CS643 - AWS Spark Wine Quality Prediction Application

Shreyas Shende (ss4897)

The aim of this project is to train a machine learning model in parallel on EC2 instances for predicting wine quality using publicly available data and then use the trained model for predictions. The project also leverages Docker to create a container image for the trained machine learning model to simplify deployments.

Links

- **GitHub Code**: GitHub Repository
- **Docker Container Image**: Docker Hub

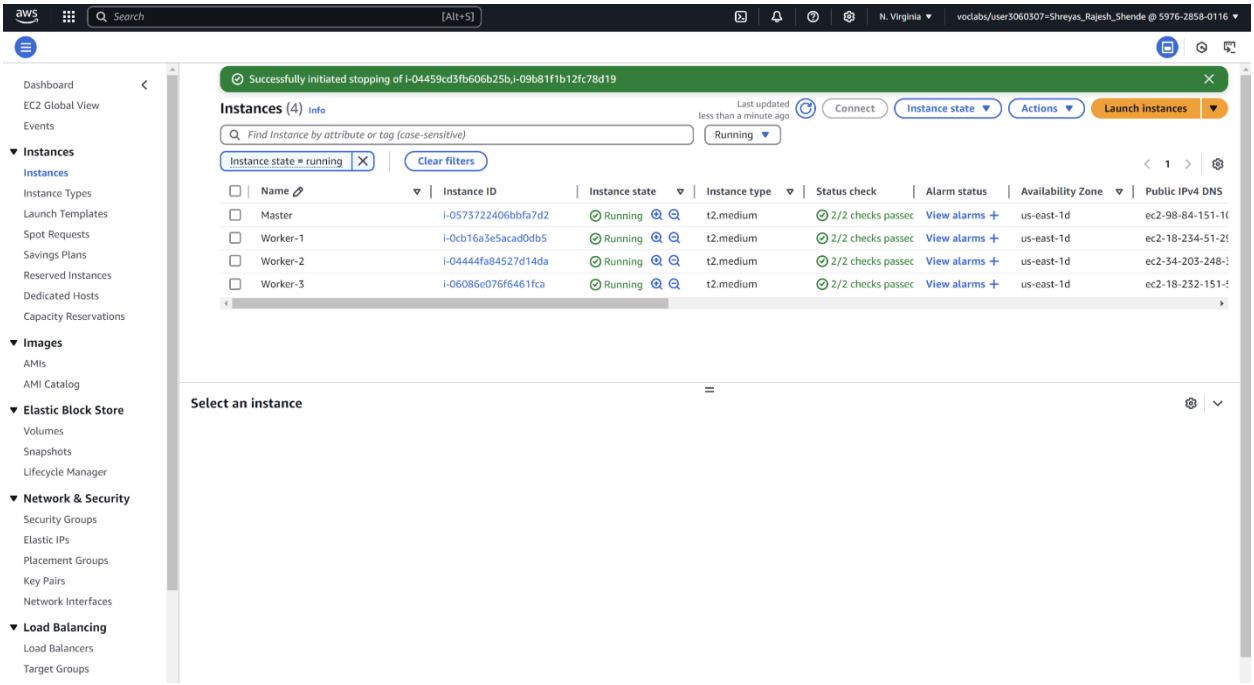
Source Files

1. WineQualityPrediction.java: Reads the training dataset from the file and creates a model.
2. WineQualityEval.java : Loads the trained model and executes it on the validation dataset, generating the F1 score.
3. Dockerfile: Creates a Docker image and runs a container for simplified deployment.

Instructions

Step 1: SSH into 4 Instances

SSH into each of the four instances using your public IP and pem key.



Step 2: Generate SSH Keys on All Instances

Generate SSH keys on all instances and note the public keys for sharing.

Step 3: Add Public Keys to authorized_keys

Add the public keys from all instances to the authorized_keys file on each instance to enable passwordless SSH between them.

Step 4: Edit /etc/hosts on All Instances

Add the IP addresses and corresponding hostnames of all instances to the /etc/hosts file on each instance.

Step 5: Install Java, Maven, and Spark

Install Java (OpenJDK 8), Maven, and Spark 3.4.1 on all instances. Configure the environment variables for Spark to make it available globally.

Step 6: Configure workers File

Edit the workers file under the Spark configuration directory and add the hostnames or IP addresses of all worker instances.

```
ubuntu@ip-172-31-17-8:~/spark-3.4.1/conf$ cat workers
#
# Licensed to the Apache Software Foundation (ASF) under one or more
# contributor license agreements.  See the NOTICE file distributed with
# this work for additional information regarding copyright ownership.
# The ASF licenses this file to You under the Apache License, Version 2.0
# (the "License"); you may not use this file except in compliance with
# the License.  You may obtain a copy of the License at
#
#    http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.
#

# A Spark Worker will be started on each of the machines listed below.
localhost
dd1
dd2
dd3
```

Step 7: Create Folders for Training and Evaluation

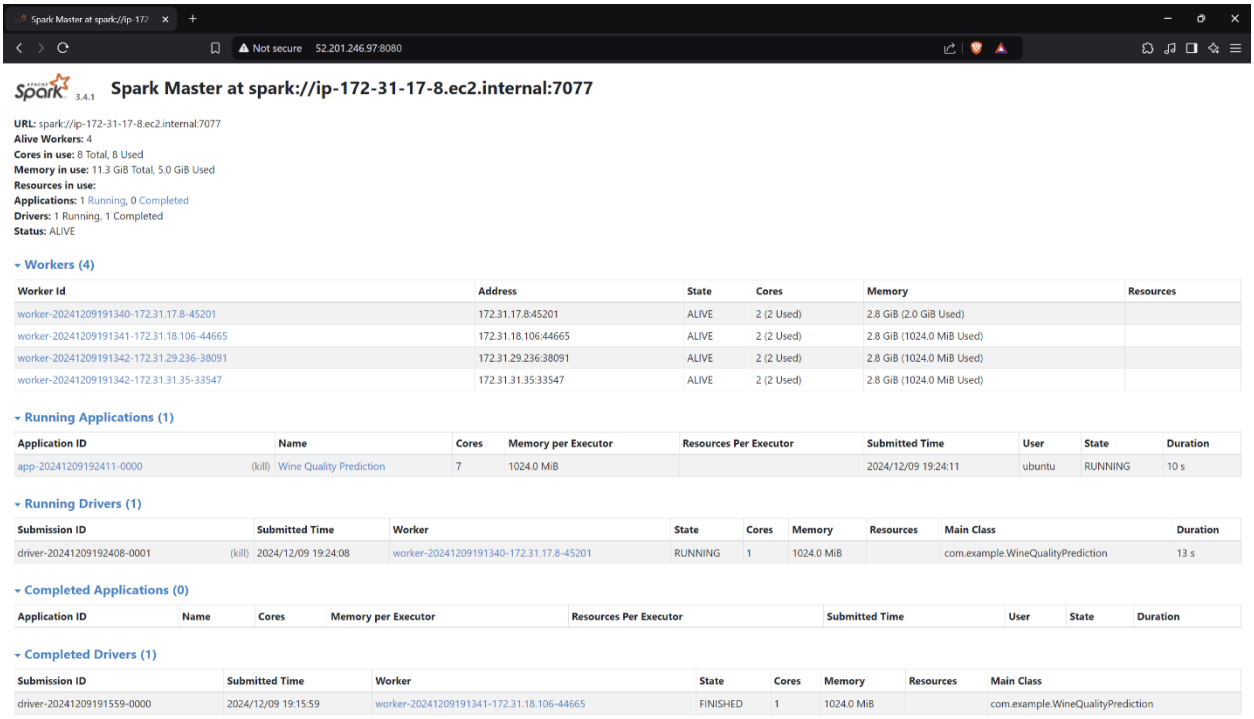
Create directories named Training and Eval on each instance. (WineQualityPrediction and WinQualityEval) Place the respective Java code files in these folders. (the code should be pasted in src/main/java/com/example/)

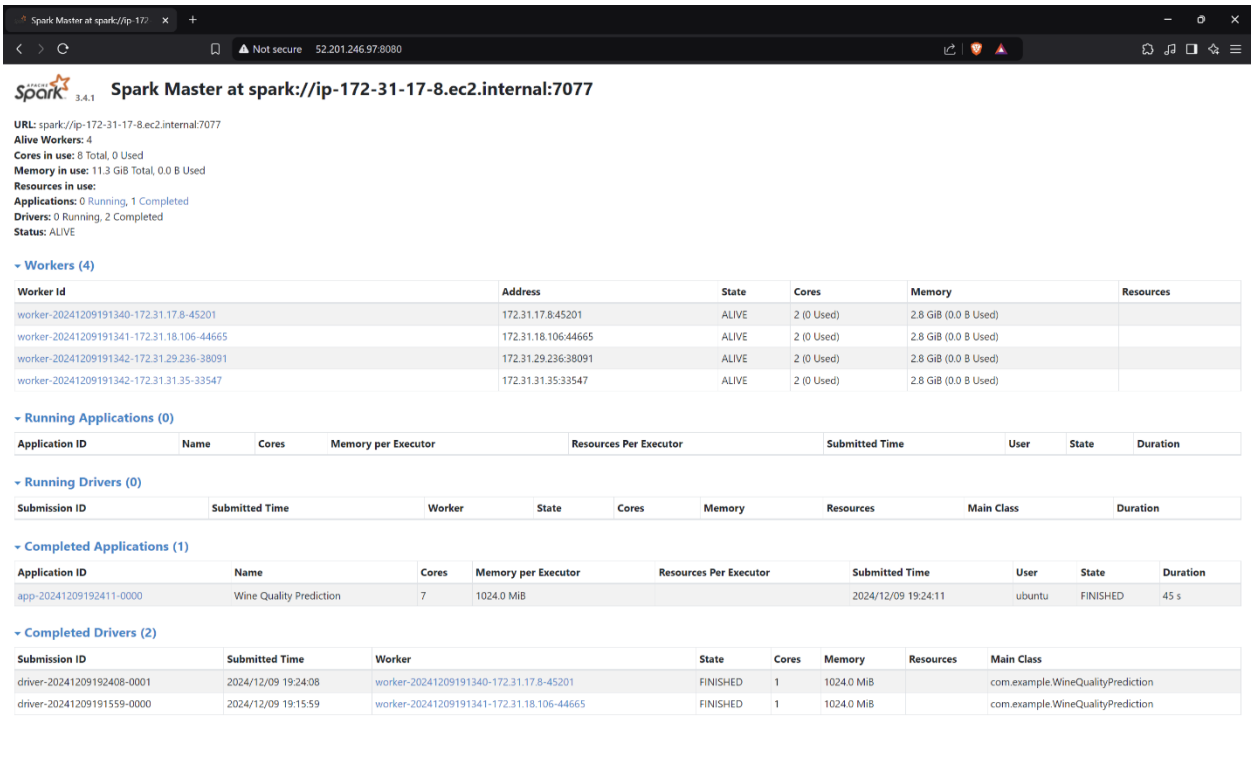
```
ubuntu@ip-172-31-17-8:~$ ls -lrt
total 379348
-rw-rw-r--  1 ubuntu ubuntu 388341449 Jun 19  2023 spark-3.4.1-bin-hadoop3.tgz
drwxr-xr-x 15 ubuntu ubuntu      4096 Dec  9 00:10 spark-3.4.1
-rw-rw-r--  1 ubuntu ubuntu     68804 Dec  9 00:14 TrainingDataset.csv
-rw-rw-r--  1 ubuntu ubuntu      8760 Dec  9 02:25 ValidationDataset.csv
drwxrwxr-x  4 ubuntu ubuntu      4096 Dec  9 03:00 WineQualityEval
drwx------  3 ubuntu ubuntu      4096 Dec  9 03:52 snap
-rw-rw-r--  1 ubuntu ubuntu      3839 Dec  9 05:24 WinePredModel.zip
drwxrwxr-x  4 ubuntu ubuntu      4096 Dec  9 19:19 WineQualityPrediction
drwxr-xr-x  4 ubuntu ubuntu      4096 Dec  9 19:24 WineQualityPredictionModel
```

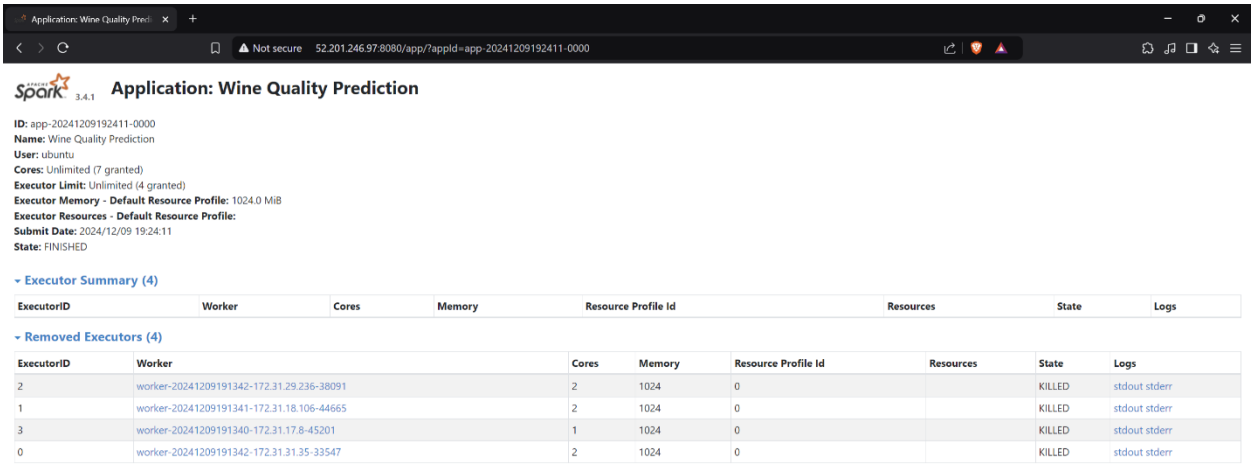
Step 8: Run the Training Code in Parallel

Use the spark-submit command to run the training code in parallel on all instances.

```
ubuntu@ip-172-31-17-8:~/WineQualityPrediction/target$ spark-submit --master spark://ip-172-31-17-8.ec2.internal:7077 --deploy-mode cluster --class com.example.WineQualityPredictio
n wine-quality-prediction-1.0-SNAPSHOT.jar
24/12/09 19:24:07 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/12/09 19:24:07 INFO SecurityManager: Changing view acls to: ubuntu
24/12/09 19:24:07 INFO SecurityManager: Changing modify acls to: ubuntu
24/12/09 19:24:07 INFO SecurityManager: Changing view acls groups to:
24/12/09 19:24:07 INFO SecurityManager: Changing modify acls groups to:
24/12/09 19:24:07 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: ubuntu; groups with view permissions: EMPTY; users
 with modify permissions: ubuntu; groups with modify permissions: EMPTY
24/12/09 19:24:07 INFO Utils: Successfully started service 'driverClient' on port 35435.
24/12/09 19:24:07 INFO TransportClientFactory: Successfully created connection to ip-172-31-17-8.ec2.internal/172.31.17.8:7077 after 40 ms (0 ms spent in bootstraps)
24/12/09 19:24:08 INFO ClientEndpoint: ... waiting before polling master for driver state
24/12/09 19:24:08 INFO ClientEndpoint: Driver successfully submitted as driver-20241209192408-0001
24/12/09 19:24:13 INFO ClientEndpoint: State of driver-20241209192408-0001 is RUNNING
24/12/09 19:24:13 INFO ClientEndpoint: Driver running on 172.31.17.8:45201 (worker-20241209191340-172.31.17.8-45201)
24/12/09 19:24:13 INFO ClientEndpoint: spark-submit not configured to wait for completion, exiting spark-submit JVM.
24/12/09 19:24:13 INFO ShutdownHookManager: Shutdown hook called
24/12/09 19:24:13 INFO ShutdownHookManager: Deleting directory /tmp/spark-9f34eed1-b0f5-431b-9052-1c88772f7219
```


Spark Master at spark://ip-172
Not secure 52.201.246.97:8080
Spark 3.4.1
Spark Master at spark://ip-172-31-17-8.ec2.internal:7077
URL: spark://ip-172-31-17-8.ec2.internal:7077
Alive Workers: 4
Cores in use: 8 Total, 8 Used
Memory in use: 11.3 GiB Total, 5.0 GiB Used
Resources in use:
Applications: 1 Running, 0 Completed
Drivers: 1 Running, 1 Completed
Status: ALIVE
Workers (4)
Worker Id | Address | State | Cores | Memory | Resources
worker-20241209191340-172.31.17.8-45201 | 172.31.17.8:45201 | ALIVE | 2 (2 Used) | 2.8 GiB (2.0 GiB Used)
worker-20241209191341-172.31.18.106-44665 | 172.31.18.106:44665 | ALIVE | 2 (2 Used) | 2.8 GiB (1024.0 MiB Used)
worker-20241209191342-172.31.29.236-38091 | 172.31.29.236:38091 | ALIVE | 2 (2 Used) | 2.8 GiB (1024.0 MiB Used)
worker-20241209191342-172.31.31.35-33547 | 172.31.31.35:33547 | ALIVE | 2 (2 Used) | 2.8 GiB (1024.0 MiB Used)
Running Applications (1)
Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration
app-20241209192411-0000 (kill) | Wine Quality Prediction | 7 | 1024.0 MiB | | 2024/12/09 19:24:11 | ubuntu | RUNNING | 10 s
Running Drivers (1)
Submission ID | Submitted Time | Worker | State | Cores | Memory | Resources | Main Class | Duration
driver-20241209192408-0001 (kill) | 2024/12/09 19:24:08 | worker-20241209191340-172.31.17.8-45201 | RUNNING | 1 | 1024.0 MiB | | com.example.WineQualityPrediction | 13 s
Completed Applications (0)
Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration
Completed Drivers (1)
Submission ID | Submitted Time | Worker | State | Cores | Memory | Resources | Main Class
driver-20241209191559-0000 | 2024/12/09 19:15:59 | worker-20241209191341-172.31.18.106-44665 | FINISHED | 1 | 1024.0 MiB | | com.example.WineQualityPrediction


Spark Master at spark://ip-172
Not secure 52.201.246.97:8080
Spark 3.4.1
Spark Master at spark://ip-172-31-17-8.ec2.internal:7077
URL: spark://ip-172-31-17-8.ec2.internal:7077
Alive Workers: 4
Cores in use: 8 Total, 0 Used
Memory in use: 11.3 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 2 Completed
Status: ALIVE
Workers (4)
Worker Id | Address | State | Cores | Memory | Resources
worker-20241209191340-172.31.17.8-45201 | 172.31.17.8:45201 | ALIVE | 2 (0 Used) | 2.8 GiB (0.0 B Used)
worker-20241209191341-172.31.18.106-44665 | 172.31.18.106:44665 | ALIVE | 2 (0 Used) | 2.8 GiB (0.0 B Used)
worker-20241209191342-172.31.29.236-38091 | 172.31.29.236:38091 | ALIVE | 2 (0 Used) | 2.8 GiB (0.0 B Used)
worker-20241209191342-172.31.31.35-33547 | 172.31.31.35:33547 | ALIVE | 2 (0 Used) | 2.8 GiB (0.0 B Used)
Running Applications (0)
Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration
Running Drivers (0)
Submission ID | Submitted Time | Worker | State | Cores | Memory | Resources | Main Class | Duration
Completed Applications (1)
Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration
app-20241209192411-0000 | Wine Quality Prediction | 7 | 1024.0 MiB | | 2024/12/09 19:24:11 | ubuntu | FINISHED | 45 s
Completed Drivers (2)
Submission ID | Submitted Time | Worker | State | Cores | Memory | Resources | Main Class
driver-20241209192408-0001 | 2024/12/09 19:24:08 | worker-20241209191340-172.31.17.8-45201 | FINISHED | 1 | 1024.0 MiB | | com.example.WineQualityPrediction
driver-20241209191559-0000 | 2024/12/09 19:15:59 | worker-20241209191341-172.31.18.106-44665 | FINISHED | 1 | 1024.0 MiB | | com.example.WineQualityPrediction



Step 9: Create Docker Image

Use the provided Dockerfile to build a Docker image that includes the trained model, validation dataset, and the evaluation code.

Use following command:

sudo docker build -t shreyasshende/wine-quality-eval:latest .

Step 10: Push Docker Image to Docker Hub

Push the Docker image to Docker Hub for easier access across instances.

Use the following command:

sudo docker push shreyasshende/wine-quality-eval:latest

Step 11: Pull the Docker Image

Pull the Docker image on the desired instances from Docker Hub.

sudo docker pull shreyasshende/wine-quality-eval:latest

```
ubuntu@ip-172-31-17-8:~$ sudo docker pull shreyasshende/wine-quality-eval:latest
latest: Pulling from shreyasshende/wine-quality-eval
Digest: sha256:b83b8d8d552897b2f5c4440707db7b96f53765562850e37d7b8505bb92c1532e
Status: Image is up to date for shreyasshende/wine-quality-eval:latest
docker.io/shreyasshende/wine-quality-eval:latest
ubuntu@ip-172-31-17-8:~$
```

Step 12: Run Docker Container

Run the Docker container on the desired instances.

sudo docker run shreyasshende/wine-quality-eval:latest

```
ubuntu@ip-172-31-17-8:~$ sudo docker run shreyasshende/wine-quality-eval:latest
spark 19:35:50.89 INFO  ==>
spark 19:35:50.90 INFO  ==> Welcome to the Bitnami spark container
spark 19:35:50.90 INFO  ==> Subscribe to project updates by watching https://github.com/bitnami/containers
spark 19:35:50.90 INFO  ==> Submit issues and feature requests at https://github.com/bitnami/containers/issues
spark 19:35:50.90 INFO  ==>

24/12/09 19:35:55 INFO SparkContext: Running Spark version 3.4.1
24/12/09 19:35:55 INFO ResourceUtils: ==============================================================
24/12/09 19:35:55 INFO ResourceUtils: No custom resources configured for spark.driver.
24/12/09 19:35:55 INFO ResourceUtils: ==============================================================
24/12/09 19:35:55 INFO SparkContext: Submitted application: Wine Quality Evaluation
24/12/09 19:35:55 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amoun
t: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/12/09 19:35:55 INFO ResourceProfile: Limiting resource is cpu
24/12/09 19:35:55 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/12/09 19:35:55 INFO SecurityManager: Changing view acls to: spark
24/12/09 19:35:55 INFO SecurityManager: Changing modify acls to: spark
24/12/09 19:35:55 INFO SecurityManager: Changing view acls groups to:
24/12/09 19:35:55 INFO SecurityManager: Changing modify acls groups to:
24/12/09 19:35:55 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: spark; groups with view permissions: EMPTY; users
with modify permissions: spark; groups with modify permissions: EMPTY
24/12/09 19:35:55 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
24/12/09 19:35:56 INFO Utils: Successfully started service 'sparkDriver' on port 38663.
24/12/09 19:35:56 INFO SparkEnv: Registering MapOutputTracker
24/12/09 19:35:56 INFO SparkEnv: Registering BlockManagerMaster
24/12/09 19:35:56 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/12/09 19:35:56 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/12/09 19:35:56 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
24/12/09 19:35:56 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-8286c006-a2e8-460b-87c0-91bf6d8a1f3c
24/12/09 19:35:56 INFO MemoryStore: MemoryStore started with capacity 434.4 MiB
24/12/09 19:35:56 INFO SparkEnv: Registering OutputCommitCoordinator
24/12/09 19:35:56 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
24/12/09 19:35:57 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/12/09 19:35:57 INFO SparkContext: Added JAR file:/app/WineQualityEval/target/wine-quality-eval-1.0-SNAPSHOT.jar at spark://356087a3566e:38663/jars/wine-quality-eval-1.0-SNAPSHO
T.jar with timestamp 1733772955543
24/12/09 19:35:57 INFO Executor: Starting executor ID driver on host 356087a3566e
```

```
F1 Score: 0.7634353741496599
24/12/09 19:36:11 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/12/09 19:36:11 INFO SparkUI: Stopped Spark web UI at http://356087a3566e:4040
24/12/09 19:36:11 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/12/09 19:36:11 INFO MemoryStore: MemoryStore cleared
24/12/09 19:36:11 INFO BlockManager: BlockManager stopped
24/12/09 19:36:11 INFO BlockManagerMaster: BlockManagerMaster stopped
24/12/09 19:36:11 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/12/09 19:36:11 INFO SparkContext: Successfully stopped SparkContext
24/12/09 19:36:11 INFO ShutdownHookManager: Shutdown hook called
24/12/09 19:36:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-120fa333-467f-4183-a998-b81084308366
24/12/09 19:36:11 INFO ShutdownHookManager: Deleting directory /tmp/spark-f348754e-125a-4ea6-a5fd-f3c719ca26d7
ubuntu@ip-172-31-17-8:~$ cd spark-3.4.1/sbin
```

You can see the result in the red box.

Model used: SVC (Support Vector Classifier)

Result:

F1 Score: 0.7634353741496599 (On Validation Dataset)