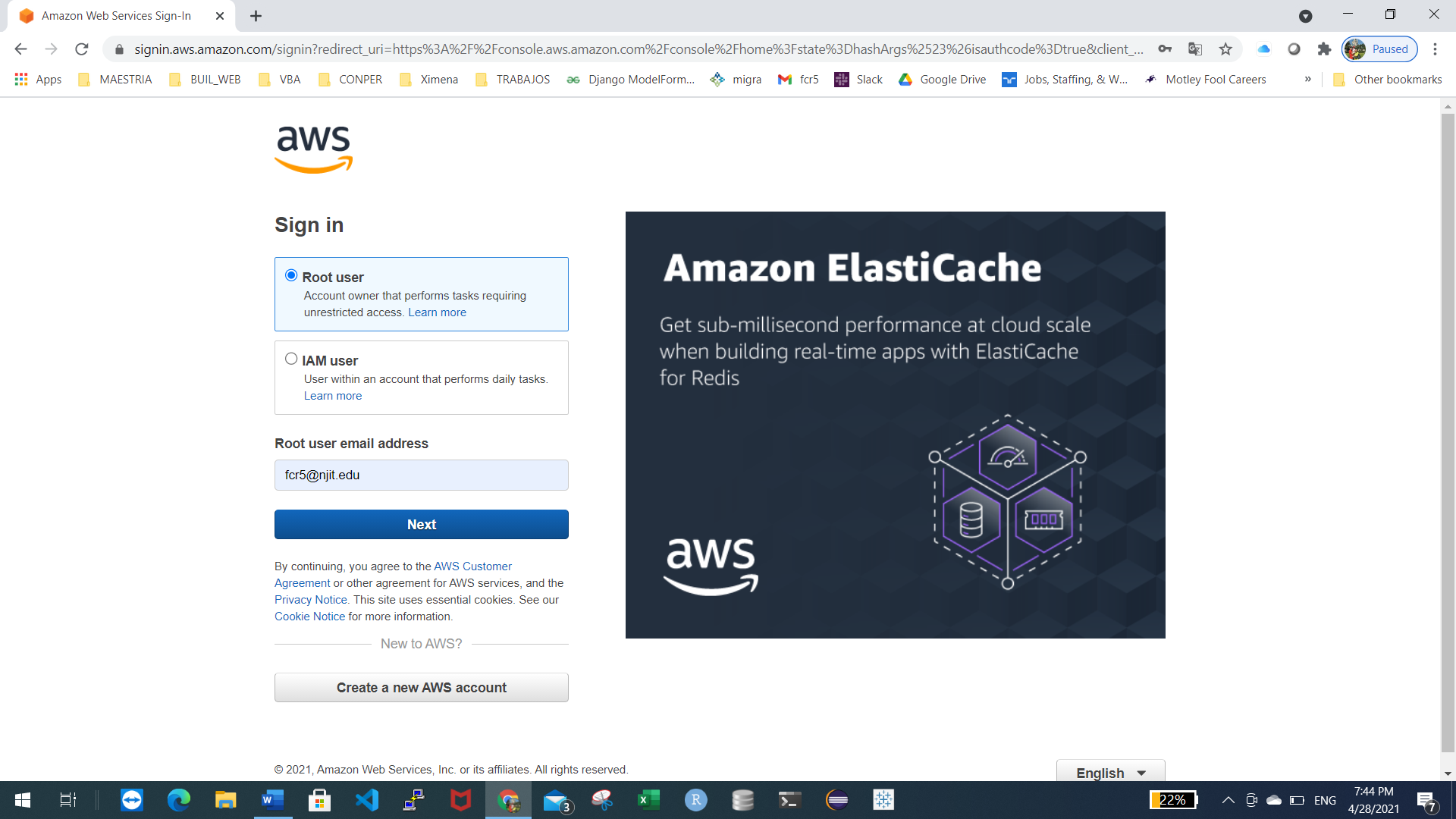
PROJECT: WINE QUALITY PREDICTION WITH SPARK

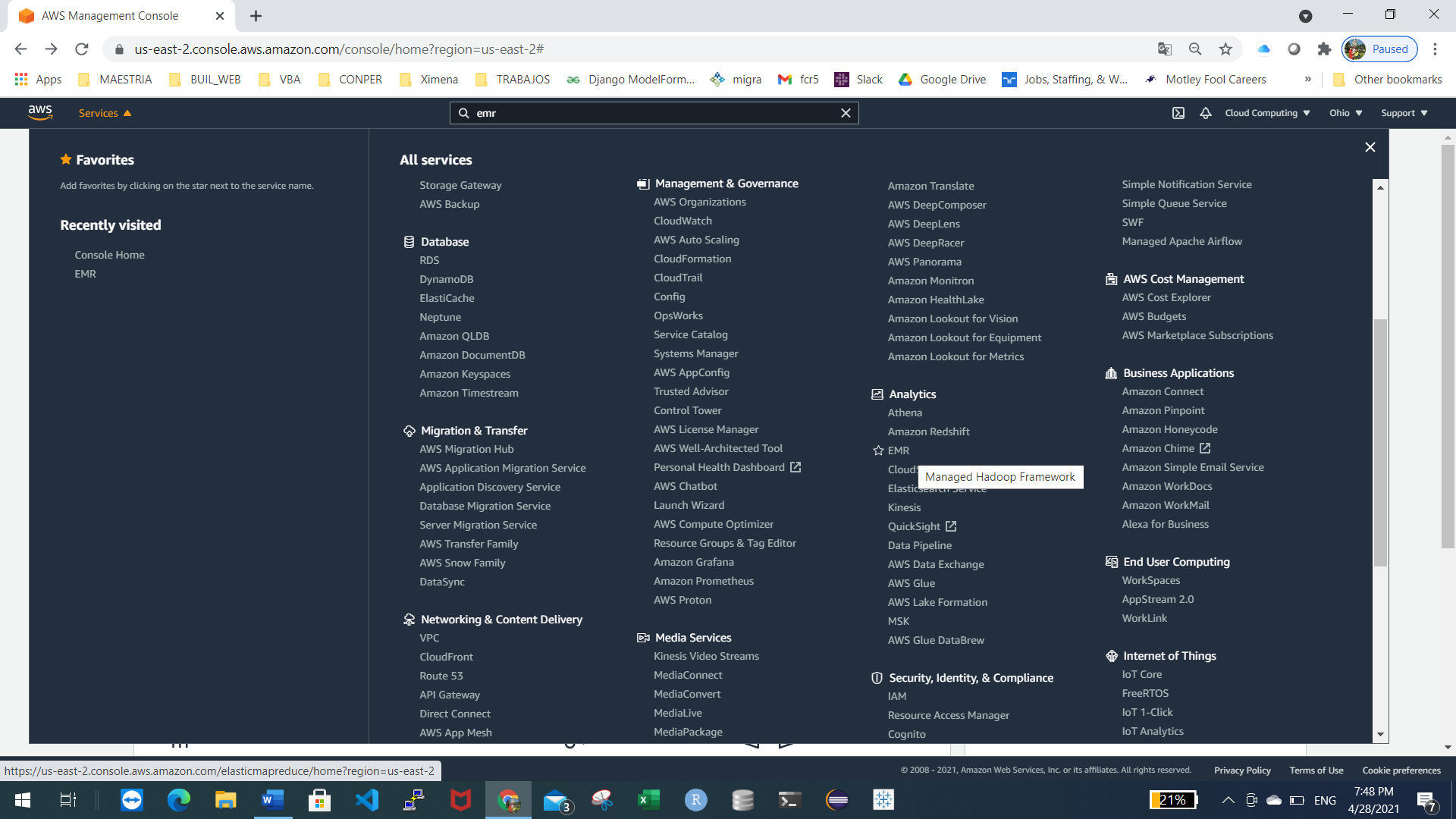
HOW TO SET UP THE CLOUD ENVIRONMENT STEP BY STEP

1. To enter to aws.com and sign in.

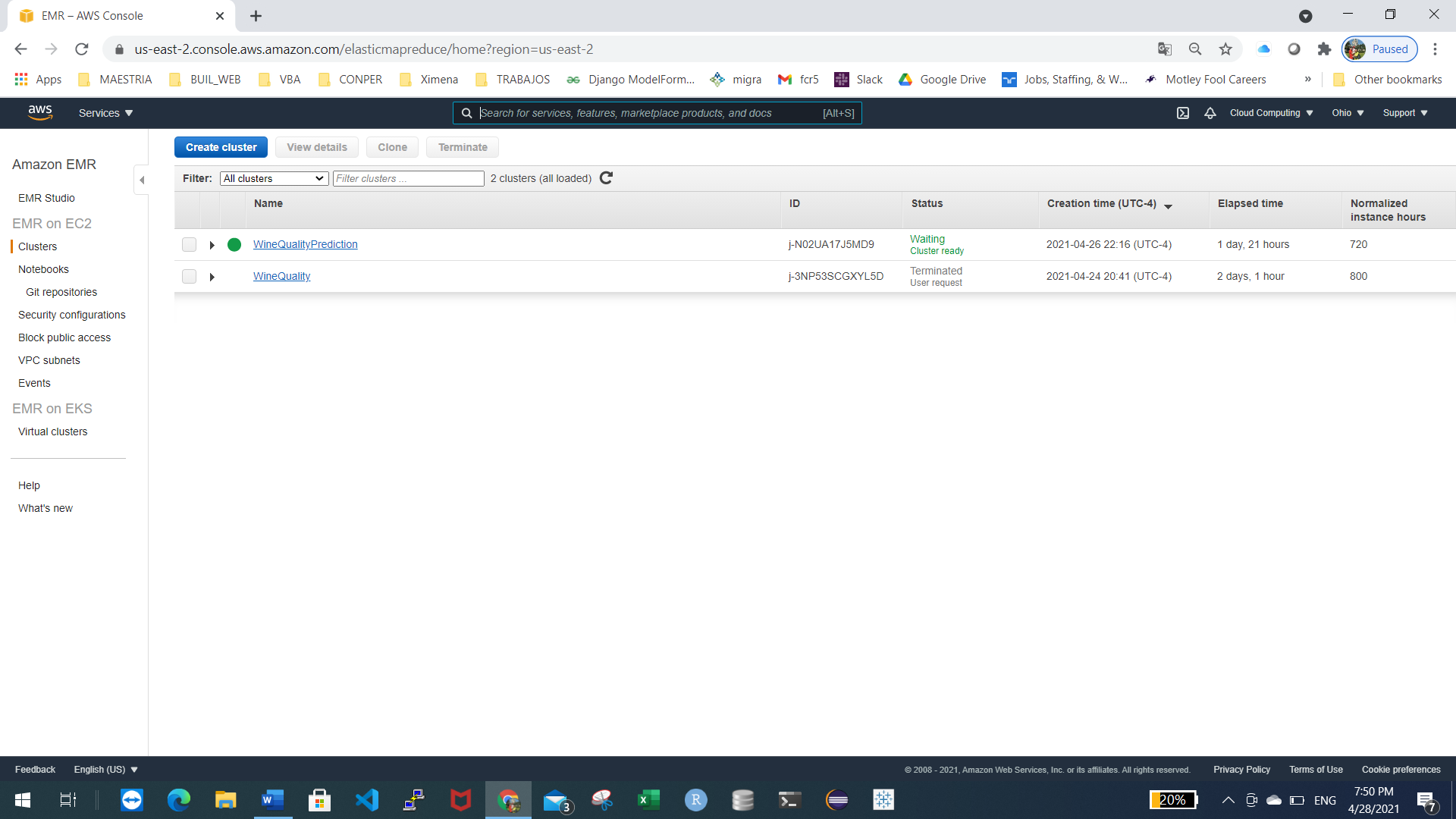


1. Create cluster with Spark installed using Quick Options

* We choose the service EMR (Elastic Map Reduce)



Choose create cluster to open the Quick Options wizard.



On the Create Cluster - Quick Options page, note the default values for Release, Instance type, Number of instances, and Permissions. These fields autopopulate with values chosen for general-purpose clusters. We must enter our values, for example:

Cluster name: WineQualityPrediction

Launch mode: cluster

Release: emr-5.33.0

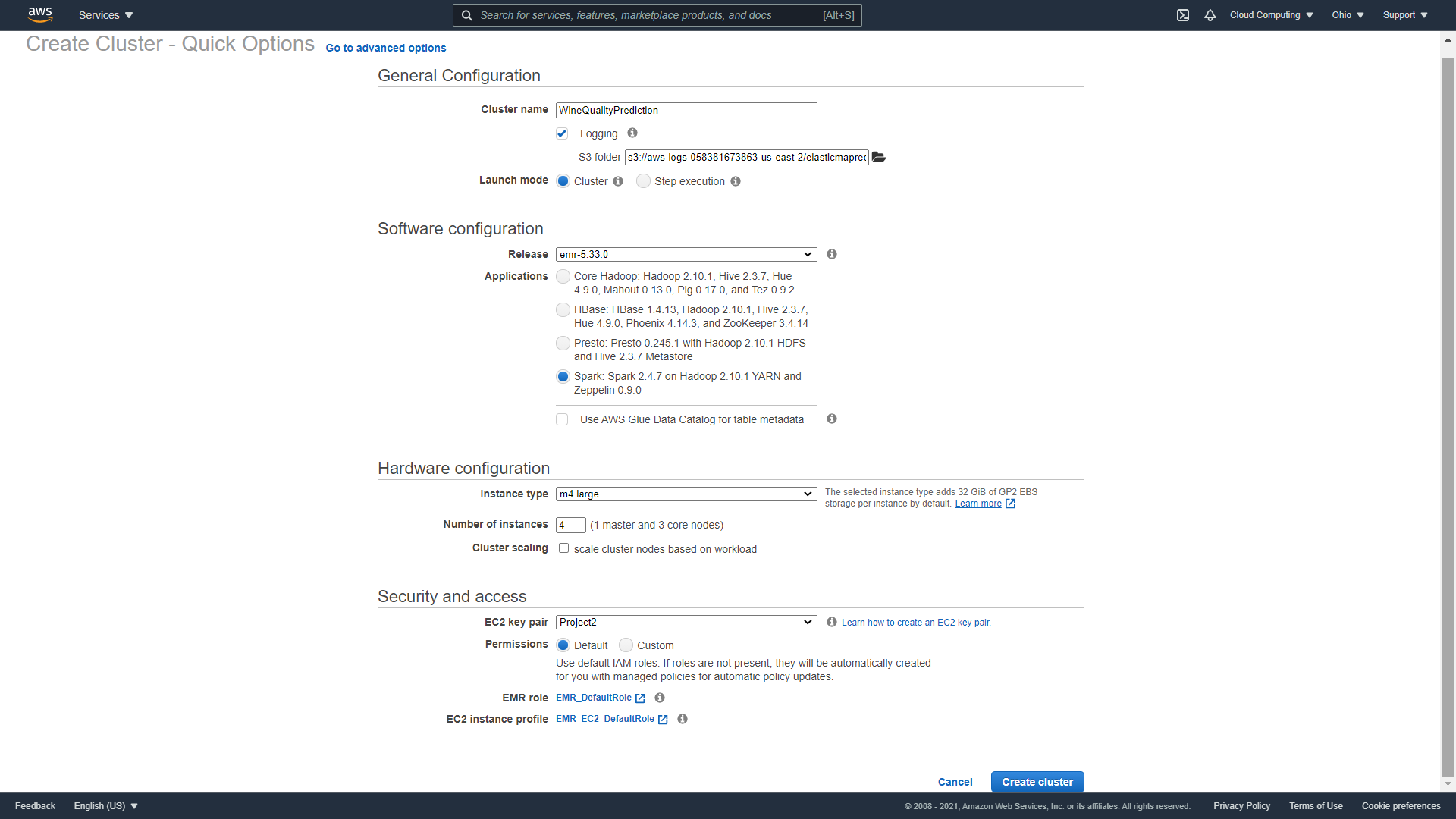
Applications: Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.9.0

Instance type: m4.large

Number of instances: 4 (1 master and 3 core nodes)

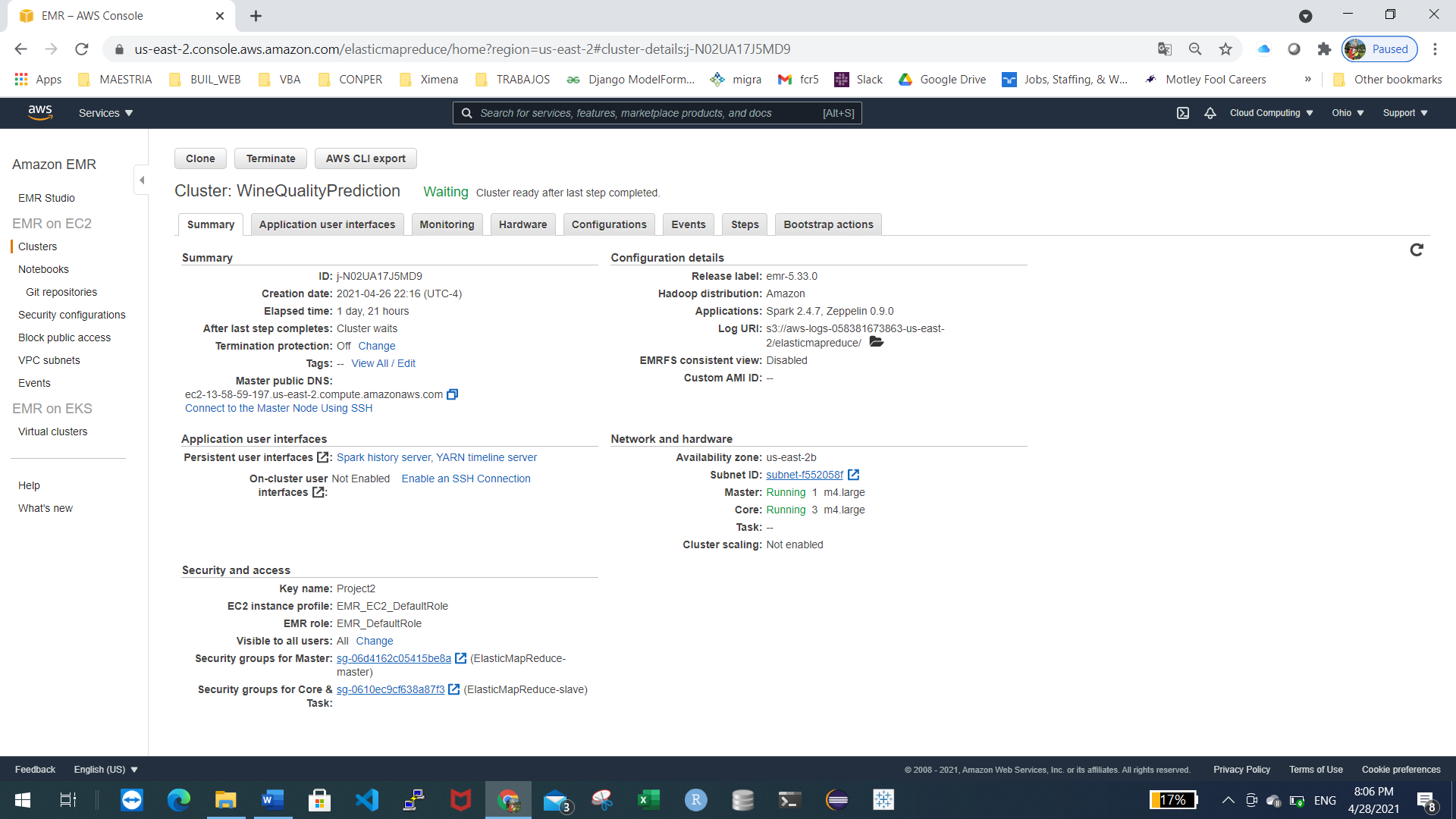
EC2 key pair: choose your key that you designated or created in Create an Amazon EC2 Key Pair for SSH. (Project2.ppk)

Click on create cluster button to launch the cluster and open the cluster status page.



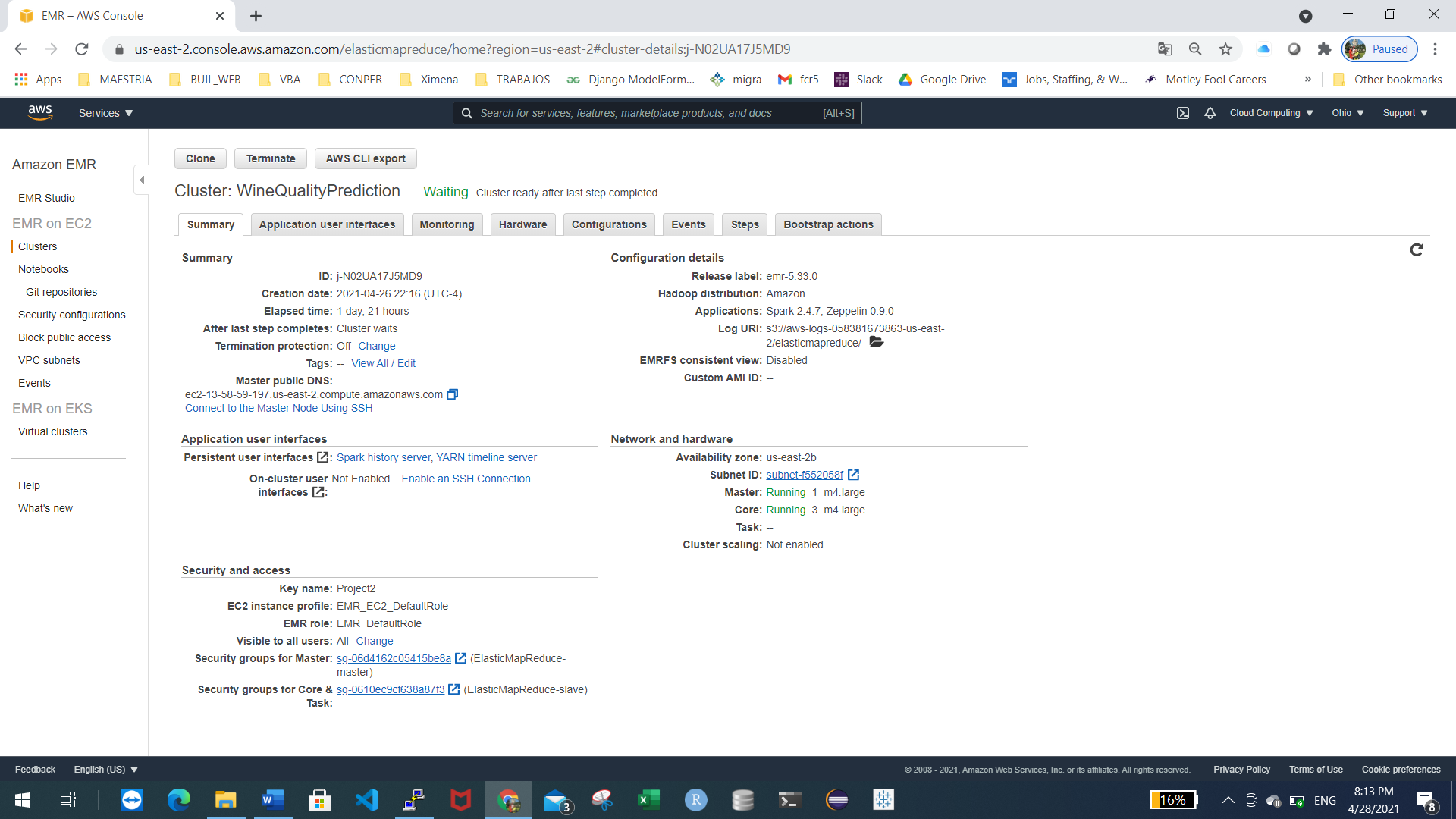
On the cluster status page, find the Status next to the cluster name. The status should change from Starting to Running to Waiting during the cluster creation process. You may need to choose the refresh icon on the right or refresh your browser to receive updates.

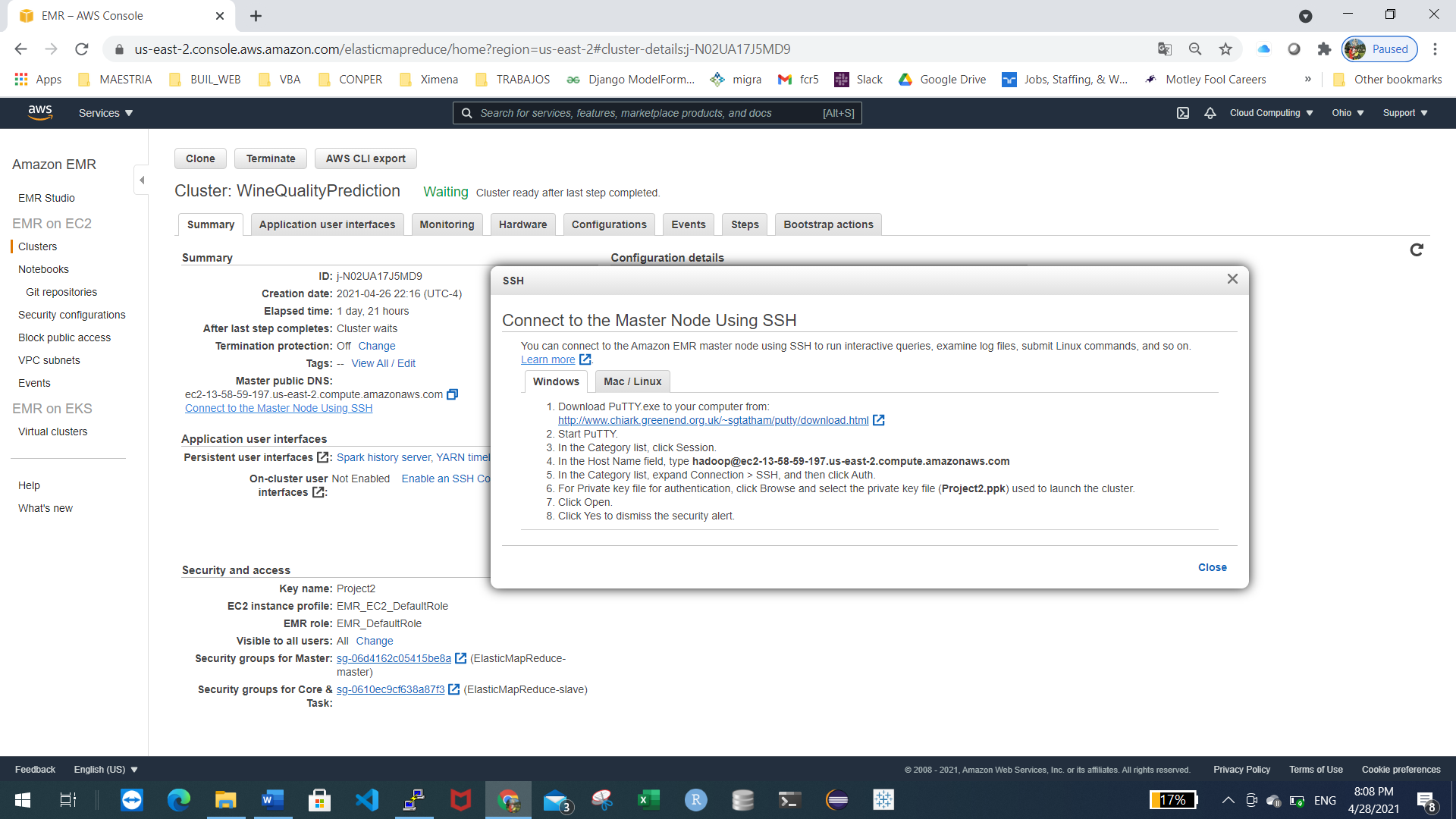
When the status progresses to Waiting, your cluster is up, running, and ready to accept work.



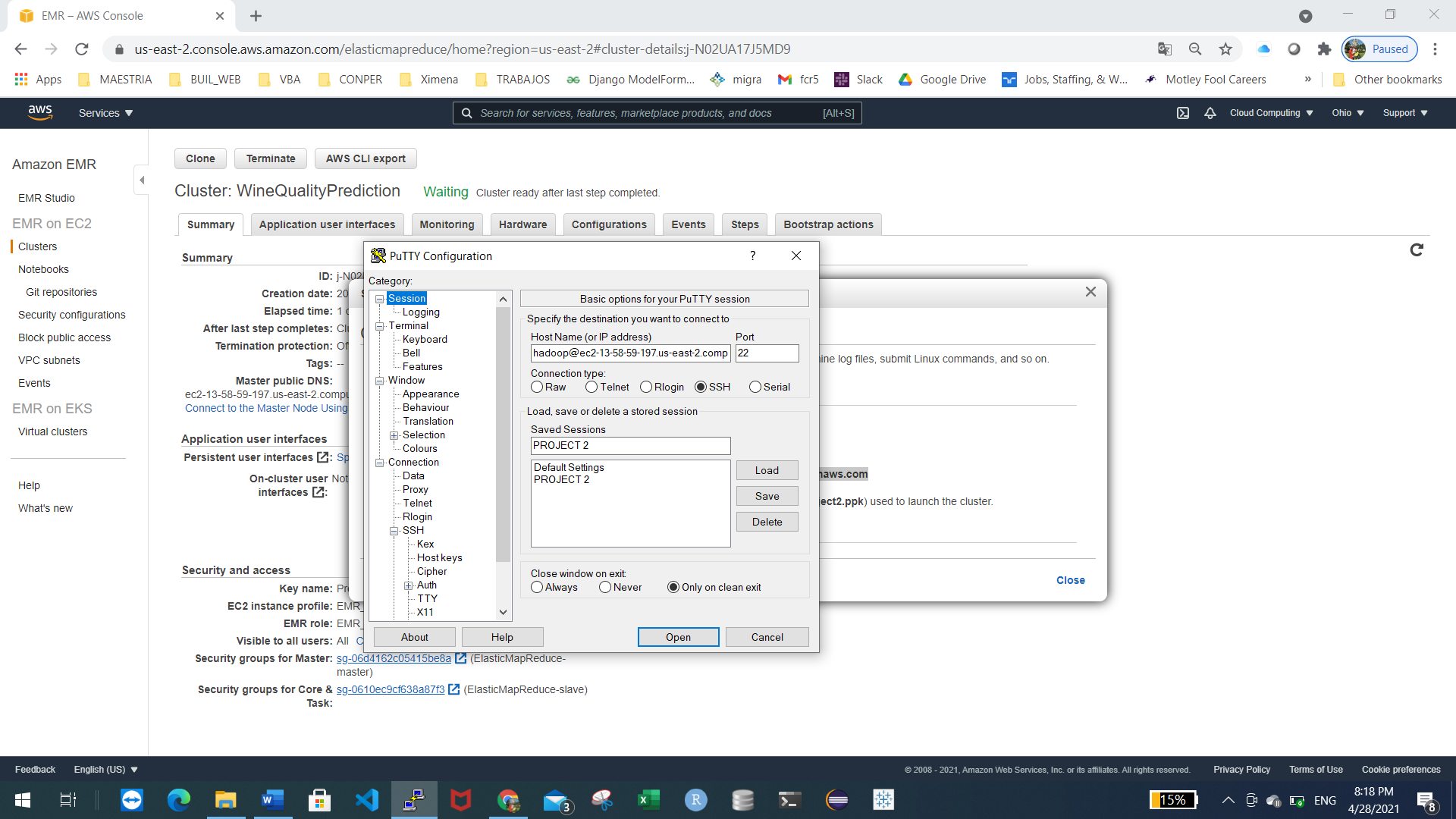
HOW TO RUN THE TRAINING MODEL

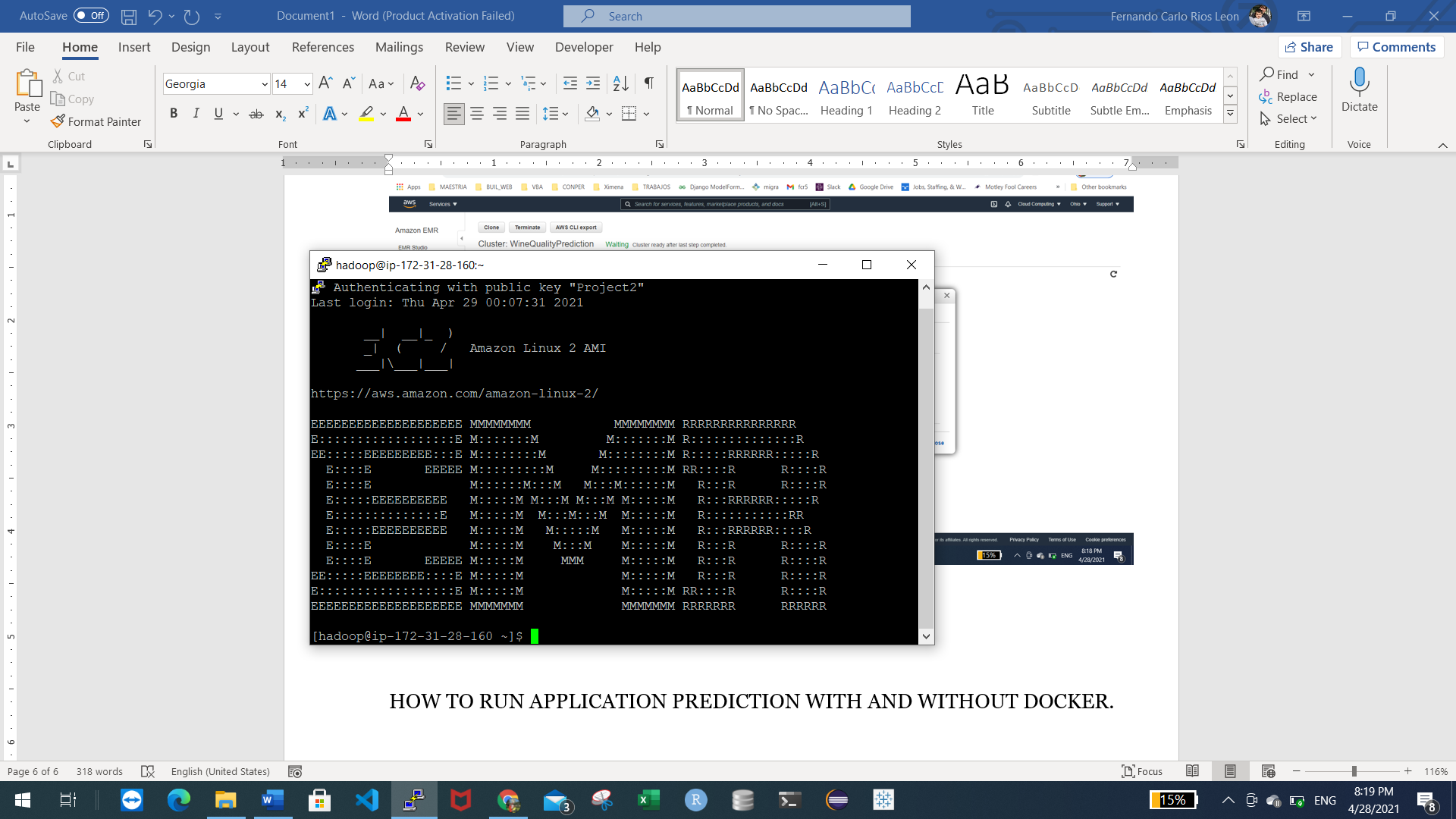
1. We must connect to the master node, I am using SSH with Putty from my desktop on windows.
   * Download PuTTY.exe to your computer from:  
     <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>
   * Start PuTTY.
   * In the Category list, click Session.
   * In the Host Name field, type your host-name (see the picture)
   * In the Category list, expand Connection > SSH, and then click Auth.
   * For Private key file for authentication, click Browse and select the private key file (Project2.ppk) used to launch the cluster.
   * Click Open.
   * Click Yes to dismiss the security alert.





Connect to master node using Putty



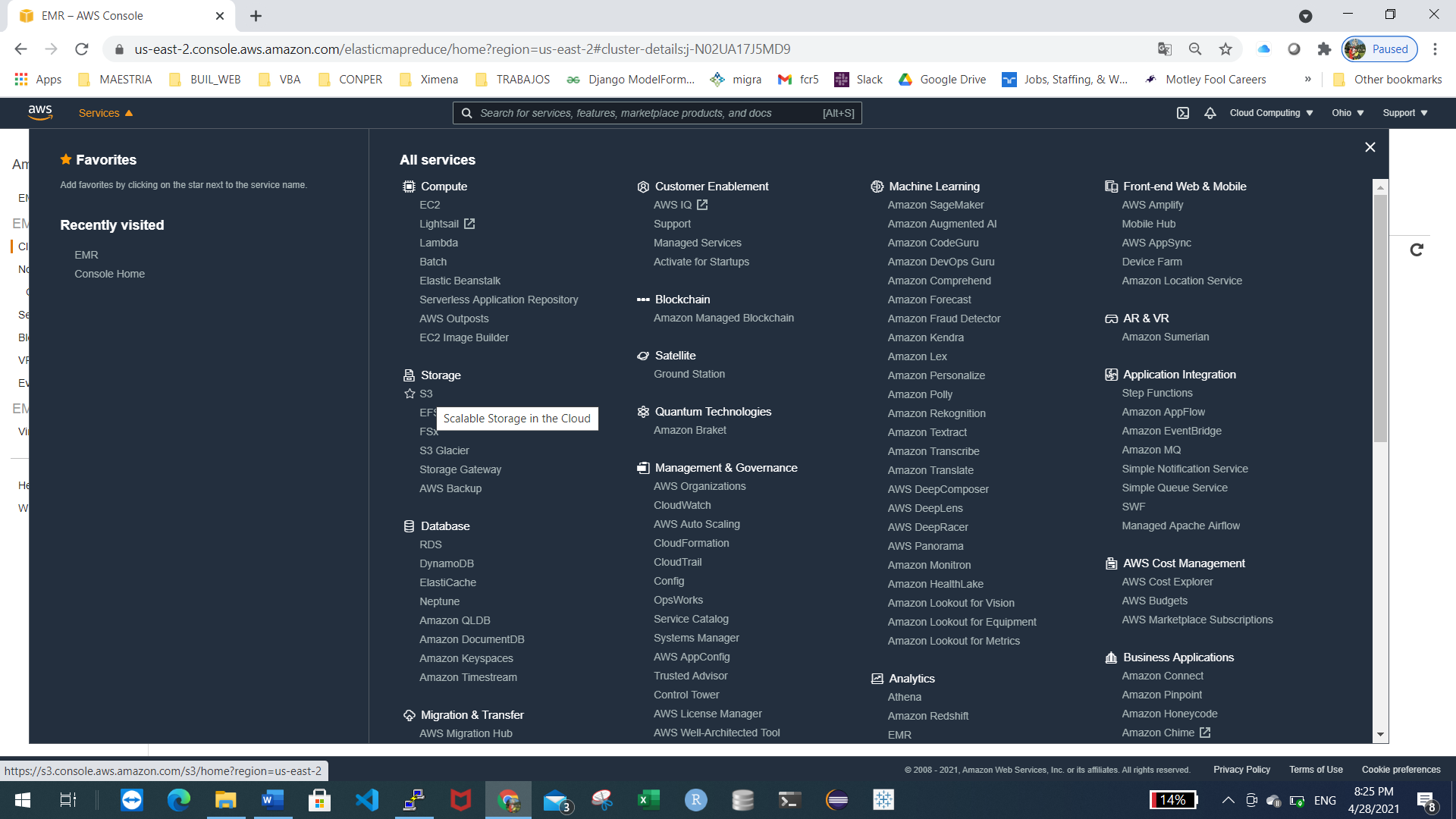


The picture above shows the connection to the master node.

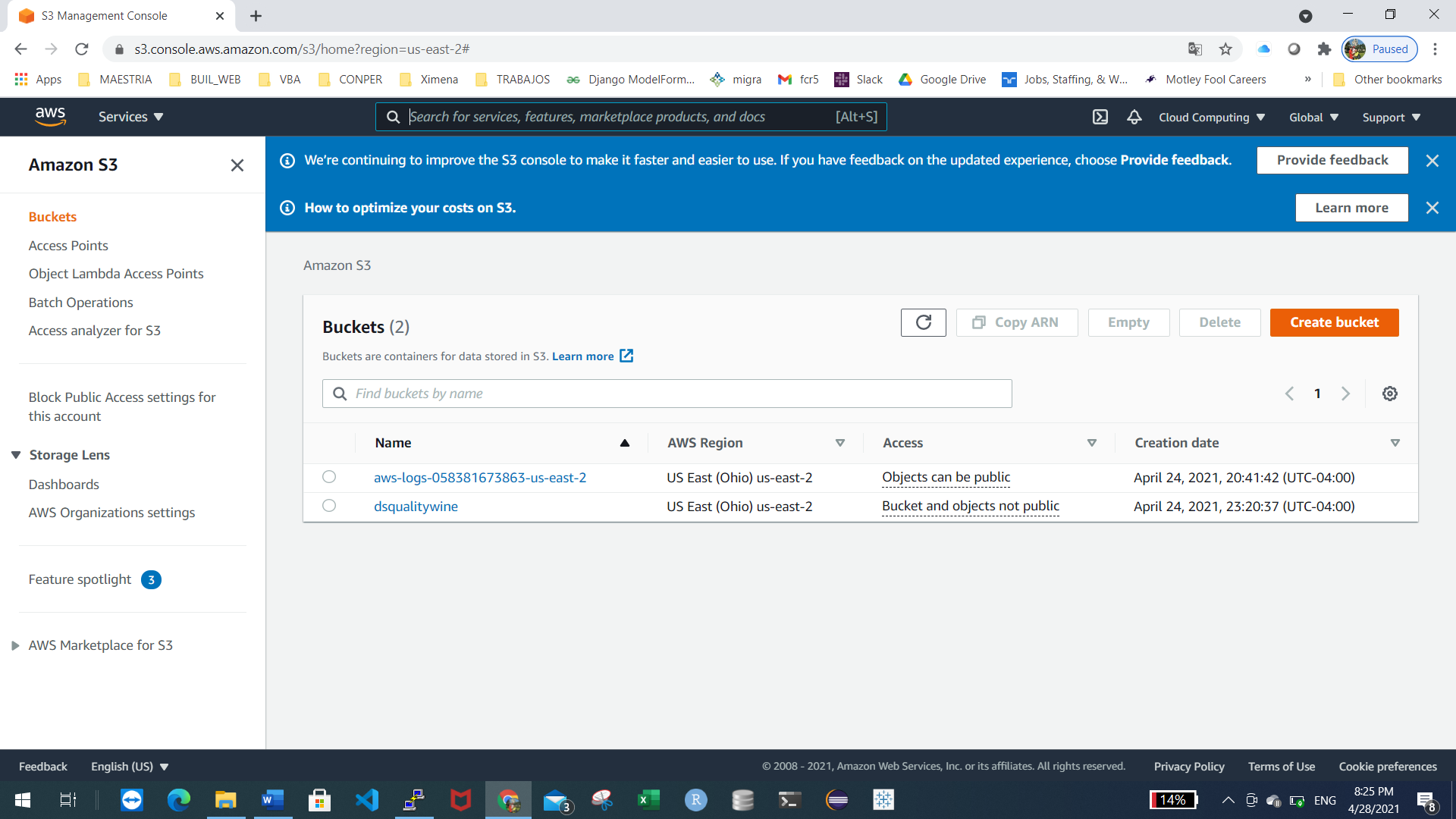
Now, we need to upload our files training and validation dataset to S3

S3 Configuration

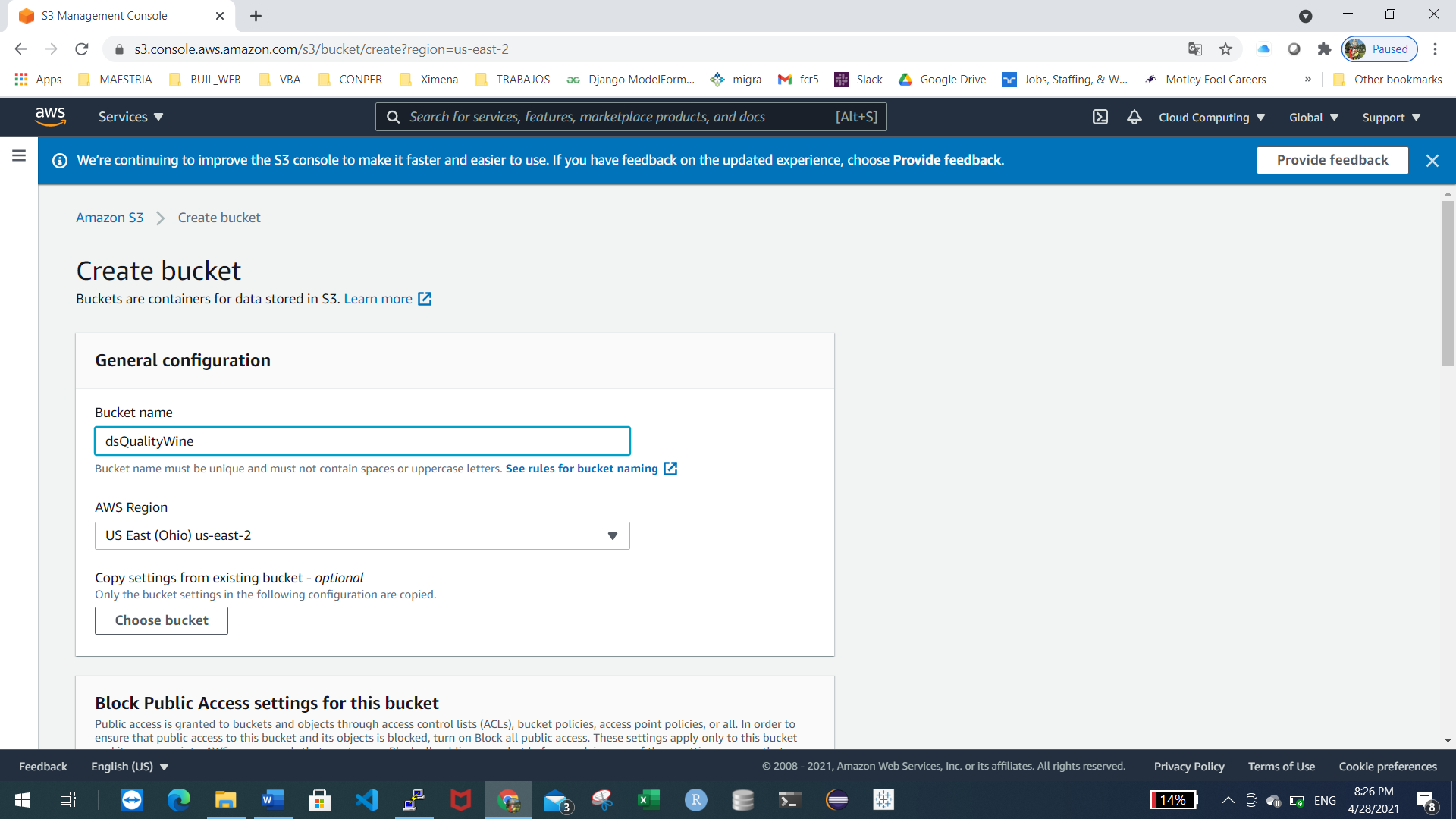
Choose S3 service



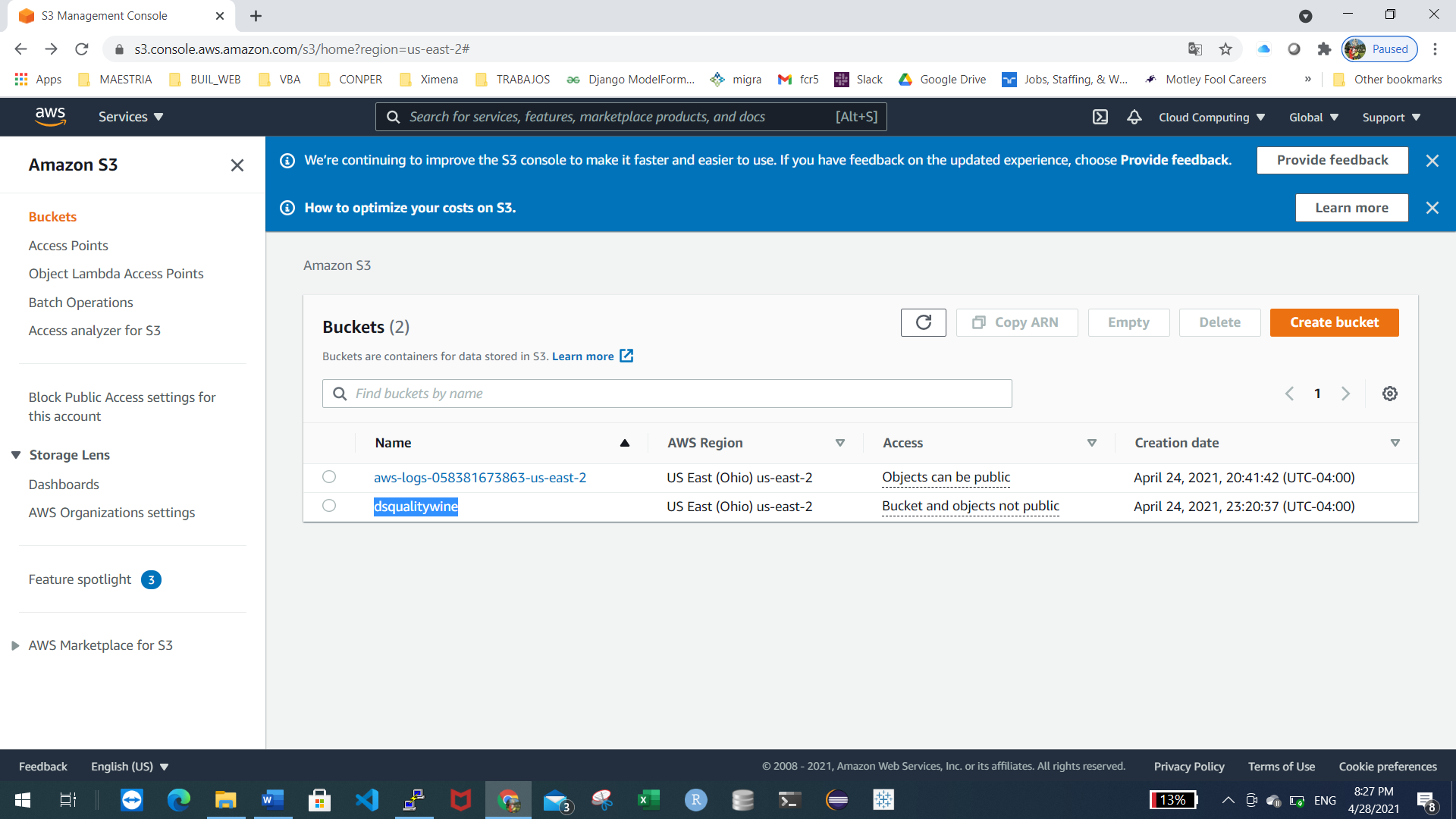
Click on create bucket button



Give a bucket name

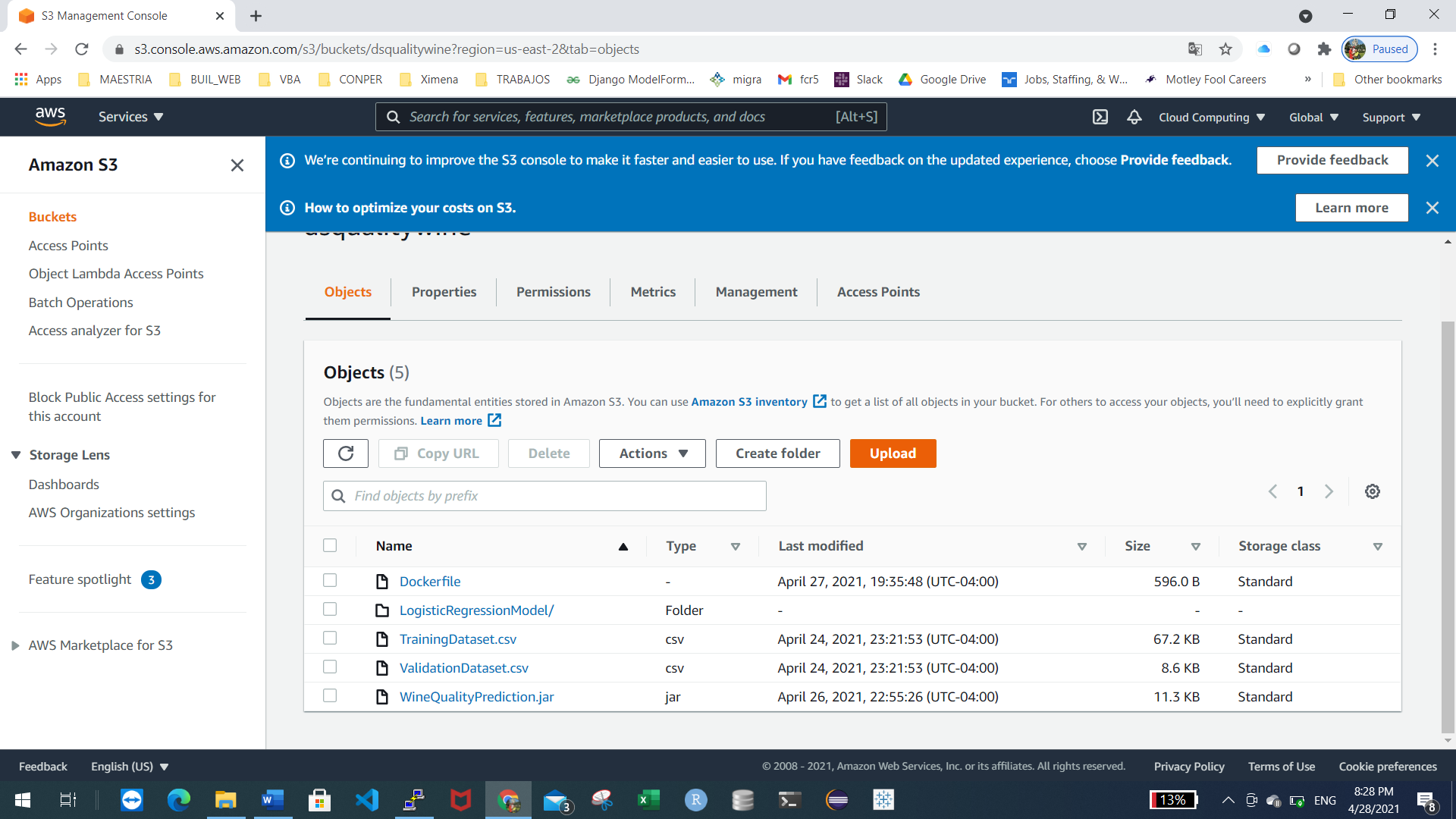


Click on create bucket button, we can see our bucket create

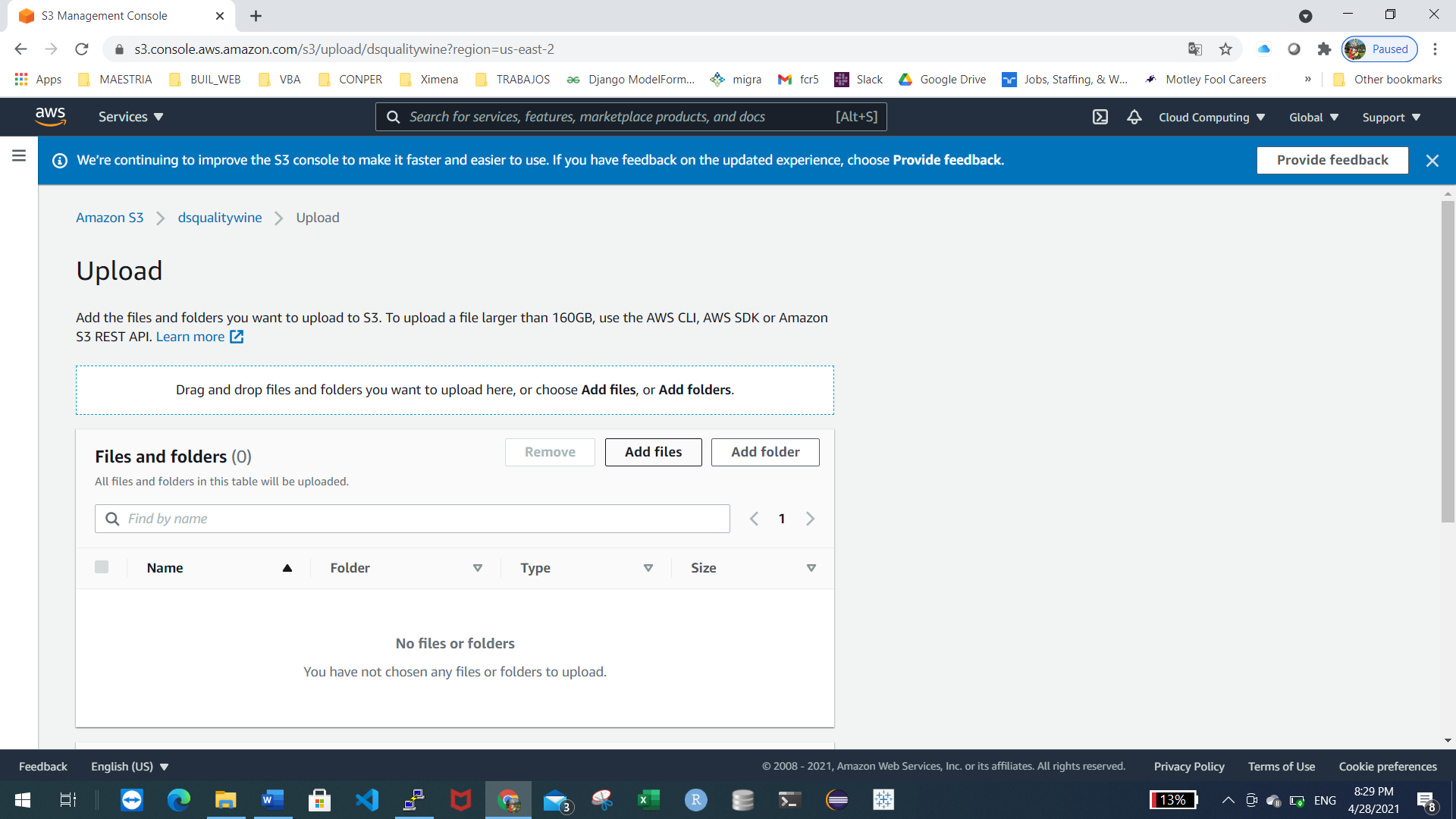


Now, upload the files:

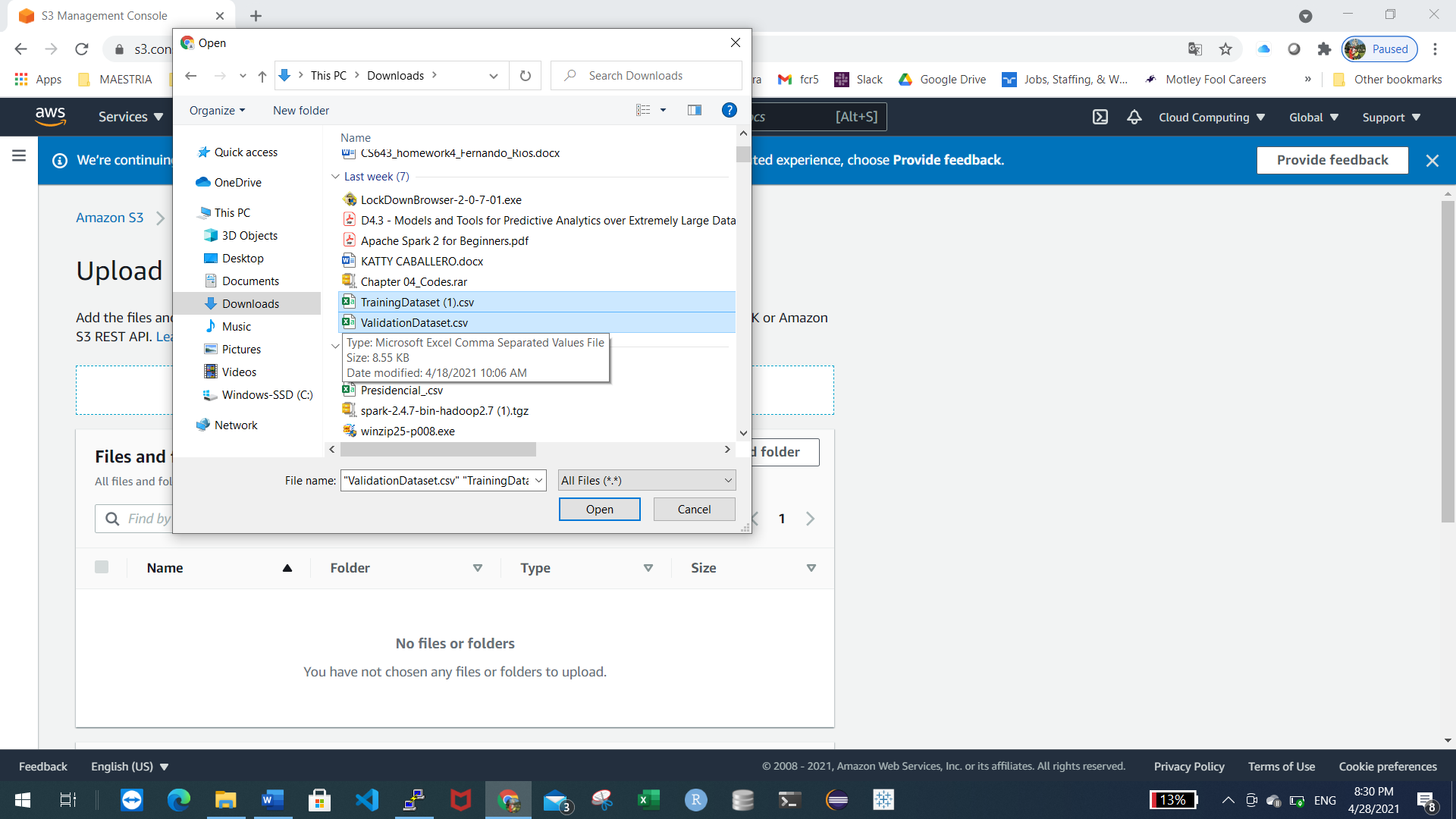
Into the bucket, click on upload button



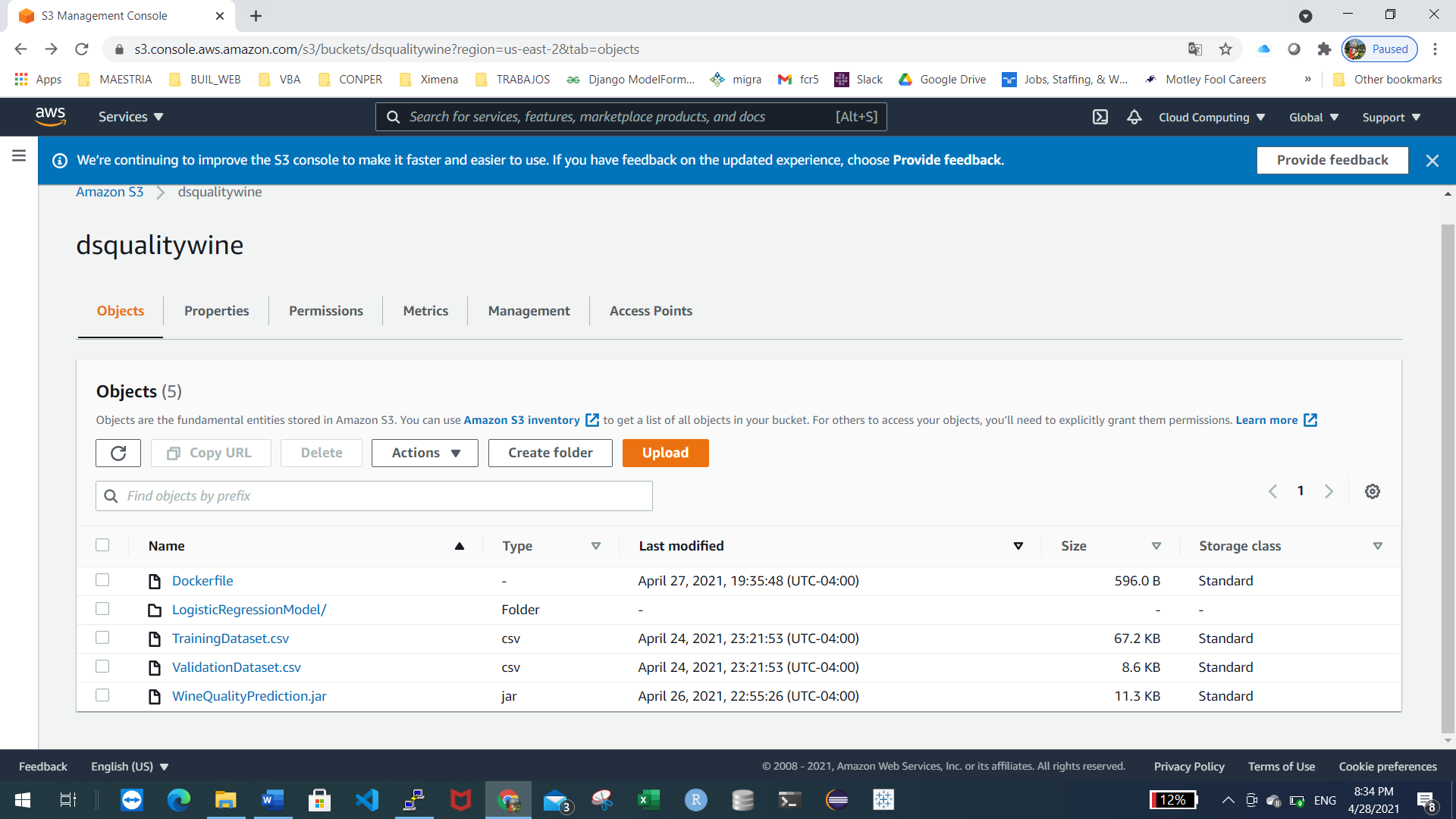
Click on Add Files



Choose the files



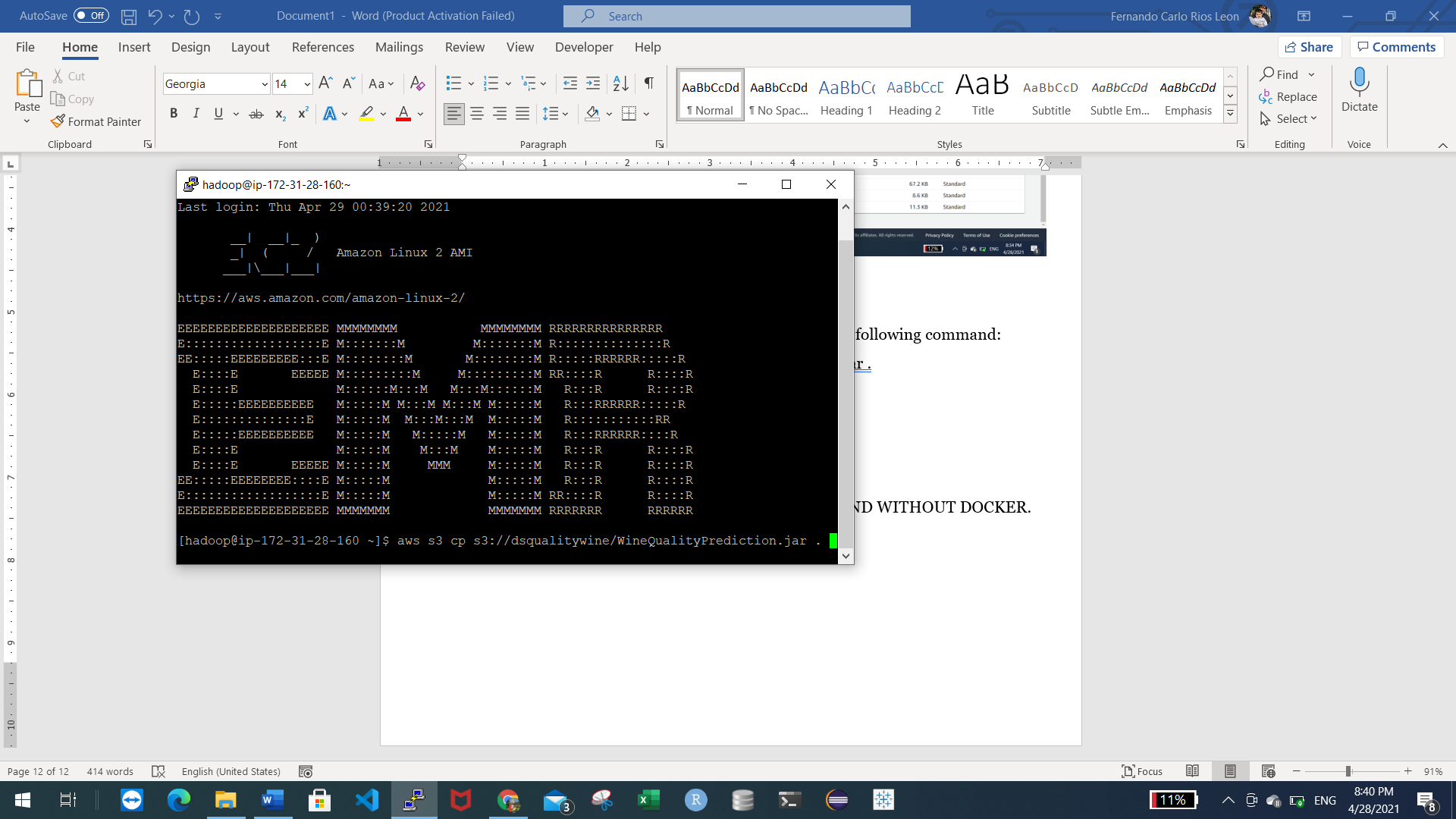
The files uploaded will appear on the list.



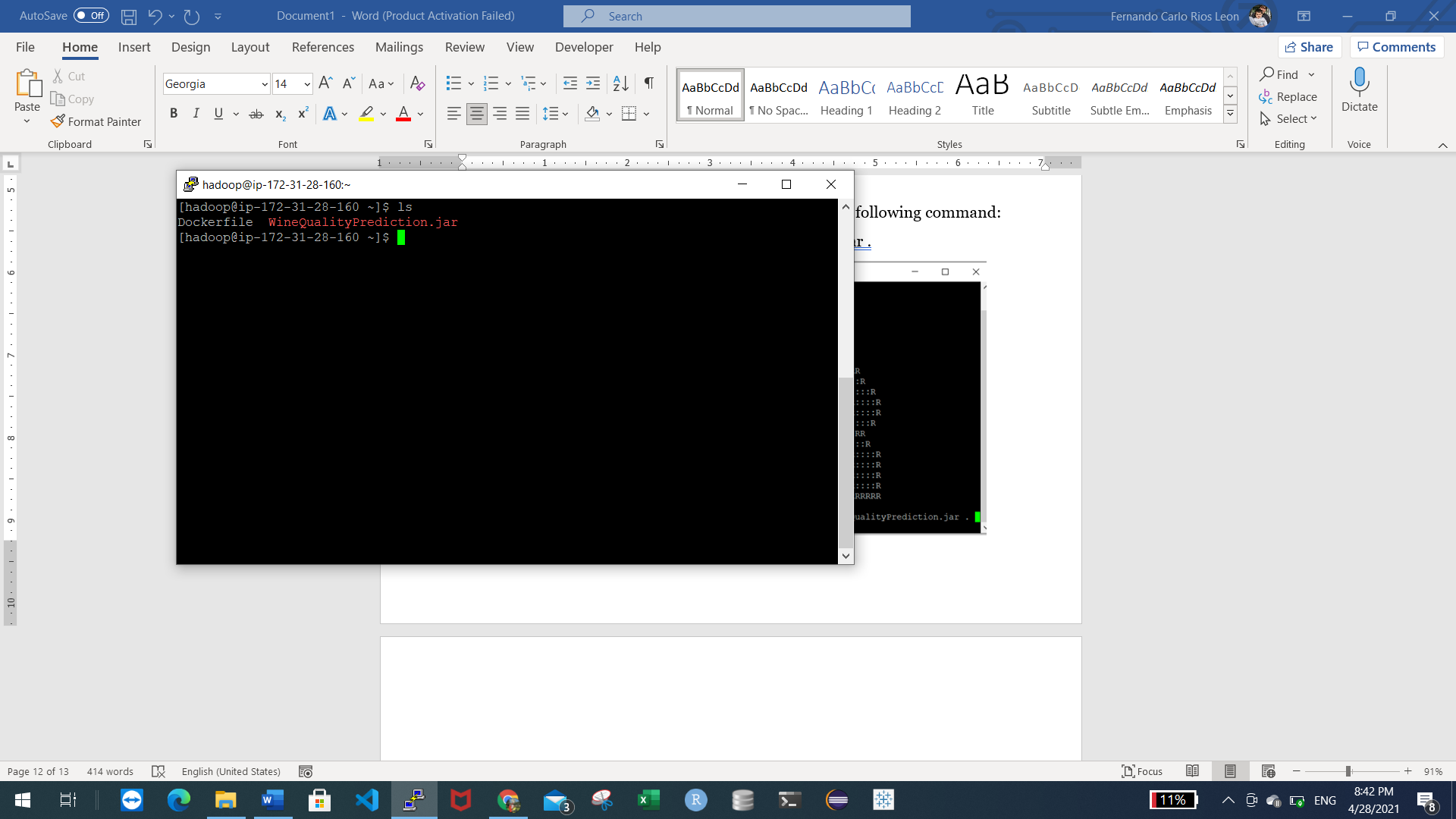
Training the model

Copy from S3 to the master node the JAR file using the following command:

aws s3 cp s3://dsqualitywine/WineQualityPrediction.jar .

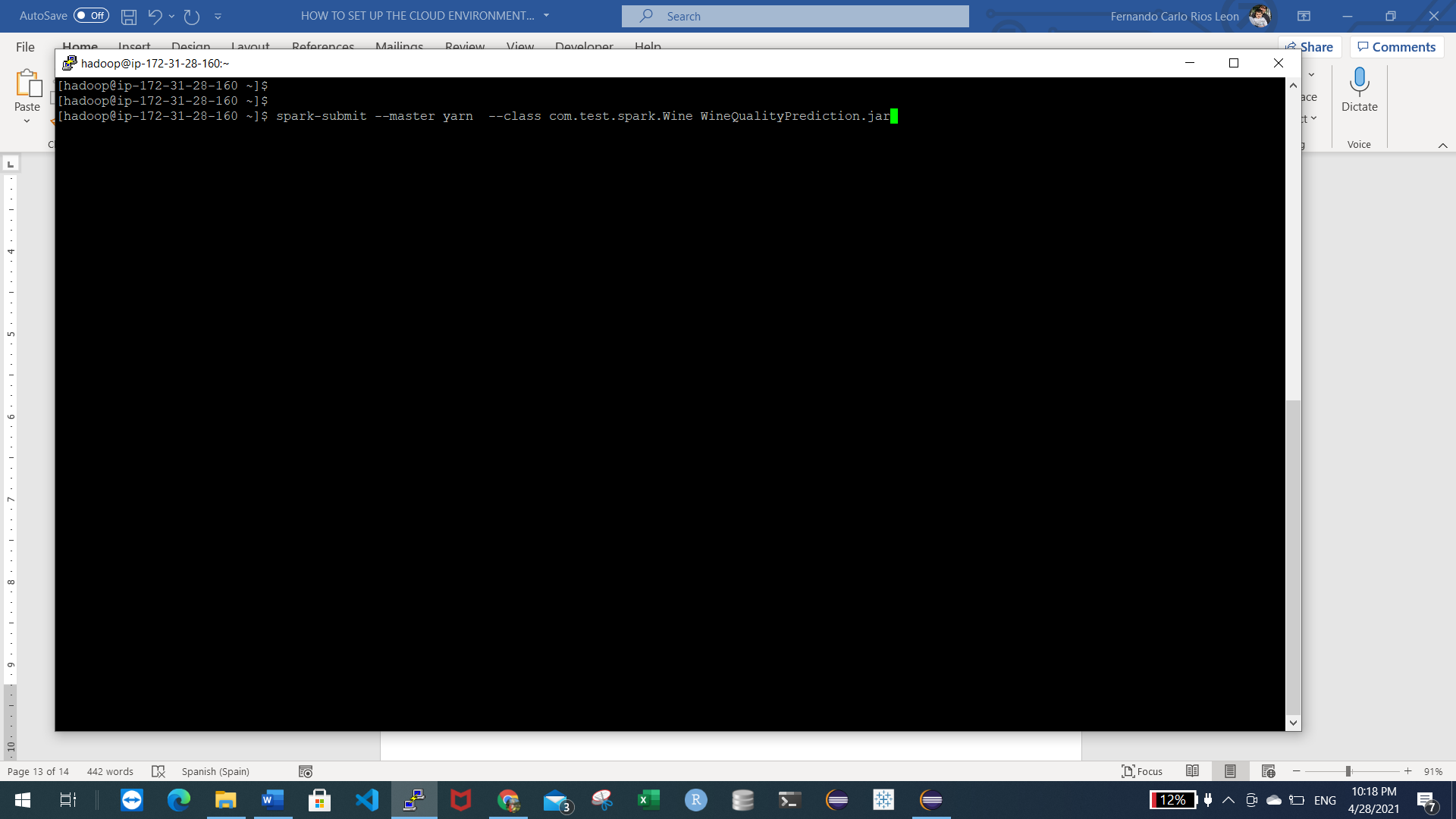


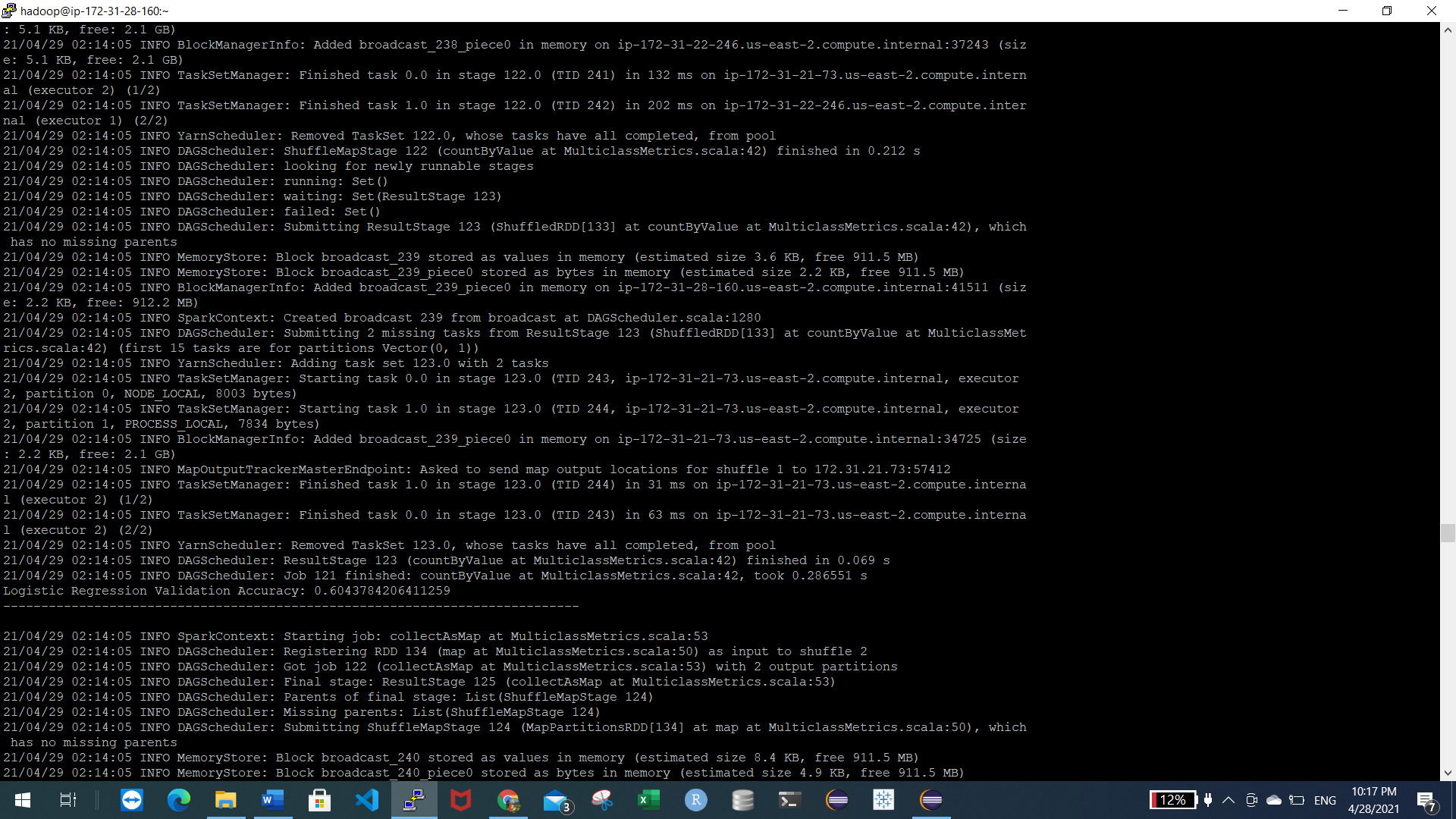
See the copied file using ls command



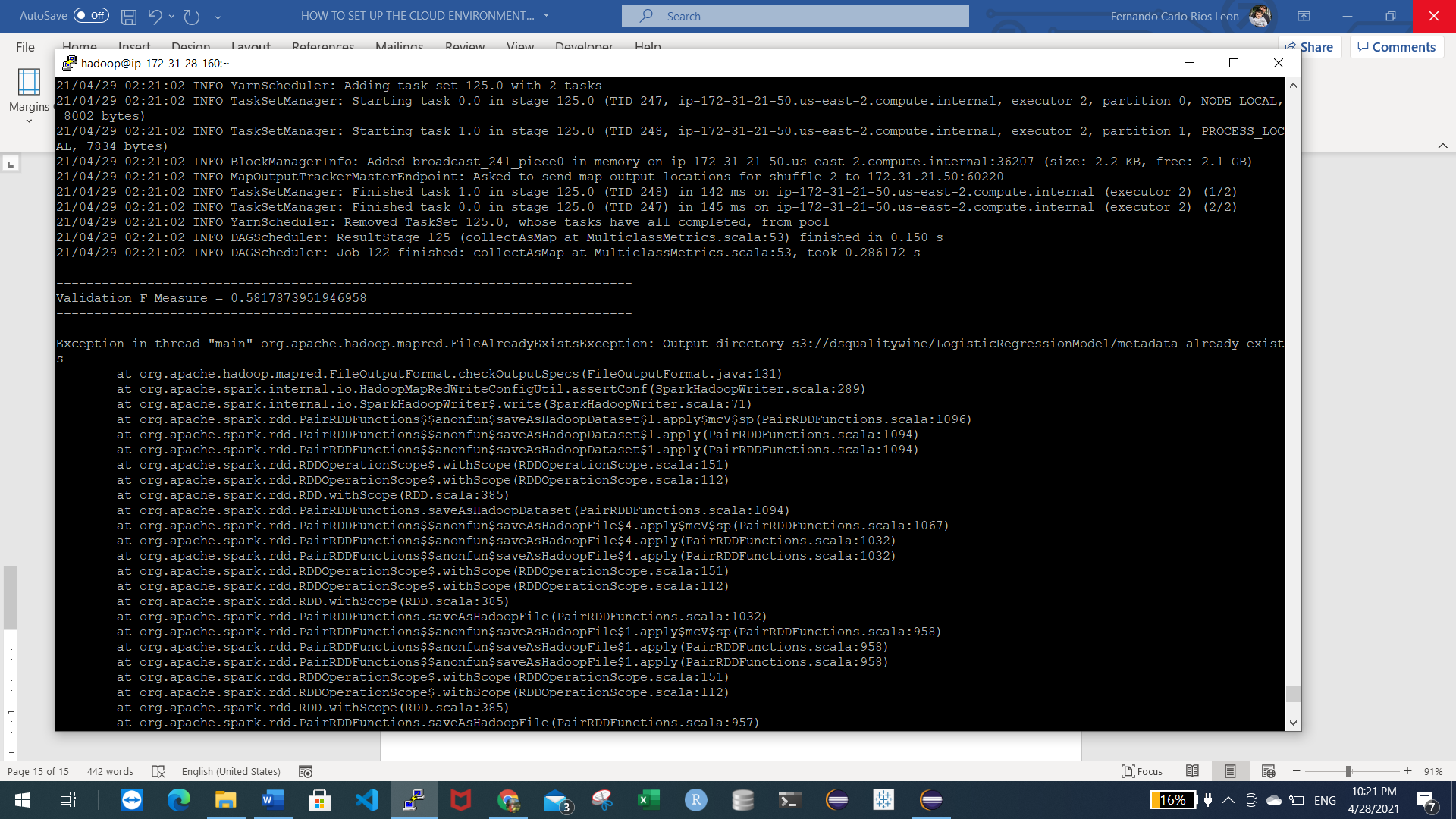
Execute the training model using the following command:

spark-submit --master yarn --class com.test.spark.Wine WineQualityPrediction.jar



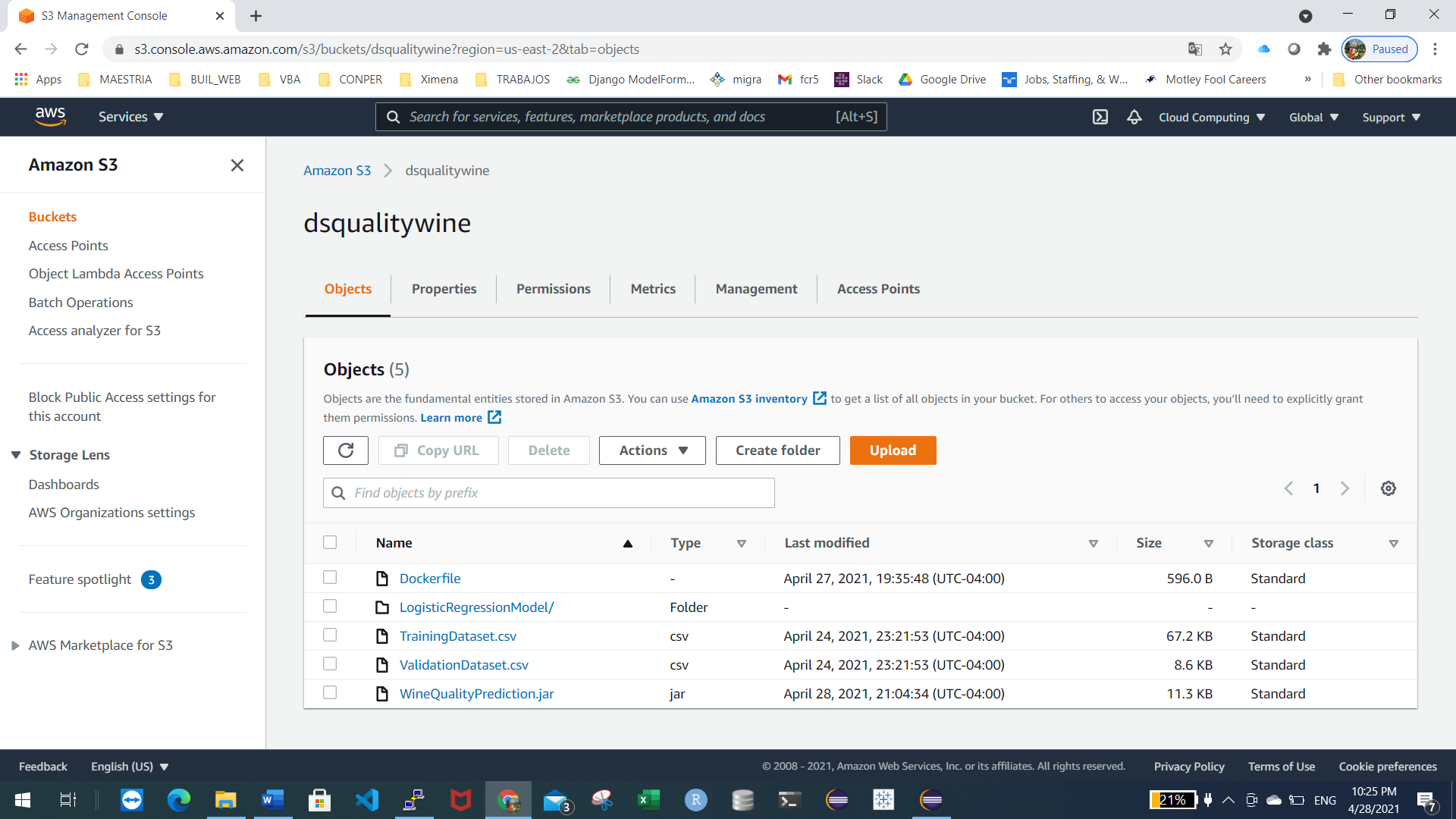


Logistic Regression Validation Accuracy: 0.60



Validation F1 Measure: 0.58

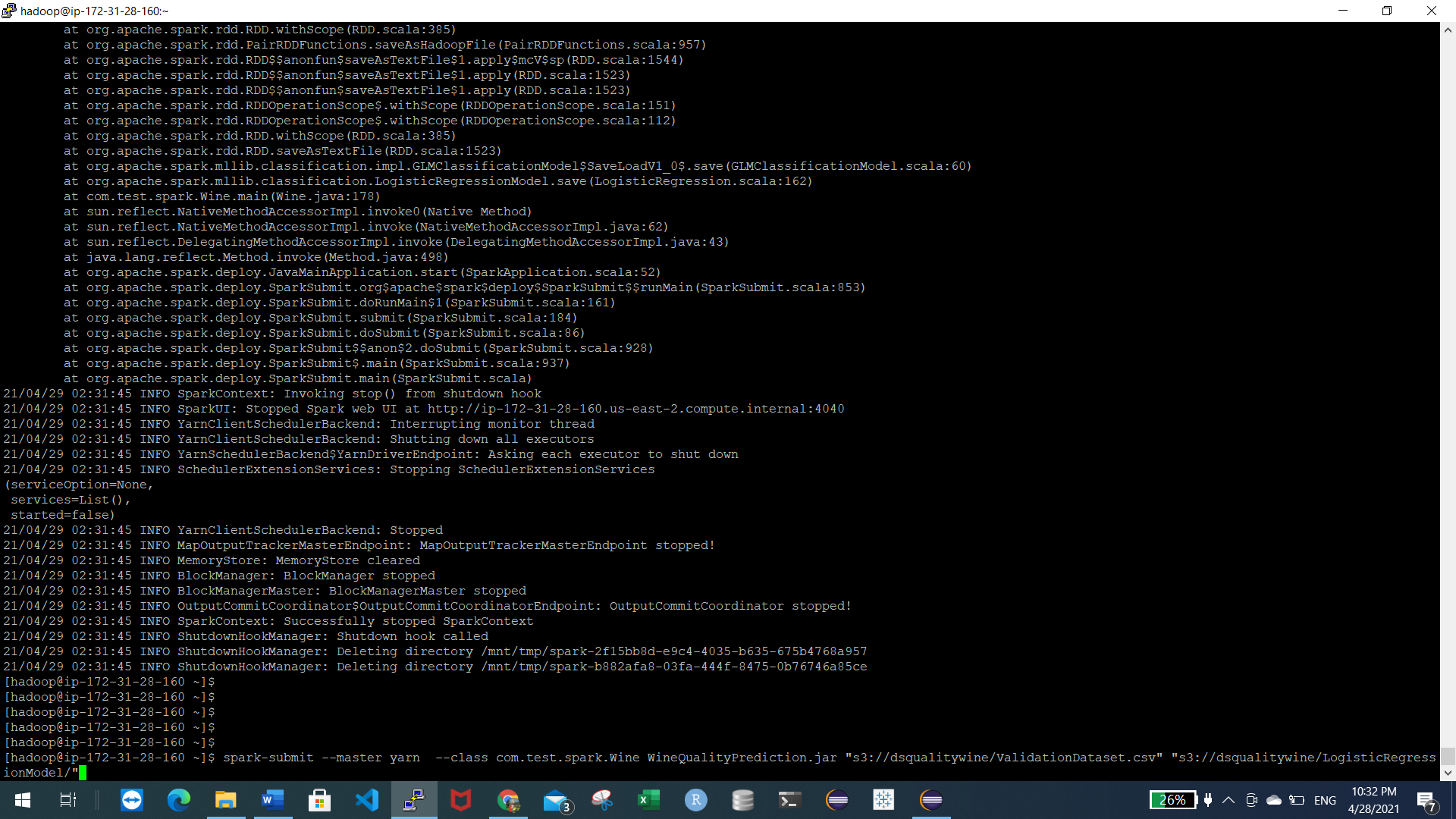
The application save the model in a folder called “LogisticRegressionModel” in S3:

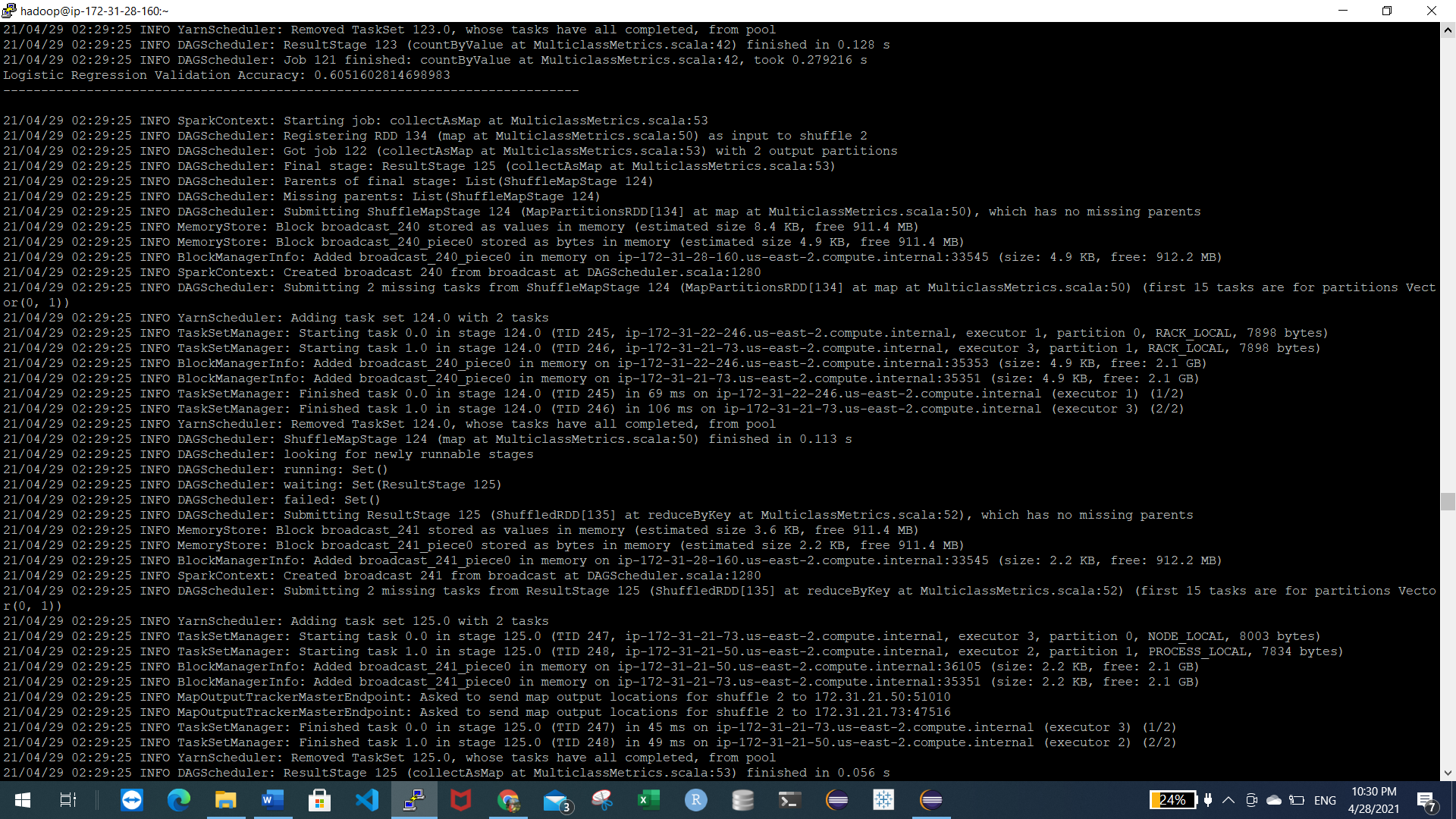


HOW TO RUN THE APPLICATION PREDICTION WITHOUT DOCKER.

Execute the model using the following command, it includes 2 parameters: the testdataset.csv and the model folder in S3

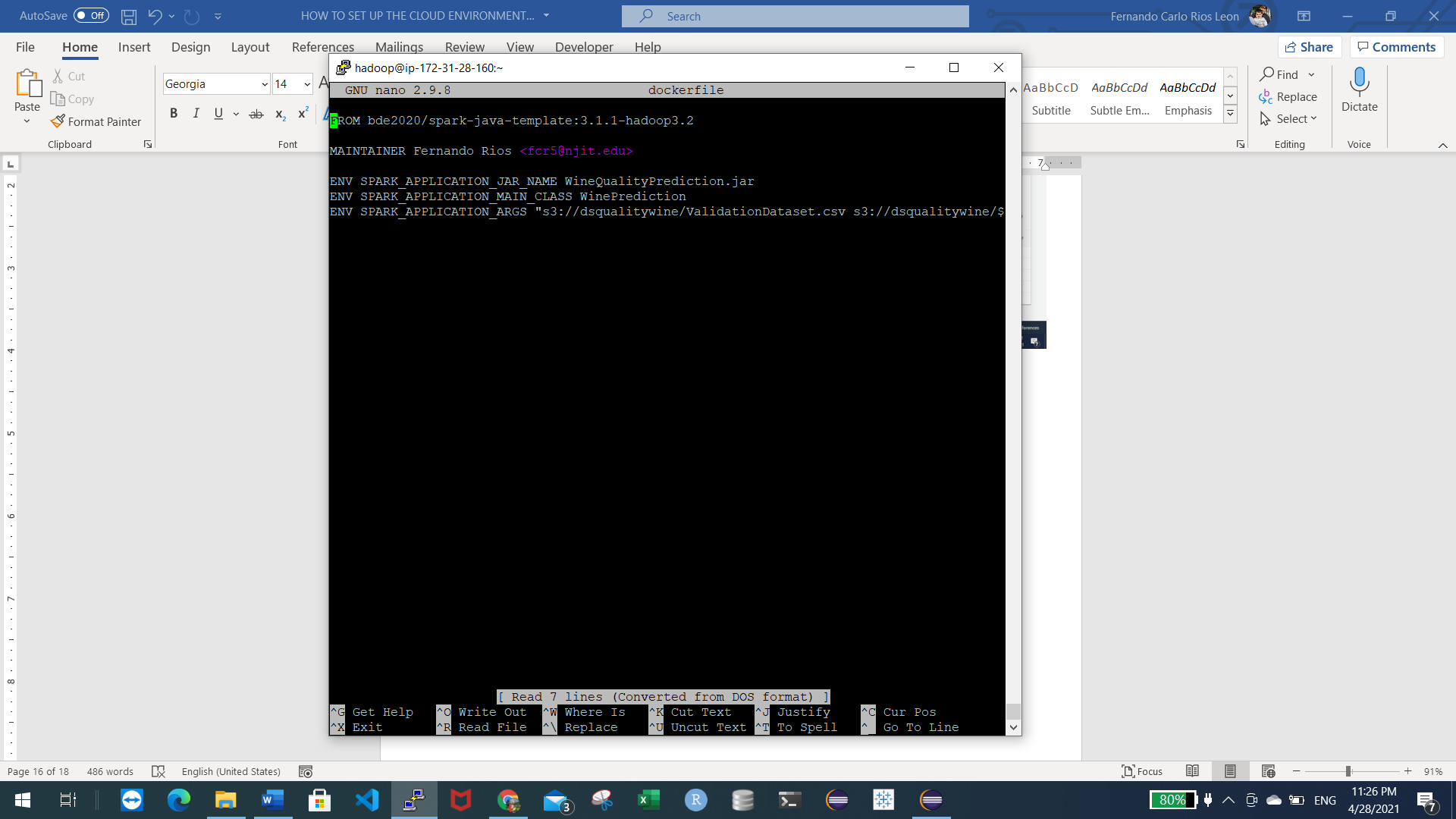
spark-submit --master yarn --class com.test.spark.Wine WineQualityPrediction.jar "s3://dsqualitywine/ValidationDataset.csv" "s3://dsqualitywine/LogisticRegressionModel/"





HOW TO RUN THE APPLICATION PREDICTION WITHOUT DOCKER.

Dockerfile view:



Build container

