

Wine Quality Prediction System using PySpark:

This project implements a distributed machine learning system for wine quality prediction using Apache Spark ML on AWS EC2. The system includes training and prediction components, containerized with Docker for easy deployment.

Dockerhub Repository : <https://hub.docker.com/repository/docker/chandra459/wine-predictor/general>

Launch EC2 Instances:

- **Instance Type:** Select an instance type like t2.large or m5.large for sufficient resources.
- **Ensure instances are in the same VPC for network connectivity.**

Environment Setup (install in all 4 ec2 instances)

Python Dependencies

- PySpark
- NumPy
- Pandas

SSH into the Instances

SSH into each EC2 instance using:

bash

Copy code

```
ssh -i "your-key.pem" ubuntu@<instance-public-ip>
```

Passphrase-less SSH

Generate a pair of authentication keys on each instance using:

```
ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa
```

Append each instance's public key(id_rsa.pub) to other instances'

```
authorized_keys
```

```
cat ~/.ssh/id_rsa.pub
```

```
nano ~/.ssh/authorized_keys
```

Correct Command to Install OpenJDK 17

bash

Copy code

```
sudo apt update && sudo apt upgrade -y
```

```
sudo apt install openjdk-17-jdk wget unzip -y
```

```
java -version # Verify Java installation
```

Install Hadoop

Download Hadoop

bash

Copy code

```
wget https://downloads.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
```

```
tar -xvzf hadoop-3.3.6.tar.gz
```

```
sudo mv hadoop-3.3.6 /usr/local/hadoop
```

Configure Hadoop Environment Variables

Add the following lines to your ~/.bashrc file:

bash

Copy code

```
export HADOOP_HOME=/usr/local/hadoop
```

```
export HADOOP_INSTALL=$HADOOP_HOME
```

```
export HADOOP_MAPRED_HOME=$HADOOP_HOME
```

```
export HADOOP_COMMON_HOME=$HADOOP_HOME
```

```
export HADOOP_HDFS_HOME=$HADOOP_HOME
```

```
export YARN_HOME=$HADOOP_HOME
```

```
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
```

```
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
```

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

Reload the environment variables:

bash

Copy code

```
source ~/.bashrc
```

Configure Spark

Edit Spark's conf/spark-env.sh:

bash

Copy code

```
cp $SPARK_HOME/conf/spark-env.sh.template $SPARK_HOME/conf/spark-env.sh
```

```
nano $SPARK_HOME/conf/spark-env.sh
```

Add:

bash

Copy code

```
export SPARK_MASTER_HOST=<master-node-private-ip>
```

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

```
export HADOOP_CONF_DIR=/usr/local/hadoop/etc/hadoop
```

Configure Worker Nodes

Edit conf/slaves on the master node and add the private IPs of all worker nodes:

bash

Copy code

```
<worker-node-1-private-ip>
```

```
<worker-node-2-private-ip>
```

Start Spark

Start the Spark master:

bash

Copy code

start-master.sh

Start Spark workers on all worker nodes:

bash

Copy code

start-slave.sh spark://<master-node-private-ip>:7077

master:

```
[ec2-user@ip-172-31-21-112 ~]$ $SPARK_HOME/sbin/start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark/spark-3.4.1-bin-hadoop3/logs/spark-ec2-user-org.apache.spark.deploy.master.Master-1-ip-172-31-21-112.ec2.internal.out
[ec2-user@ip-172-31-21-112 ~]$
```

Worker1:

```
[ec2-user@ip-172-31-21-106 spark]$ $SPARK_HOME/sbin/start-worker.sh spark://172.31.21.112:7077
starting org.apache.spark.deploy.worker.Worker, logging to /home/ec2-user/spark/logs/spark-ec2-user-org.apache.spark.deploy.worker.Worker-1-ip-172-31-21-106.ec2.internal.out
[ec2-user@ip-172-31-21-106 spark]$
```

Worker2:

```
[ec2-user@ip-172-31-20-163 ~]$ $SPARK_HOME/sbin/start-worker.sh spark://172.31.21.112:7077
starting org.apache.spark.deploy.worker.Worker, logging to /home/ec2-user/spark/logs/spark-ec2-user-org.apache.spark.deploy.worker.Worker-1-ip-172-31-20-163.ec2.internal.out
[ec2-user@ip-172-31-20-163 ~]$
```

Worker3:

```
[ec2-user@ip-172-31-20-9 ~]$ $SPARK_HOME/sbin/start-worker.sh spark://172.31.21.112:7077
/home/ec2-user/spark/conf/spark-env.sh: line 1: GNU: command not found
starting org.apache.spark.deploy.worker.Worker, logging to /home/ec2-user/spark/logs/spark-ec2-user-org.apache.spark.deploy.worker.Worker-1-ip-172-31-20-9.ec2.internal.out
[ec2-user@ip-172-31-20-9 ~]$
```

Upload dataset to each instance using below cmd:

scp -i "your-key.pem" file-to-upload ubuntu@<instance-public-ip>:<remote-path>

Train the model:

bash

Copy

On master node

Python3 wine_training.py

Build and run Docker container for prediction:

bash

Copy

Build Docker image

docker build -t wine-predictor.py .

Run prediction

docker run – wine-predictor.py .

Docker Built Successfull Image:

```
at java.base/java.lang.Thread.run(Thread.java:840)

[ec2-user@ip-172-31-21-112 ~]$ docker build -t wine-predictor .
[+] Building 0.3s (13/13) FINISHED                                docker:default
=> [internal] load build definition from Dockerfile              0.0s
=> => transferring dockerfile: 2.63kB                             0.0s
=> [internal] load metadata for docker.io/library/eclipse-temurin:17-jre 0.1s
=> [auth] library/eclipse-temurin:pull token for registry-1.docker.io 0.0s
=> [internal] load .dockerignore                                  0.0s
=> => transferring context: 2B                                     0.0s
=> [1/7] FROM docker.io/library/eclipse-temurin:17-jre@sha256:14ea79cda7bf92f2cfc24a7e4a3143c8a23c722b499b1c1d083e07815467f6c 0.0s
=> [internal] load build context                                 0.0s
=> => transferring context: 384B                                   0.0s
=> CACHED [2/7] RUN set -ex && apt-get update && apt-get install -y curl bzip2 procps --no-install-recommends && curl -s -L --url "https://repo.continuum.io/miniconda/M 0.0s
=> CACHED [3/7] WORKDIR /mlprog                                  0.0s
=> CACHED [4/7] COPY wine_train.py .                             0.0s
=> CACHED [5/7] COPY wine_test.py .                             0.0s
=> CACHED [6/7] COPY TrainingDataset.csv .                       0.0s
=> CACHED [7/7] COPY ValidationDataset.csv .                     0.0s
=> exporting to image                                           0.0s
=> exporting layers                                             0.0s
=> writing image sha256:191bb18a4af123154661d37c0355c0b362f3b8e122252a18e8ba2ab26b54b212 0.0s
=> naming to docker.io/library/wine-predictor                   0.0s
```

Wine_Predictor_Results:

Result img 1:

```
24/12/08 20:46:19 INFO DAGScheduler: Job 123 finished: collectAnMap at MulticlassMetrics.scala:61, took 0.090715 s
[Test] F1 score = 0.5576016539163332
[Test] Accuracy = 0.5625
```

Result img 2:

```
24/12/08 20:46:18 INFO SparkContext: Created broadcast 245 from broadcast at DAGScheduler.scala:1513
24/12/08 20:46:18 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 129 (MapPartitionsRDD[207] at map at MulticlassMetrics.scala:52) (first 15 tasks are for partitions Vec
tor(0))
24/12/08 20:46:18 INFO TaskSchedulerImpl: Adding task set 129.0 with 1 tasks resource profile 0
24/12/08 20:46:18 INFO TaskSetManager: Starting task 0.0 in stage 129.0 (TID 126) (402abf26a424, executor driver, partition 0, PROCESS_LOCAL, 4896 bytes) taskResourceAssignments Map()
24/12/08 20:46:18 INFO Executor: Running task 0.0 in stage 129.0 (TID 126)
24/12/08 20:46:18 INFO FileScanRDD: Reading File path: file:///mlprog/ValidationDataset.csv, range: 0-8760, partition values: [empty row]
24/12/08 20:46:18 INFO Executor: Finished task 0.0 in stage 129.0 (TID 126). 1826 bytes result sent to driver
24/12/08 20:46:18 INFO TaskSetManager: Finished task 0.0 in stage 129.0 (TID 126) in 48 ms on 402abf26a424 (executor driver) (1/1)
24/12/08 20:46:18 INFO TaskSchedulerImpl: Removed TaskSet 129.0, whose tasks have all completed, from pool
24/12/08 20:46:18 INFO DAGScheduler: ShuffleMapStage 129 (map at MulticlassMetrics.scala:52) finished in 0.056 s
24/12/08 20:46:18 INFO DAGScheduler: looking for newly runnable stages
24/12/08 20:46:18 INFO DAGScheduler: running: Set()
24/12/08 20:46:18 INFO DAGScheduler: waiting: Set(resultStage 130)
24/12/08 20:46:18 INFO DAGScheduler: failed: Set()
24/12/08 20:46:18 INFO DAGScheduler: Submitting ResultStage 130 (ShuffledRDD[208] at reduceByKey at MulticlassMetrics.scala:61), which has no missing parents
24/12/08 20:46:18 INFO MemoryStore: Block broadcast_246 stored as values in memory (estimated size 4.7 KiB, free 432.8 MiB)
24/12/08 20:46:18 INFO MemoryStore: Block broadcast_246 piece0 stored as bytes in memory (estimated size 2.8 KiB, free 432.8 MiB)
24/12/08 20:46:18 INFO BlockManagerInfo: Added broadcast_246 piece0 in memory on 402abf26a424:38133 (size: 2.8 KiB, free: 434.1 MiB)
24/12/08 20:46:18 INFO SparkContext: Created broadcast 246 from broadcast at DAGScheduler.scala:1513
24/12/08 20:46:18 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 130 (ShuffledRDD[208]) at reduceByKey at MulticlassMetrics.scala:61 (first 15 tasks are for partitions Vect
or(0))
24/12/08 20:46:18 INFO TaskSchedulerImpl: Adding task set 130.0 with 1 tasks resource profile 0
24/12/08 20:46:18 INFO TaskSetManager: Starting task 0.0 in stage 130.0 (TID 127) (402abf26a424, executor driver, partition 0, NODE_LOCAL, 4271 bytes) taskResourceAssignments Map()
24/12/08 20:46:18 INFO Executor: Running task 0.0 in stage 130.0 (TID 127)
24/12/08 20:46:18 INFO ShuffleBlockFetcherIterator: Getting 1 (490.0 B) non-empty blocks including 1 (490.0 B) local and 0 (0.0 B) host-local and 0 (0.0 B) push-merged-local and 0 (0.0 B)
remote blocks
24/12/08 20:46:18 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
24/12/08 20:46:18 INFO Executor: Finished task 0.0 in stage 130.0 (TID 127). 2162 bytes result sent to driver
24/12/08 20:46:18 INFO TaskSetManager: Finished task 0.0 in stage 130.0 (TID 127) in 17 ms on 402abf26a424 (executor driver) (1/1)
24/12/08 20:46:18 INFO TaskSchedulerImpl: Removed TaskSet 130.0, whose tasks have all completed, from pool
24/12/08 20:46:18 INFO DAGScheduler: ResultStage 130 (collectAnMap at MulticlassMetrics.scala:61) finished in 0.024 s
24/12/08 20:46:18 INFO DAGScheduler: Job 123 is finished. Cancelling potential speculative or zombie tasks for this job
24/12/08 20:46:18 INFO TaskSchedulerImpl: Killing all running tasks in stage 130: Stage finished
24/12/08 20:46:18 INFO DAGScheduler: Job 123 finished: collectAnMap at MulticlassMetrics.scala:61, took 0.090715 s
[Test] F1 score = 0.5576016539163332
[Test] Accuracy = 0.5625
```

Docker push

docker push <dockerhub-username>/wine-predictor:latest