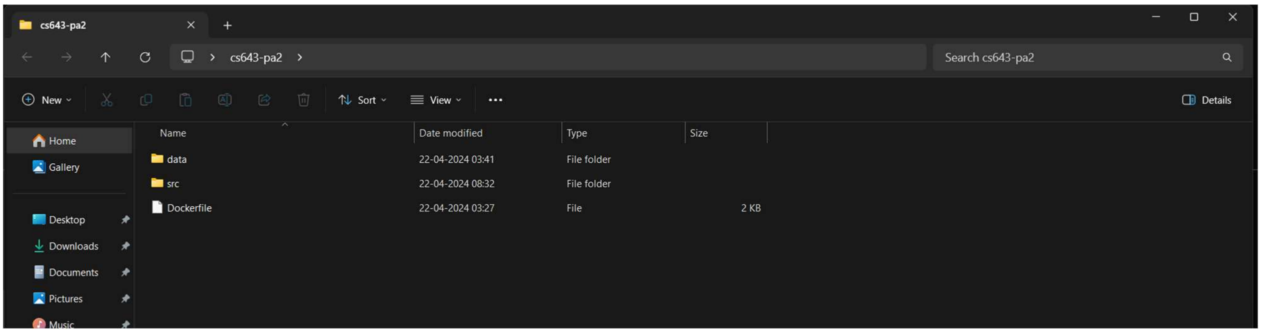
Objects (2)

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
□	metadata/	Folder	-	-	-
□	stages/	Folder	-	-	-

Steps for running the ML Model with Docker:

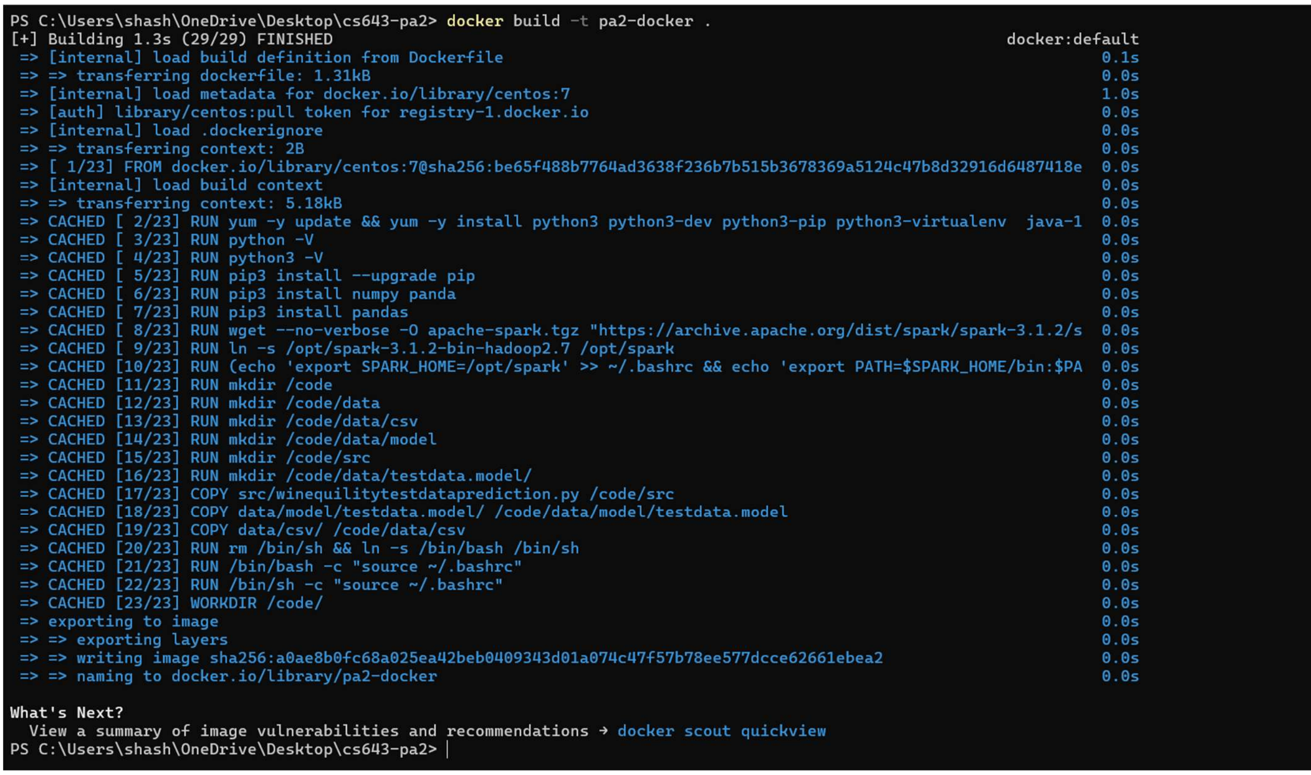
- Open Terminal and navigate to the folder where the Docker File is present.



Name	Date modified	Type	Size
data	22-04-2024 03:41	File folder	
src	22-04-2024 08:32	File folder	
Dockerfile	22-04-2024 03:27	File	2 KB

- Run the Docker build command to build the Docker image.

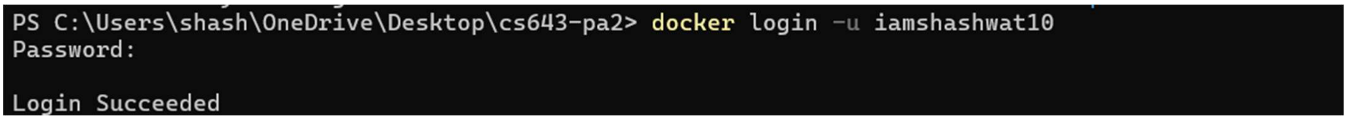
  Command: docker build -t pa2-docker .



```
PS C:\Users\shash\OneDrive\Desktop\cs643-pa2> docker build -t pa2-docker .
[+] Building 1.3s (29/29) FINISHED                                                            docker:default
 => [internal] load build definition from Dockerfile                                                   0.1s
 => => transferring dockerfile: 1.31kB                                                                 0.0s
 => [internal] load metadata for docker.io/library/centos:7                                            1.0s
 => [auth] library/centos:pull token for registry-1.docker.io                                          0.0s
 => [internal] load .dockerignore                                                                      0.0s
 => => transferring context: 2B                                                                        0.0s
 => [ 1/23] FROM docker.io/library/centos:7@sha256:be65f488b7764ad3638f236b7b515b3678369a5124c47b8d32916d6487418e  0.0s
 => [internal] load build context                                                                      0.0s
 => => transferring context: 5.18kB                                                                    0.0s
 => CACHED [ 2/23] RUN yum -y update && yum -y install python3 python3-dev python3-pip python3-virtualenv  java-1  0.0s
 => CACHED [ 3/23] RUN python -V                                                                       0.0s
 => CACHED [ 4/23] RUN python3 -V                                                                      0.0s
 => CACHED [ 5/23] RUN pip3 install --upgrade pip                                                      0.0s
 => CACHED [ 6/23] RUN pip3 install numpy panda                                                        0.0s
 => CACHED [ 7/23] RUN pip3 install pandas                                                             0.0s
 => CACHED [ 8/23] RUN wget --no-verbose -O apache-spark.tgz "https://archive.apache.org/dist/spark/spark-3.1.2/s  0.0s
 => CACHED [ 9/23] RUN ln -s /opt/spark-3.1.2-bin-hadoop2.7 /opt/spark                                 0.0s
 => CACHED [10/23] RUN (echo 'export SPARK_HOME=/opt/spark' >> ~/.bashrc && echo 'export PATH=$SPARK_HOME/bin:$PA  0.0s
 => CACHED [11/23] RUN mkdir /code                                                                     0.0s
 => CACHED [12/23] RUN mkdir /code/data                                                                0.0s
 => CACHED [13/23] RUN mkdir /code/data/csv                                                            0.0s
 => CACHED [14/23] RUN mkdir /code/data/model                                                          0.0s
 => CACHED [15/23] RUN mkdir /code/src                                                                 0.0s
 => CACHED [16/23] RUN mkdir /code/data/testdata.model/                                                0.0s
 => CACHED [17/23] COPY src/winequilitytestdataprediction.py /code/src                                 0.0s
 => CACHED [18/23] COPY data/model/testdata.model/ /code/data/model/testdata.model                     0.0s
 => CACHED [19/23] COPY data/csv/ /code/data/csv                                                       0.0s
 => CACHED [20/23] RUN rm /bin/sh && ln -s /bin/bash /bin/sh                                           0.0s
 => CACHED [21/23] RUN /bin/bash -c "source ~/.bashrc"                                                 0.0s
 => CACHED [22/23] RUN /bin/sh -c "source ~/.bashrc"                                                   0.0s
 => CACHED [23/23] WORKDIR /code/                                                                      0.0s
 => exporting to image                                                                                 0.0s
 => => exporting layers                                                                                0.0s
 => => writing image sha256:a0ae8b0fc68a025ea42beb0409343d01a074c47f57b78ee577dcce62661ebea2            0.0s
 => => naming to docker.io/library/pa2-docker                                                          0.0s

What's Next?
  View a summary of image vulnerabilities and recommendations → docker scout quickview
PS C:\Users\shash\OneDrive\Desktop\cs643-pa2> |
```

- Now login to Docker hub to push the image.

  Command: docker login -u iamshashwat10



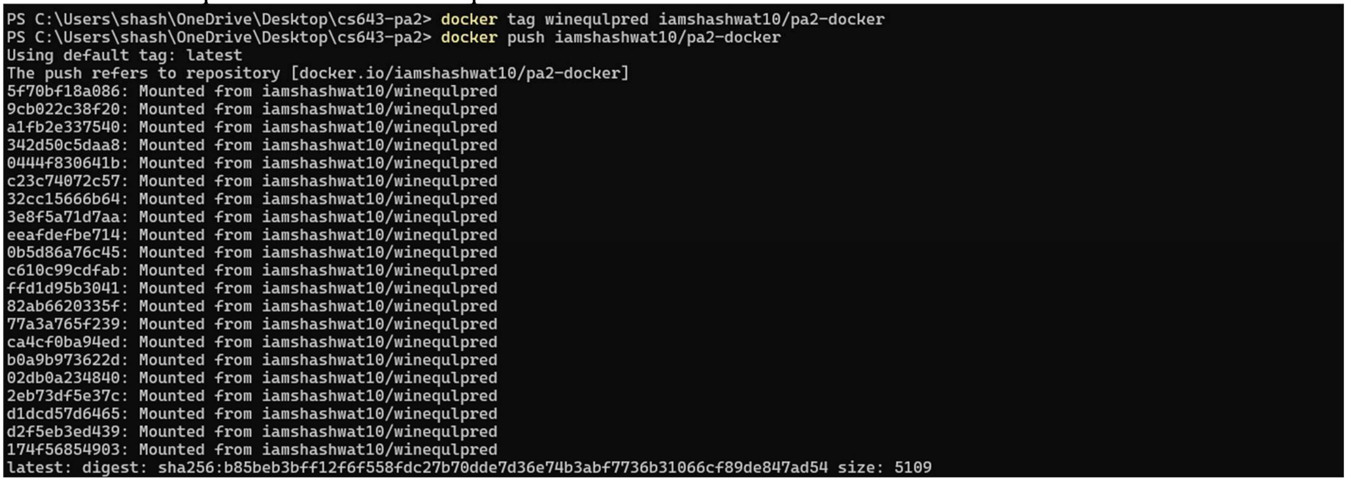
```
PS C:\Users\shash\OneDrive\Desktop\cs643-pa2> docker login -u iamshashwat10
Password:

Login Succeeded
```

- Run the below command to push the image on Docker hub.

  Command:

  docker tag pa2-docker iamshashwat10/pa2-docker

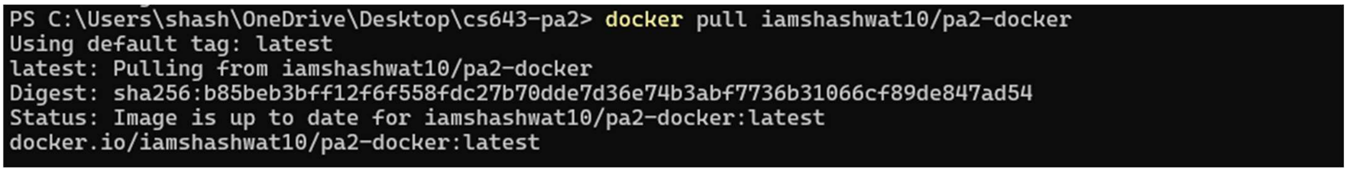
  docker push iamshashwat10/pa2-docker



```
PS C:\Users\shash\OneDrive\Desktop\cs643-pa2> docker tag winequlpred iamshashwat10/pa2-docker
PS C:\Users\shash\OneDrive\Desktop\cs643-pa2> docker push iamshashwat10/pa2-docker
Using default tag: latest
The push refers to repository [docker.io/iamshashwat10/pa2-docker]
5f70bf18a086: Mounted from iamshashwat10/winequlpred
9cb022c38f20: Mounted from iamshashwat10/winequlpred
a1fb2e337540: Mounted from iamshashwat10/winequlpred
342d50c5daa8: Mounted from iamshashwat10/winequlpred
0444f830641b: Mounted from iamshashwat10/winequlpred
c23c74072c57: Mounted from iamshashwat10/winequlpred
32cc15666b64: Mounted from iamshashwat10/winequlpred
3e8f5a71d7aa: Mounted from iamshashwat10/winequlpred
eeafdefbe714: Mounted from iamshashwat10/winequlpred
0b5d86a76c45: Mounted from iamshashwat10/winequlpred
c610c99cdfab: Mounted from iamshashwat10/winequlpred
ffd1d95b3041: Mounted from iamshashwat10/winequlpred
82ab6620335f: Mounted from iamshashwat10/winequlpred
77a3a765f239: Mounted from iamshashwat10/winequlpred
ca4cf0ba94ed: Mounted from iamshashwat10/winequlpred
b0a9b973622d: Mounted from iamshashwat10/winequlpred
02db0a234840: Mounted from iamshashwat10/winequlpred
2eb73df5e37c: Mounted from iamshashwat10/winequlpred
d1dcd57d6465: Mounted from iamshashwat10/winequlpred
d2f5eb3ed439: Mounted from iamshashwat10/winequlpred
174f56854903: Mounted from iamshashwat10/winequlpred
latest: digest: sha256:b85beb3bff12f6f558fdc27b70dde7d36e74b3abf7736b31066cf89de847ad54 size: 5109
```

- Now pull the image from the docker hub on the machine where you want to run the Docker image.

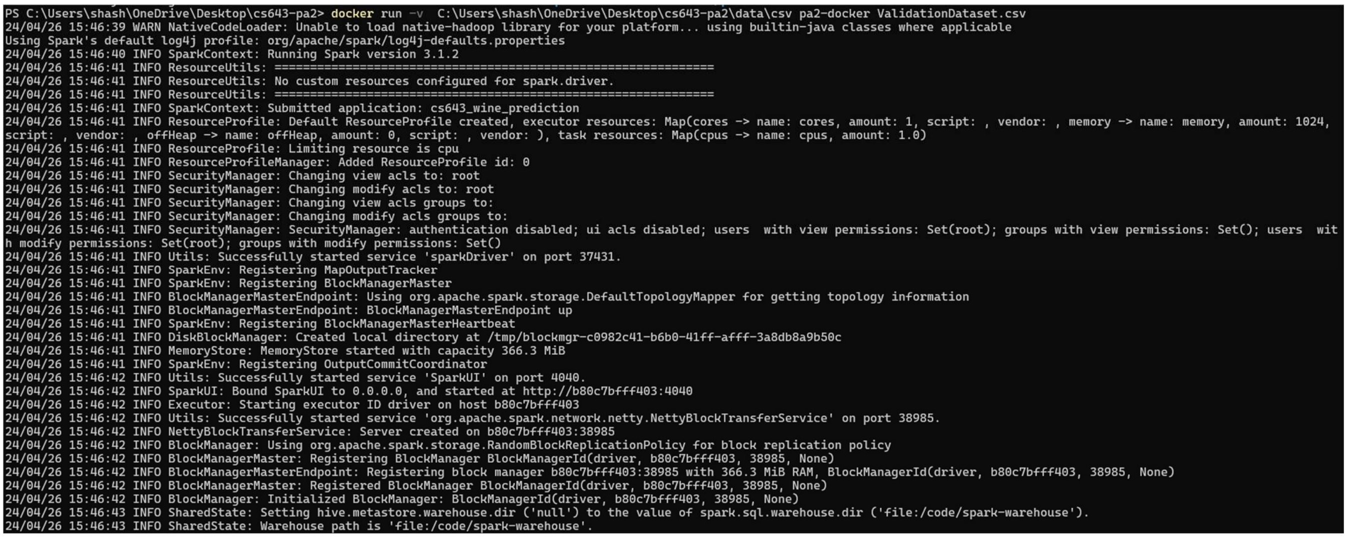
  Command: docker pull iamshashwat10/pa2-docker



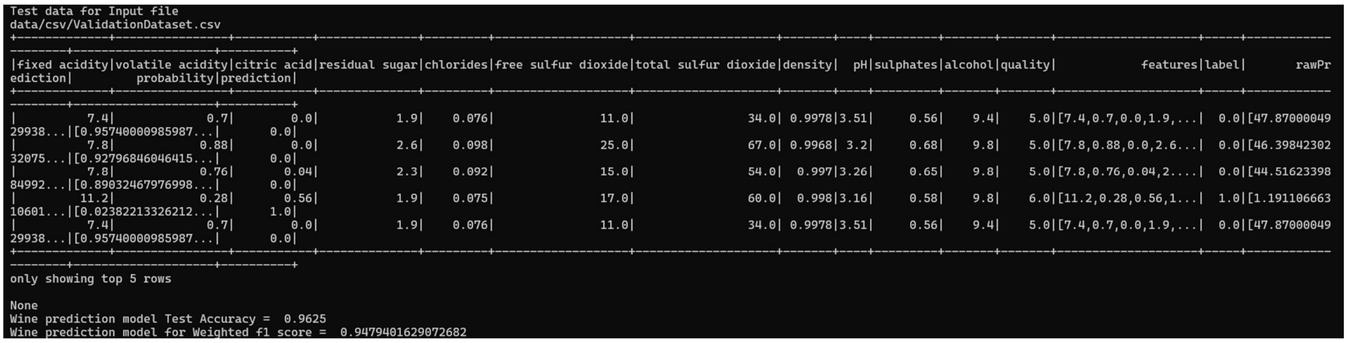
```
PS C:\Users\shash\OneDrive\Desktop\cs643-pa2> docker pull iamshashwat10/pa2-docker
Using default tag: latest
latest: Pulling from iamshashwat10/pa2-docker
Digest: sha256:b85beb3bff12f6f558fdc27b70dde7d36e74b3abf7736b31066cf89de847ad54
Status: Image is up to date for iamshashwat10/pa2-docker:latest
docker.io/iamshashwat10/pa2-docker:latest
```

- Now run the Docker run command to execute the image and see the results.

  Command: docker run -v C:\Users\shash\OneDrive\Desktop\cs643-pa2\data\csv pa2-docker ValidationDataset.csv



```
PS C:\Users\shash\OneDrive\Desktop\cs643-pa2> docker run -v  C:\Users\shash\OneDrive\Desktop\cs643-pa2\data\csv pa2-docker ValidationDataset.csv
24/04/26 15:46:39 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
24/04/26 15:46:40 INFO SparkContext: Running Spark version 3.1.2
24/04/26 15:46:41 INFO ResourceUtils: ==============================================================
24/04/26 15:46:41 INFO ResourceUtils: No custom resources configured for spark.driver.
24/04/26 15:46:41 INFO ResourceUtils: ==============================================================
24/04/26 15:46:41 INFO SparkContext: Submitted application: cs643_wine_prediction
24/04/26 15:46:41 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024,
script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/04/26 15:46:41 INFO ResourceProfile: Limiting resource is cpu
24/04/26 15:46:41 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/04/26 15:46:41 INFO SecurityManager: Changing view acls to: root
24/04/26 15:46:41 INFO SecurityManager: Changing modify acls to: root
24/04/26 15:46:41 INFO SecurityManager: Changing view acls groups to:
24/04/26 15:46:41 INFO SecurityManager: Changing modify acls groups to:
24/04/26 15:46:41 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(root); groups with view permissions: Set(); users  wit
h modify permissions: Set(root); groups with modify permissions: Set()
24/04/26 15:46:41 INFO Utils: Successfully started service 'sparkDriver' on port 37431.
24/04/26 15:46:41 INFO SparkEnv: Registering MapOutputTracker
24/04/26 15:46:41 INFO SparkEnv: Registering BlockManagerMaster
24/04/26 15:46:41 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/04/26 15:46:41 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/04/26 15:46:41 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/26 15:46:41 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-c0982c41-b6b0-41ff-afff-3a8db8a9b50c
24/04/26 15:46:41 INFO MemoryStore: MemoryStore started with capacity 366.3 MiB
24/04/26 15:46:41 INFO SparkEnv: Registering OutputCommitCoordinator
24/04/26 15:46:42 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/26 15:46:42 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://b80c7bfff403:4040
24/04/26 15:46:42 INFO Executor: Starting executor ID driver on host b80c7bfff403
24/04/26 15:46:42 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 38985.
24/04/26 15:46:42 INFO NettyBlockTransferService: Server created on b80c7bfff403:38985
24/04/26 15:46:42 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/04/26 15:46:42 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, b80c7bfff403, 38985, None)
24/04/26 15:46:42 INFO BlockManagerMasterEndpoint: Registering block manager b80c7bfff403:38985 with 366.3 MiB RAM, BlockManagerId(driver, b80c7bfff403, 38985, None)
24/04/26 15:46:42 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, b80c7bfff403, 38985, None)
24/04/26 15:46:42 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, b80c7bfff403, 38985, None)
24/04/26 15:46:43 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('file:/code/spark-warehouse').
24/04/26 15:46:43 INFO SharedState: Warehouse path is 'file:/code/spark-warehouse'.
```



```
Test data for Input file
data/csv/ValidationDataset.csv
+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+----+---------+-------+-------+--------------------+-----+------------
--------+--------------------+----------+
|fixed acidity|volatile acidity|citric acid|residual sugar|chlorides|free sulfur dioxide|total sulfur dioxide|density|  pH|sulphates|alcohol|quality|            features|label|       rawPr
ediction|         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|[47.87000049
29938...|[0.95740000985987...|       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|[46.39842302
32075...|[0.92796846046415...|       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|[44.51623398
84992...|[0.89032467976998...|       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|[1.191106663
10601...|[0.02382213326212...|       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|[47.87000049
29938...|[0.95740000985987...|       0.0|
+-------------+----------------+-----------+--------------+---------+-------------------+--------------------+-------+----+---------+-------+-------+--------------------+-----+------------
--------+--------------------+----------+
only showing top 5 rows

None
Wine prediction model Test Accuracy =  0.9625
Wine prediction model for Weighted f1 score =  0.9479401629072682
```