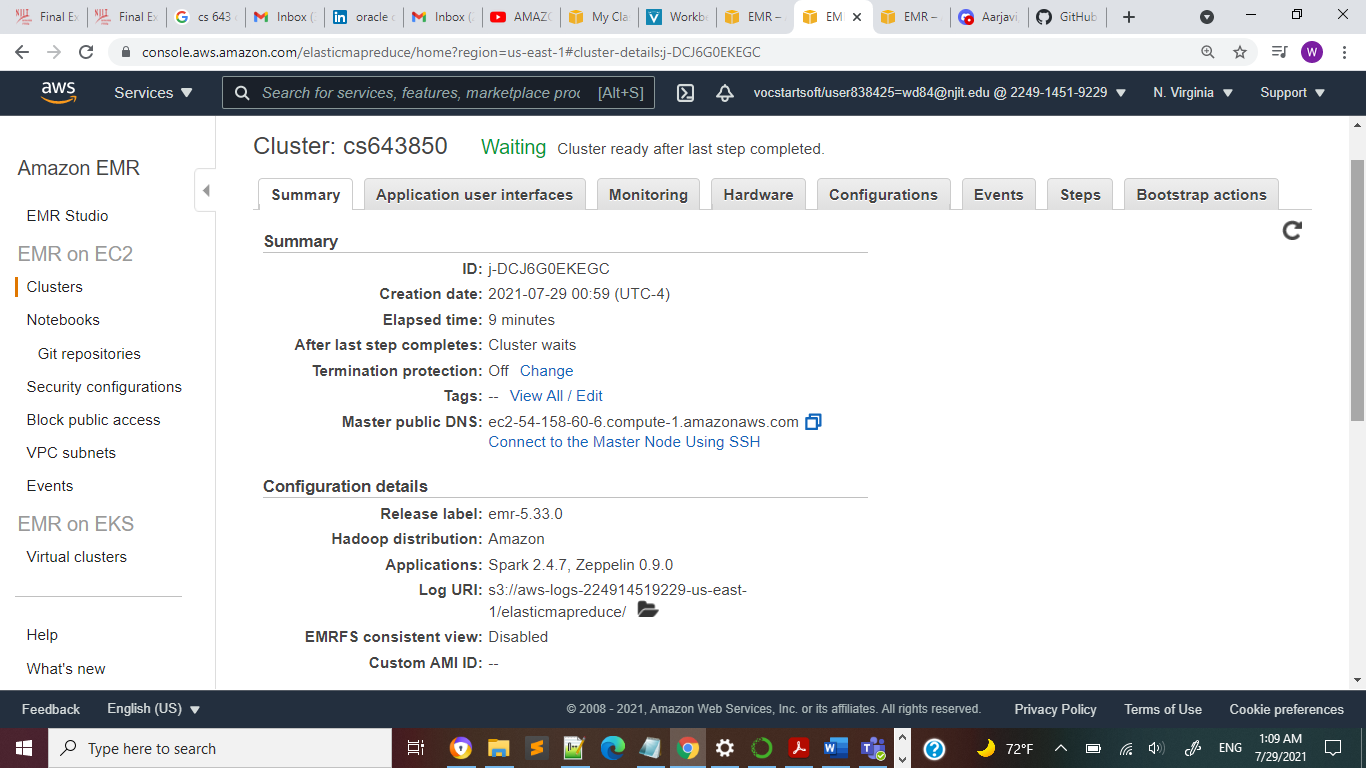
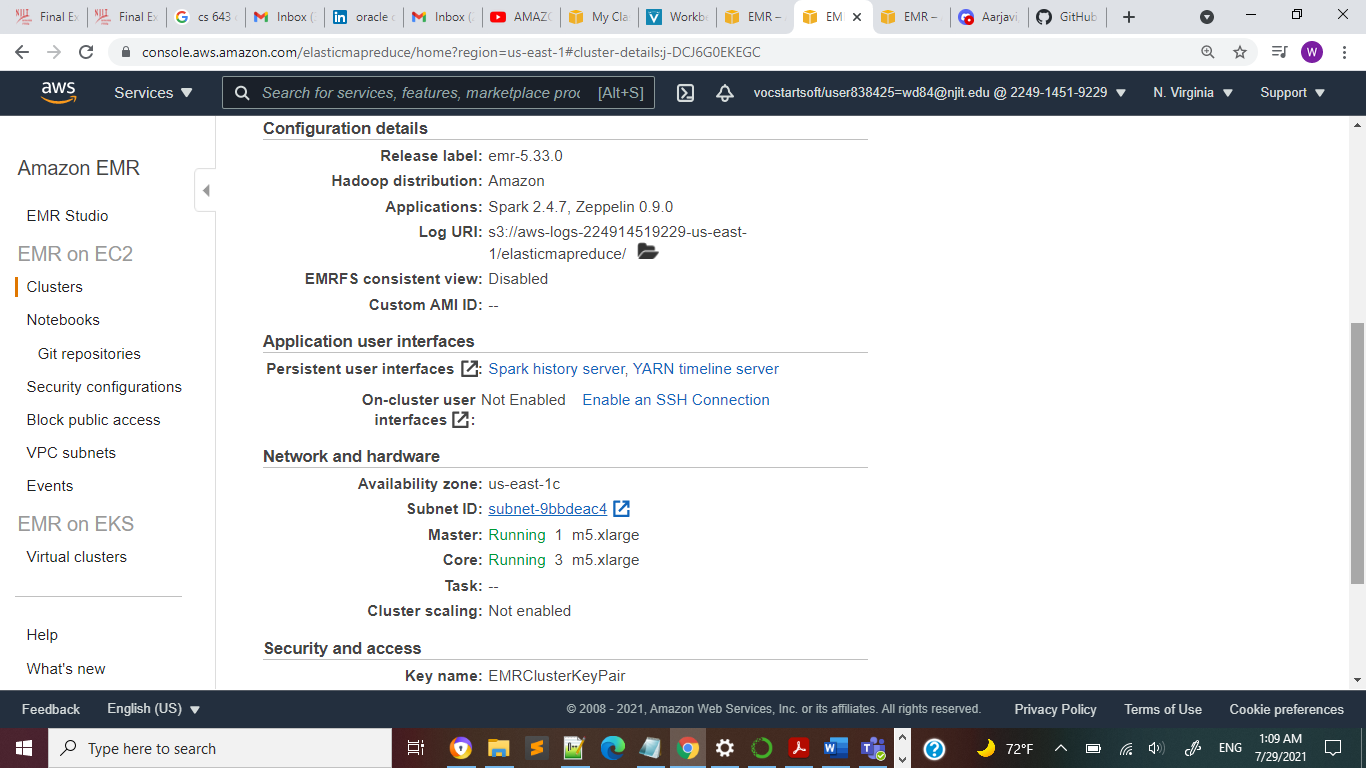
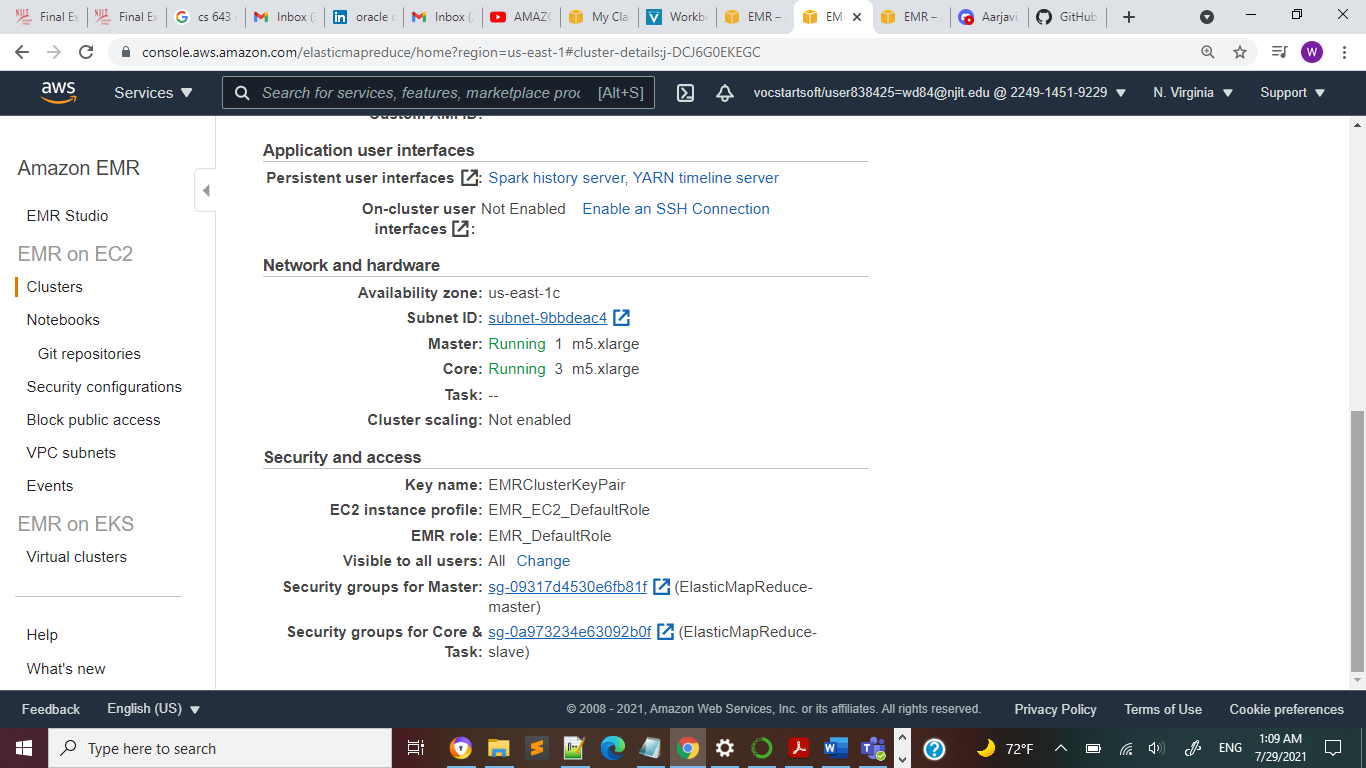
**PROGRAMING ASSIGNMENT 2**

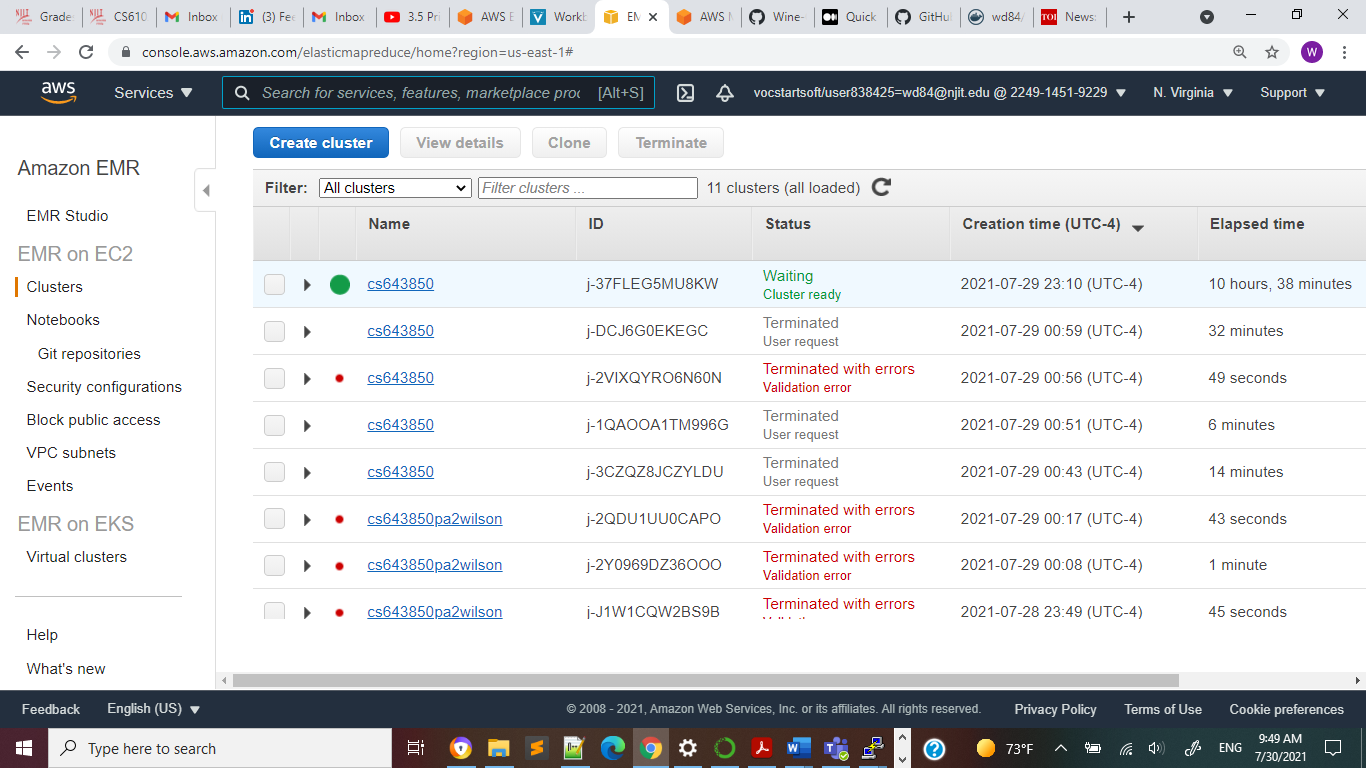
Following are the steps that were taken during the course of this Programing Assignment.

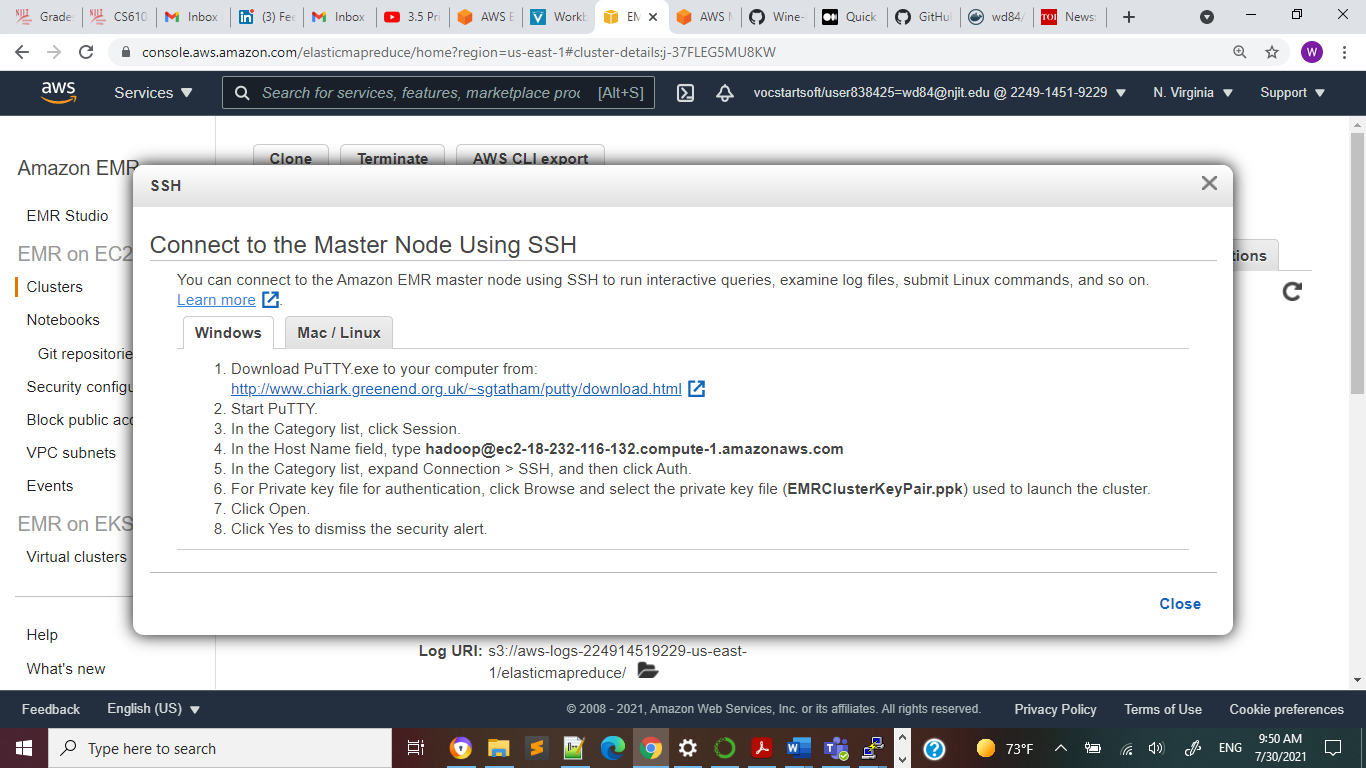
**STEP 1**: Setting up the EMR cluster

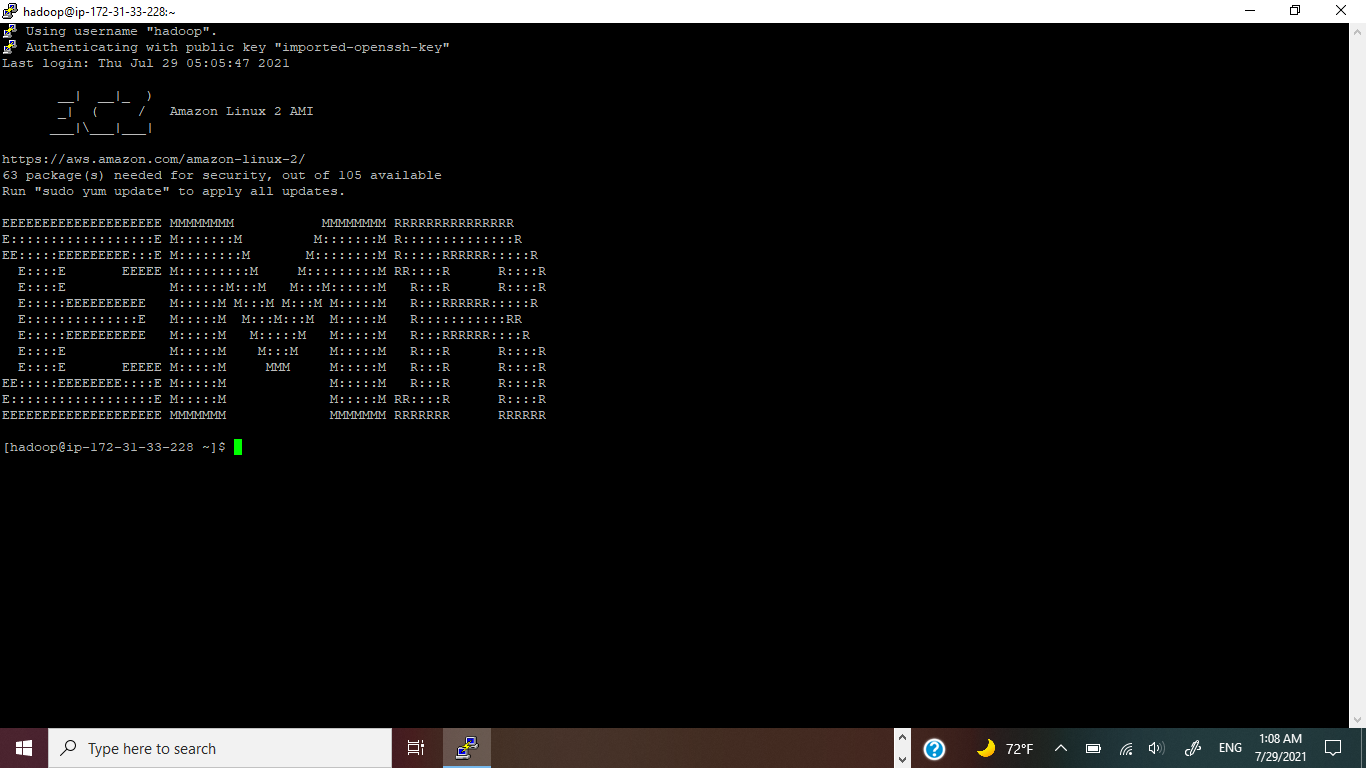












arn:aws:s3:::cs643850pa2wilson

EMR Cluster

[hadoop@ec2-18-232-116-132.compute-1.amazonaws.com](mailto:hadoop@ec2-18-232-116-132.compute-1.amazonaws.com)

After you have set up the EMR cluster following are the steps that were done

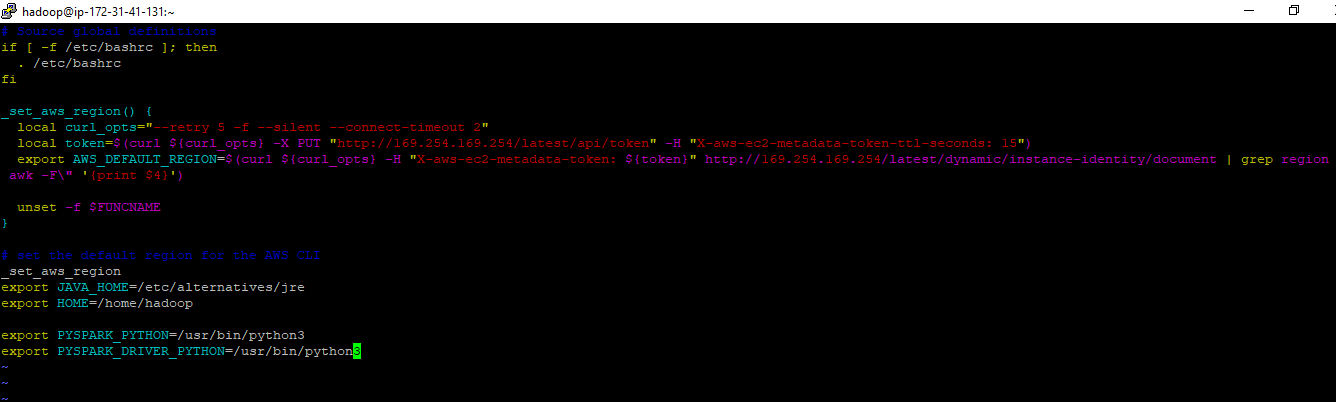
1. Create an EMR Spark cluster in AWS with 4 EC2-instances
2. Login into the master instance and do the following to Install pyspark using

***pip install pyspark***

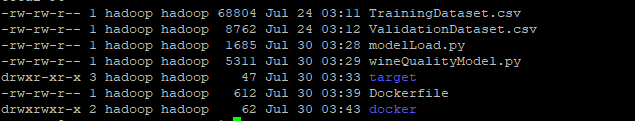
1. Set environment variables

***export PYSPARK\_PYTHON=/usr/bin/python3***

***export PYSPARK\_DRIVER\_PYTHON=/usr/bin/python3***

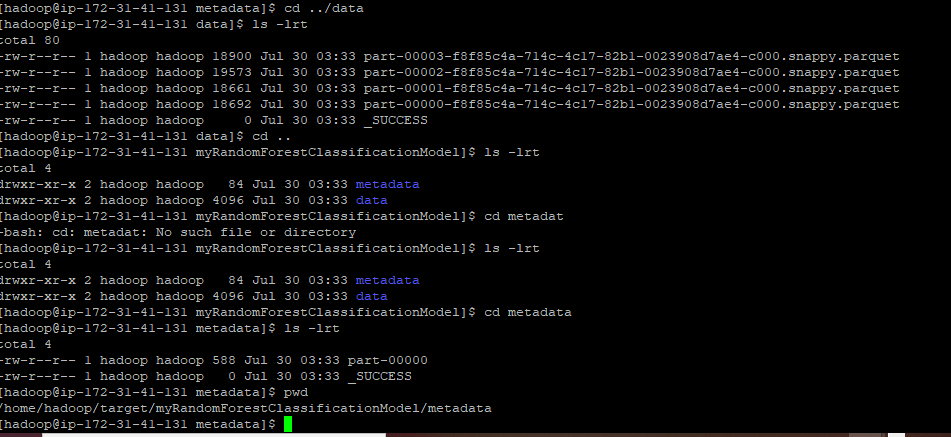


1. Now, copy wineQualityModel.py, TrainingDataset.csv and ValidationDataset.csv to the master instance and execute the following command to generate the model,



python3 wineQualityModel.py

1. The above command will generate *target* folder which has the source folder to store the model.



1. Later, install Docker in EMR instance using the below commands,

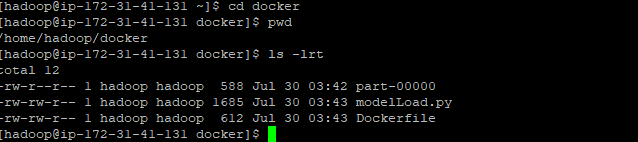
***sudo yum update -y***

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

***sudo yum install docker***

***sudo service docker start sudo usermod -a -G docker Hadoop***

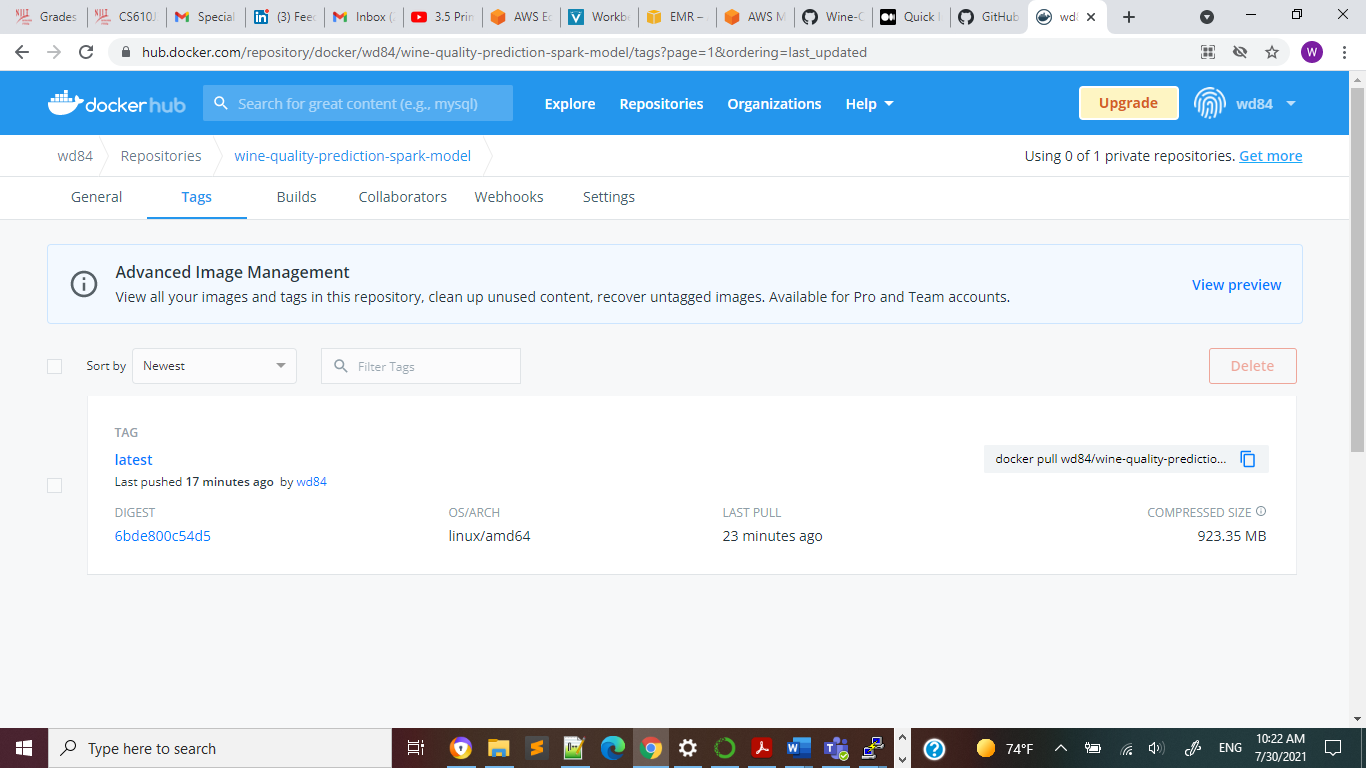
1. Now, we need to create the image. So, copy the Dockerfile, target (model) and modelLoad.py into a new directory in master instance and execute the following command to build the image,



Command to build the image is as below

***sudo docker build -t winemodel***

1. Now we need to create docker hub account and repository

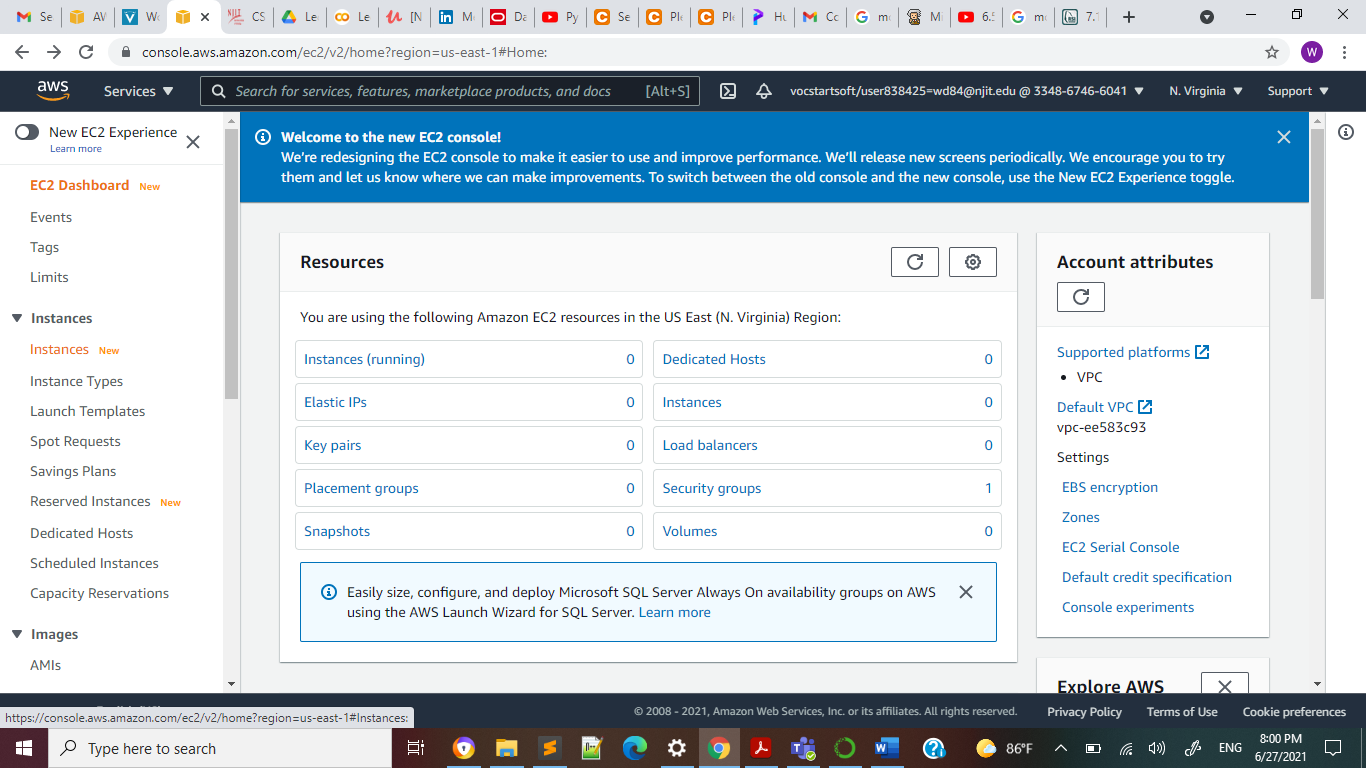


1. Now, we need to tag and push this image to DockerHub using the following command. Wd84 is my docket account ID and wine-quality-prediction-spark-model is the repository inside the account

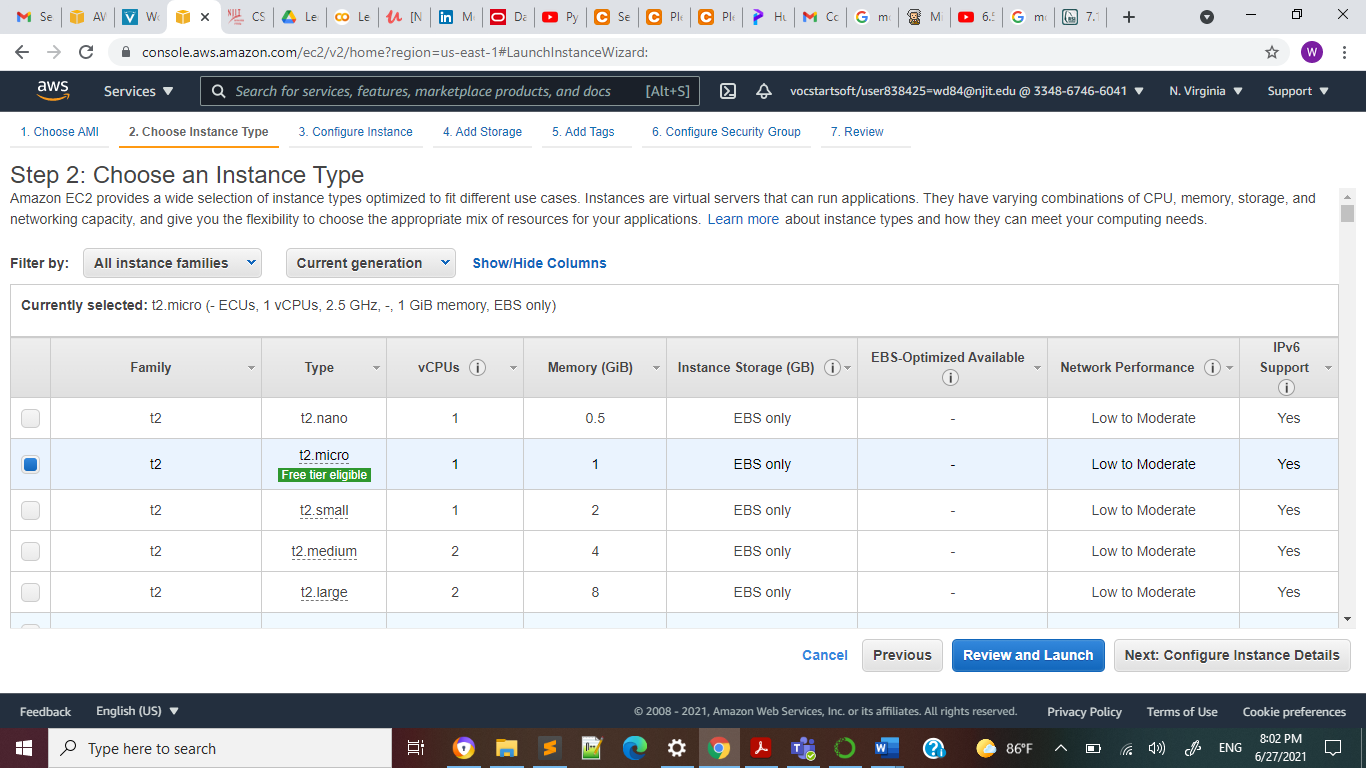
docker tag winemodel wd84/wine-quality-prediction-spark-model

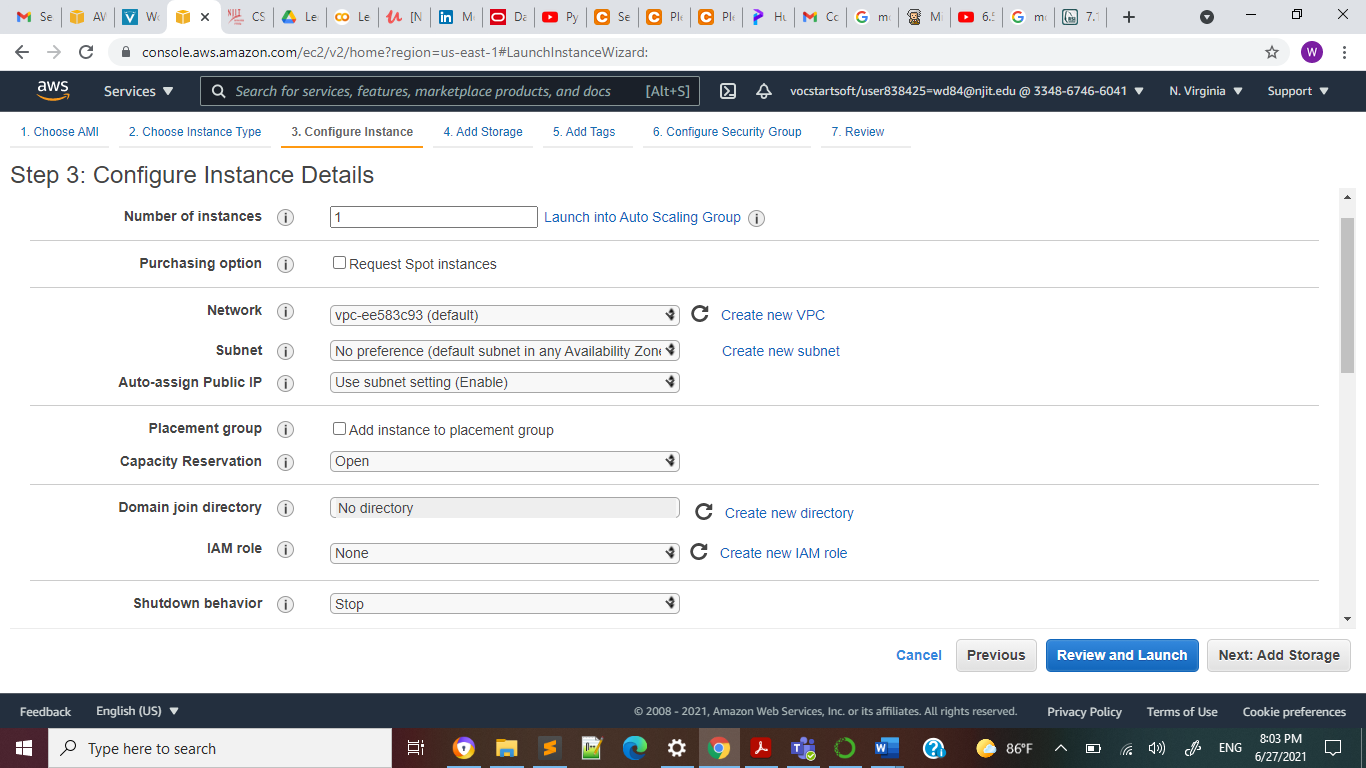
docker push wd84/wine-quality-prediction-spark-model:latest

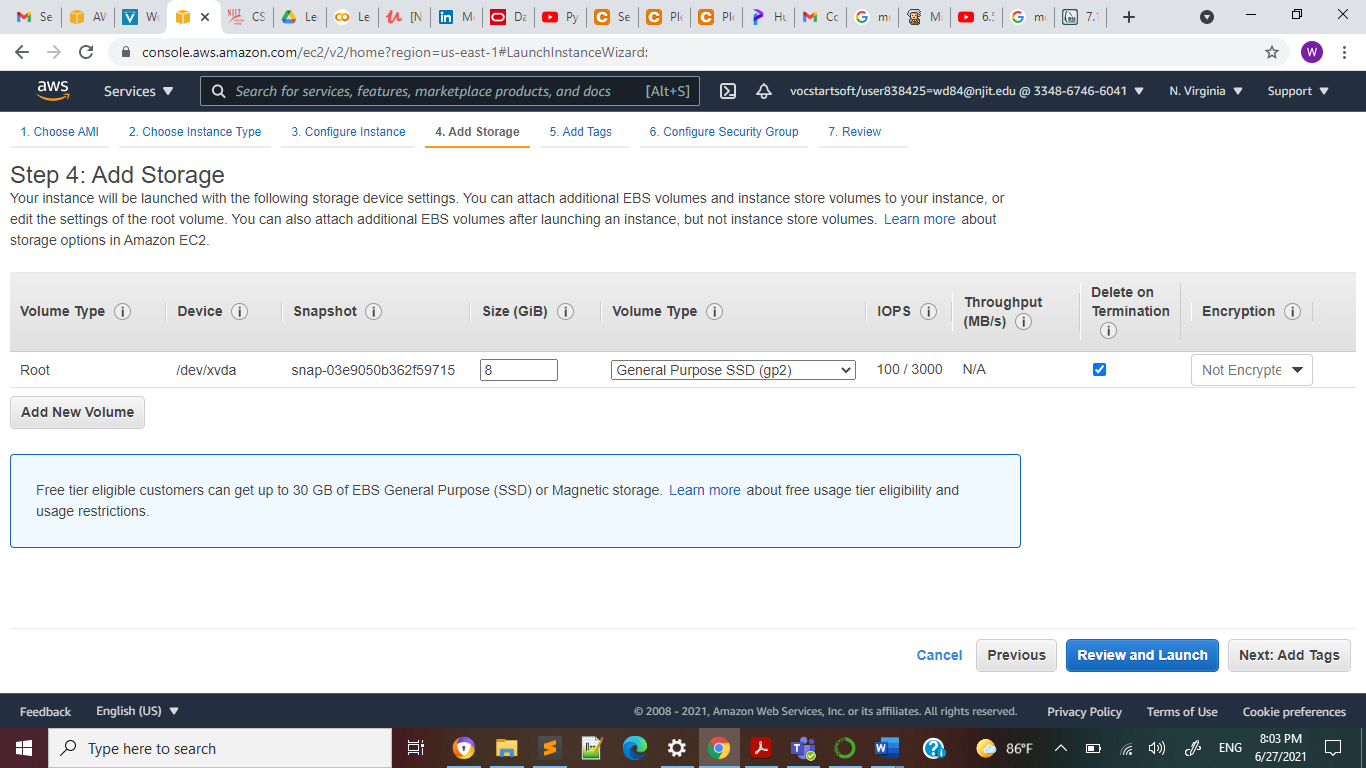
1. Now, create a new ec2 instance and login to that new ec2 instance as **ec2-user**

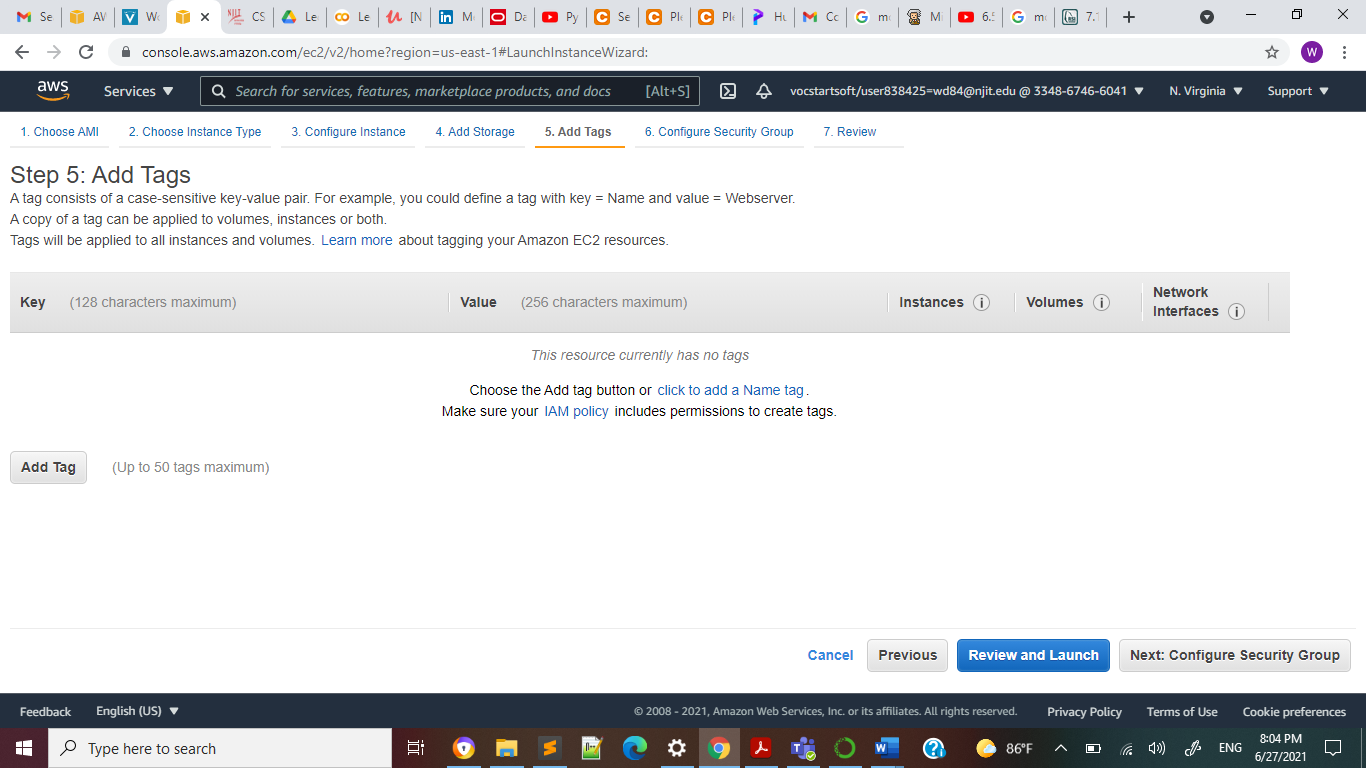


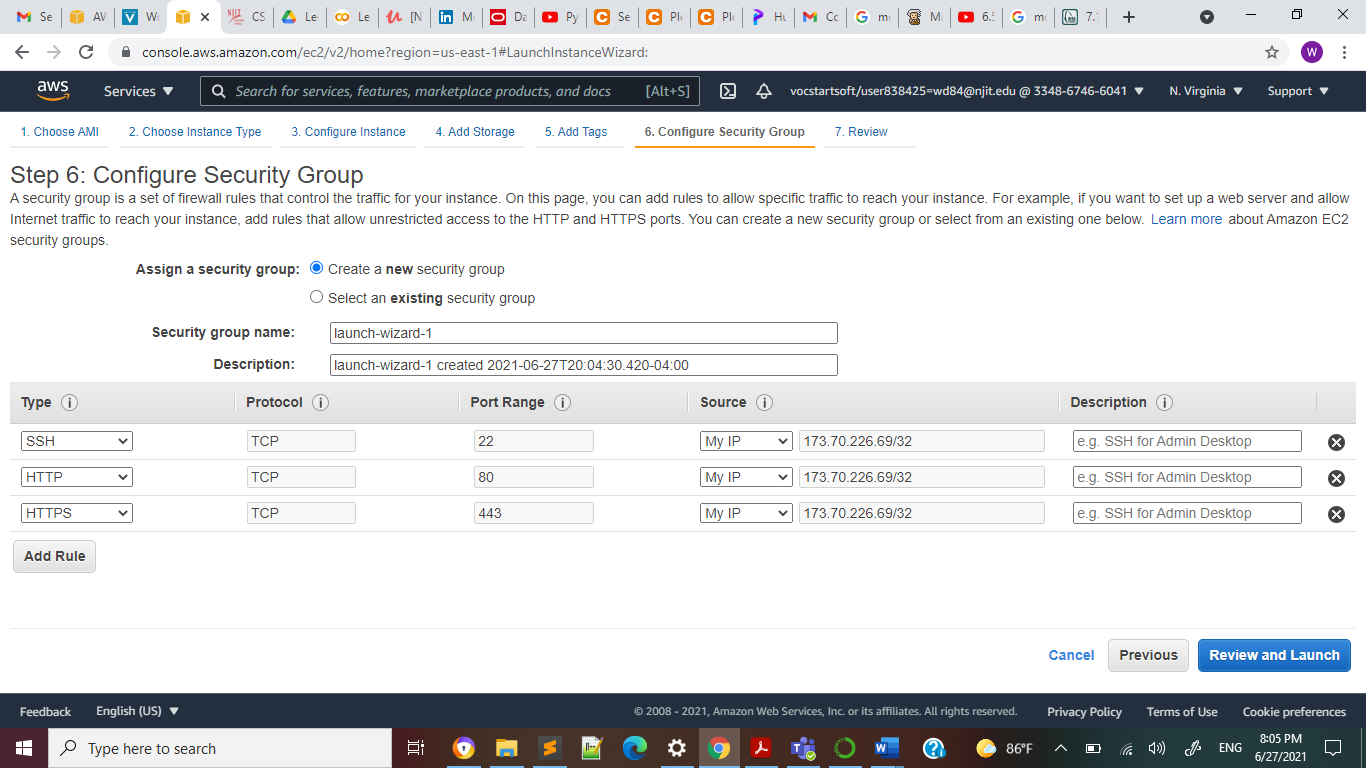


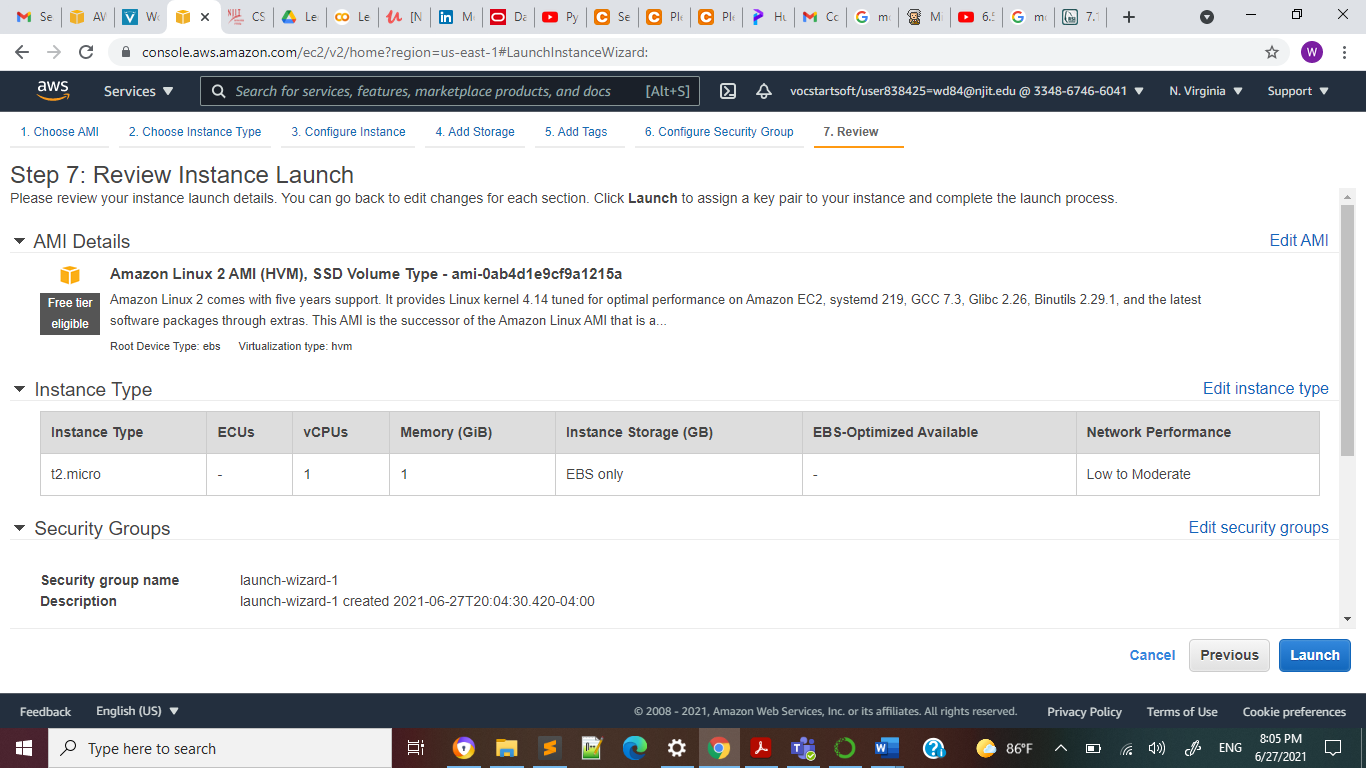


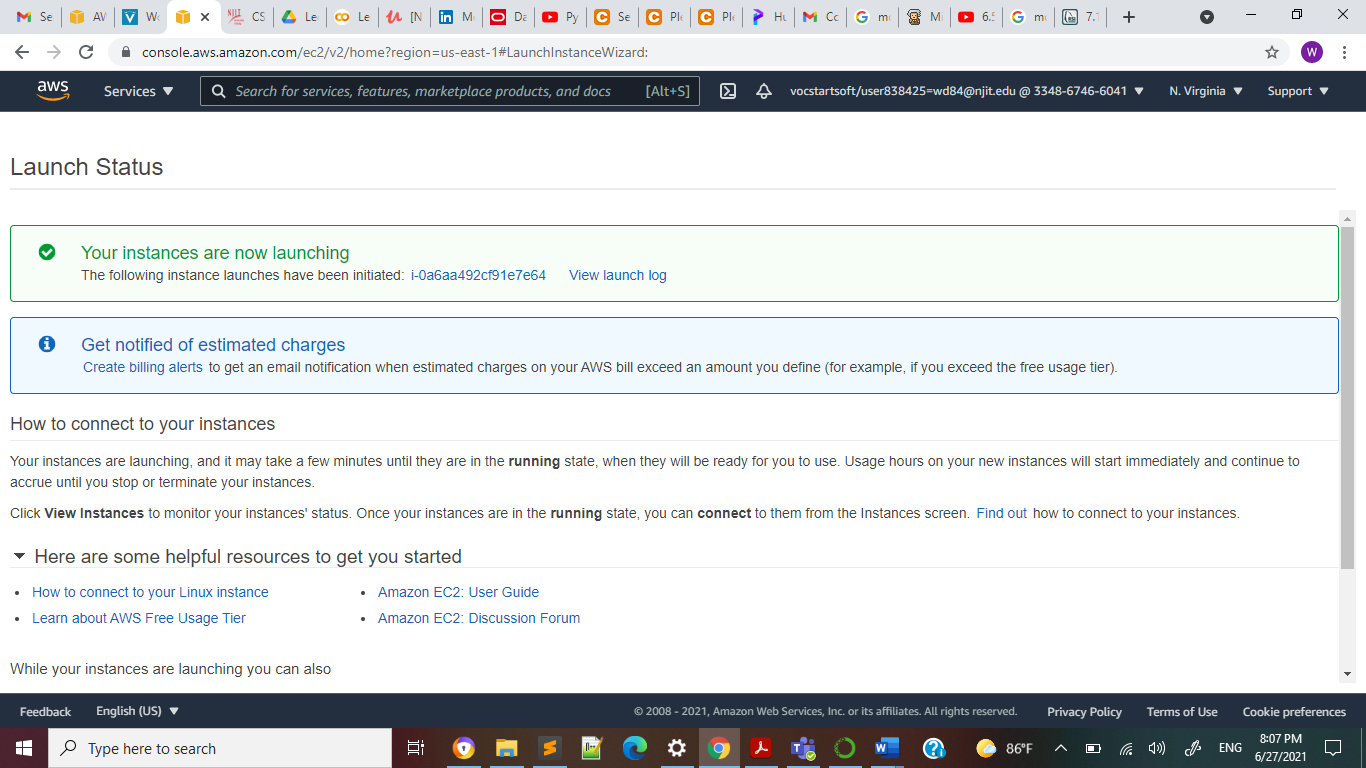


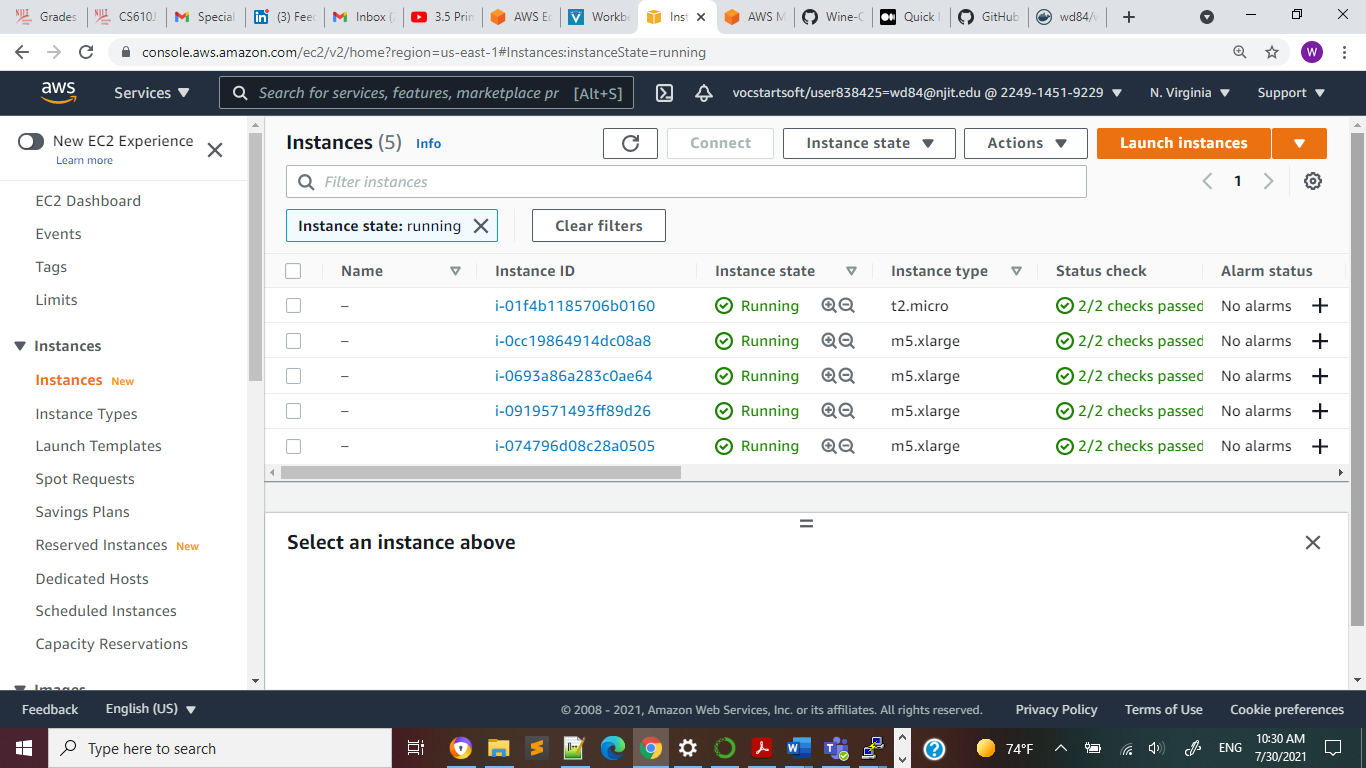
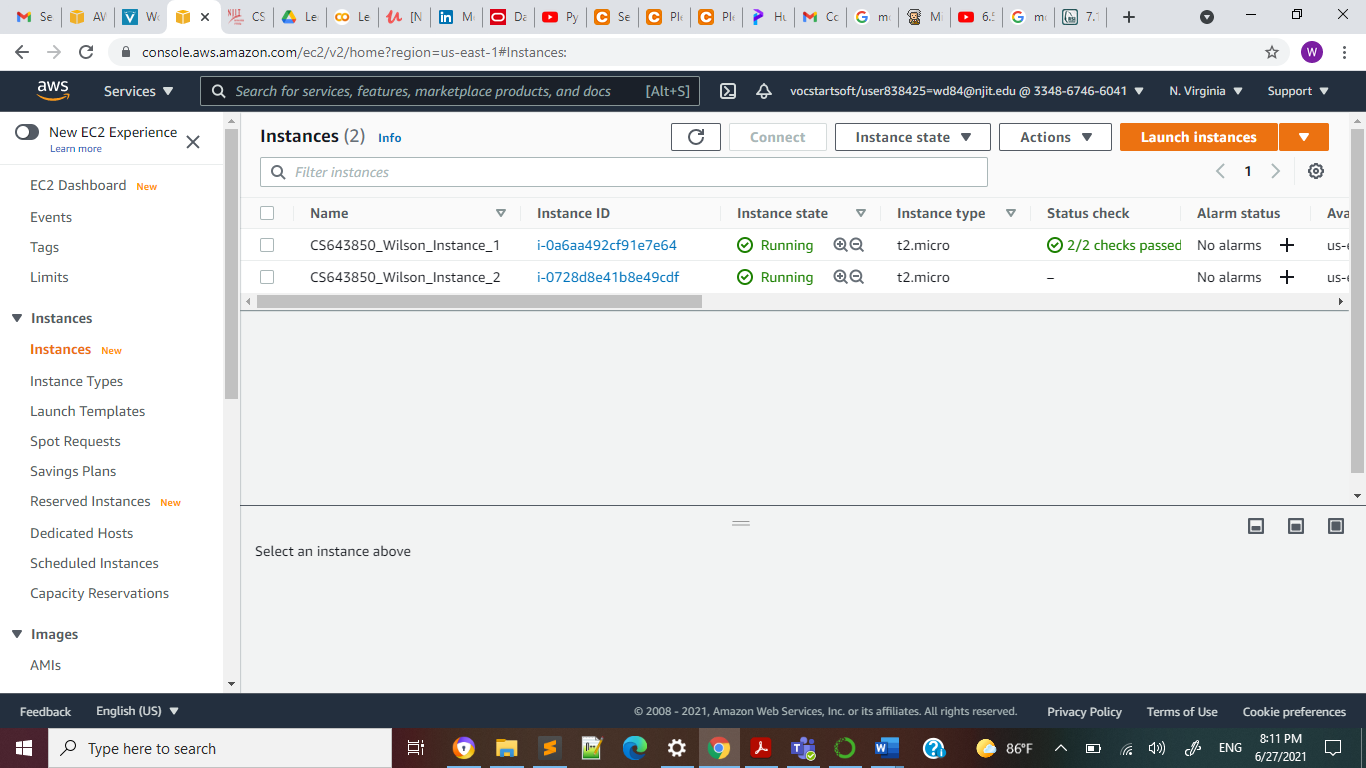




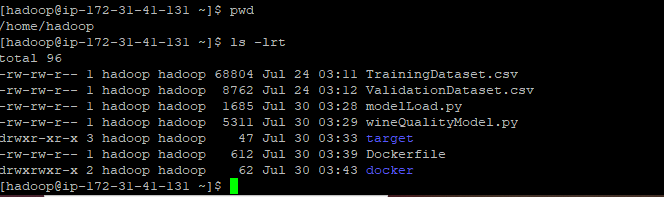




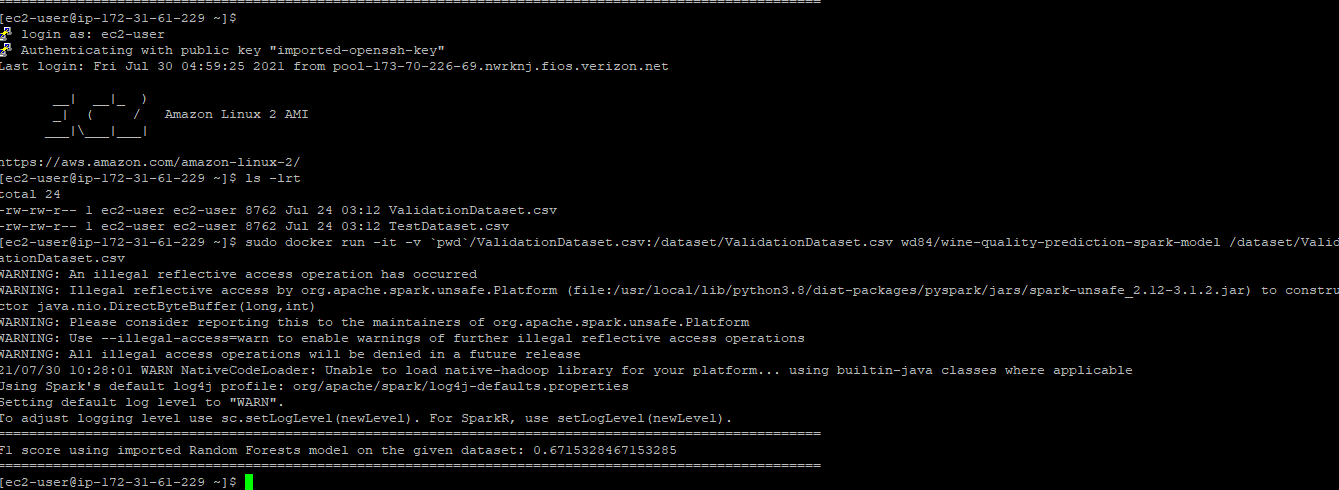


1. Copy the ValidationDataset.csv into your present working directory in ec2 instance and execute the following command



***sudo docker run -it -v `pwd`/ValidationDataset.csv:/dataset/ValidationDataset.csv wd84/wine-quality-prediction-spark-model /dataset/ValidationDataset.csv***



**Summary Statistics**

**To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).**

**============================================================================**

**F1 score using imported Random Forests model on the given dataset: 0.6715328467153285**

**NOTE: To run the docker image with your individual file please put your test file in the directory and then run the below command. Replace ValidationDataset.csv (highlighted in yellow) with the file name that you want to test**

***sudo docker run -it -v `pwd`/ValidationDataset.csv:/dataset/ValidationDataset.csv wd84/wine-quality-prediction-spark-model /dataset/ValidationDataset.csv***