**Link to the GitHub repository :-**

<https://github.com/RutujaTikare13/winequality>

**Link to the DockerHub :-**

<https://hub.docker.com/repository/docker/rutujatkr/winequality>

**--------------------------------------------------------------------------------------------------------------------**

**Steps to run the python program for model training:-**

1. Create an ec2 instance to run the flintrock server.

2. SSH into the flintrock ec2 instance using the below command

**ssh -i <path to pem file> ec2-user@<IPV4 address of the instance>**

3. Create a aws credentials file on the flintrock instance.

4. Copy the pem file from the local to the flintrock ec2 instance using the command.

**scp -i < path to pem file> ec2-user@< IPV4 address of the instance>:/home/ec2- user/pem**

5. Install flintrock on the ec2 instance using the below command.

**pip3 install flintrock**

6. Configure the flintrock using the below command.

**flintrock configure**

7. Open the config.yaml file in editor mode and set the keyname to the pem filename(without .pem) and path to the pem file path. Change the instance type to m4.large with slave instances as 3. Set install hdfs and spark, set spark to true after uncommenting.

8. Launch the flintrock cluster.

**flintrock launch <cluster\_name>**

9. Log into spark master using:

**flintrock login < cluster\_name>**

10. Create alias for python to python 3.7 and pip to pip3 in the ~/.bashrc file

11. Install the following libraries:

**sudo yum install python-devel**

**sudo yum install python-pip**

**pip install pandas**

**pip install matplotlib**

**pip install -U scikit-learn scipy matplotlib**

**sudo yum install gcc-gfortran**

12. Set the inbound and outbound rules to allow ssh/scp access to flintrock master

13. Copy the python code, training dataset and Validation dataset from the local to the master ec2 instance. Copy the CSV to the Hadoop cluster:

**hadoop fs -mkdir -p /hdfs/csv**

**hadoop fs -put ./ValidationDataset.csv /hdfs/csv**

**hadoop fs -put ./TrainingDataset.csv /hdfs/csv**

**hadoop fs -ls /hdfs/csv/**

14. Run the python file using the following command.

**~/spark/bin/spark-submit --master spark://<IPv4 internal IP>:7077 --deploy-mode client <python filename>** hdfs://<flintrock-master-internal-ipv4>:9000/hdfs/csv/ hdfs://<flintrock-master-internal-ipv4>:9000/hdfs/output/

15. Copy output from hdfs to flintrock-master:

hadoop fs -get /hdfs/output/ .

16. Secure copy output to local:

scp -r -i ~/wine-quality.pem ec2-user@<Flintrock Master Public Ipv4>:/home/ec2-user/Edited/output .

**--------------------------------------------------------------------------------------------------------------------**

**Steps to run the docker on single ec2 instance:-**

1. Create an ec2 instance.

2. SSH into the created instance.

3. Install docker using the below command.

**sudo amazon-linux-extras install docker**

4. Copy the ValidationDataset.csv file on to the instance.

5. Pull the docker file using the below command.

**sudo** **docker pull rutujatkr/winequality**

6. If you encounter problem that Cannot connect to the Docker daemon then

**sudo service docker stop**

**sudo service docker start**

7. Run the docker using the below command.

**sudo docker run -v <ABSOLUTE Path toValidationDataset.csv>:/app/ValidationDataset.csv rutujatkr/winequality ValidationDataset.csv**

**--------------------------------------------------------------------------------------------------------------------**

**Steps to run the model without docker:-**

1. Create an ec2 instance.

2. SSH into the created instance.

3. Install the following libraries:

**pip3.7 install pyspark**

**pip3.7 install pandas**

**pip3.7 install matplotlib**

**pip3.7 install -U scikit-learn scipy matplotlib**

**pip3.7 install findspark**

4. Install java.

**sudo amazon-linux-extras install java-openjdk11**

5. Set JAVA\_HOME and PATH.

export JAVA\_HOME="/usr/lib/jvm/java-11-openjdk-11.0.12.0.7-0.amzn2.0.2.x86\_64/"  
 export PATH=$JAVA\_HOME/bin:$PATH

6. Copy the model, ValidationDataset.csv, python file to the instance created

6. Run the python file using the below command.

python3.7 wine\_quality\_docker.py ValidationDataset.csv