

## Programming Assignment 2

### Cloud Computing

Name – Gayatri Aavula

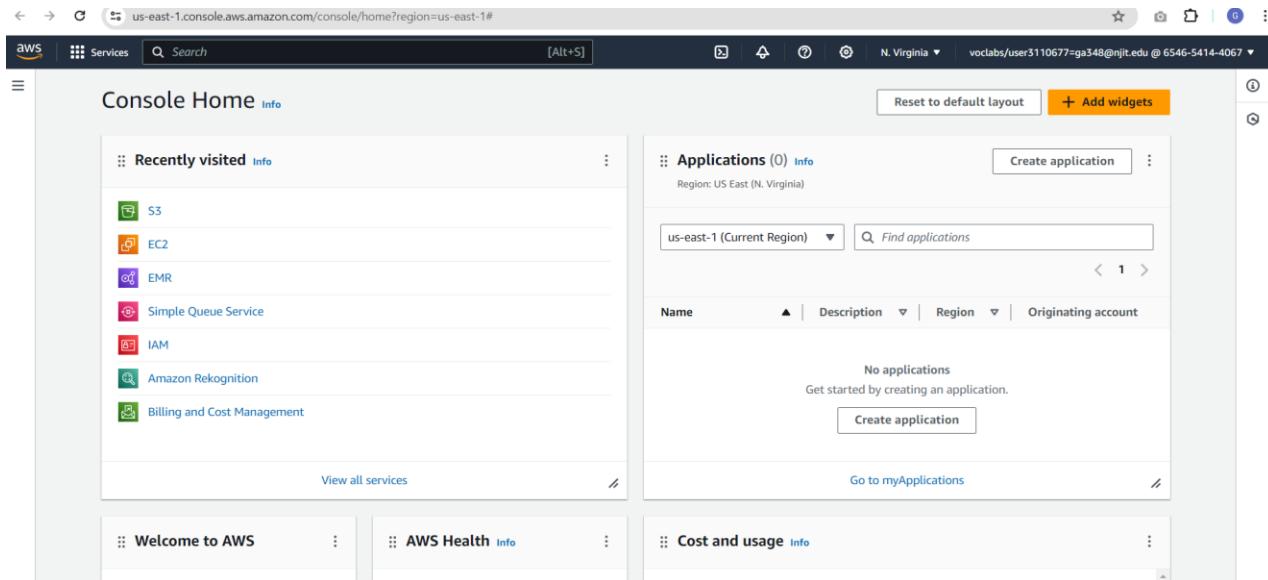
UCID – ga348@njit.edu

GitHub – [https://github.com/gayatriaavula/winepred\\_quality.git](https://github.com/gayatriaavula/winepred_quality.git)

DockerHub - <https://hub.docker.com/repository/docker/ga348/cs643-programming-assignment-2/>

1. Log into the AWS Management Console.

2. From the AWS Management Console, select the EMR service from the list of available services. Then, choose EMR on EC2 Clusters.



3. You'll be taken to the Clusters page where you'll see that there are currently no active clusters.

4. To create a new cluster, click on the "Create cluster" button.

Amazon EMR > EMR on EC2: Clusters

Clusters (0) <a href="#">Info</a>		<a href="#">C</a>	<a href="#">View details</a>	<a href="#">Terminate</a>	<a href="#">Clone</a>	<a href="#">Create cluster</a>
<a href="#">Filter clusters by status</a> <a href="#">▼</a>		<a href="#"> Find clusters</a>				
<a href="#"> Filter clusters by creation date-time</a>						
<a href="#"></a>		<a href="#">Cluster ID</a>	<a href="#">▼</a>	<a href="#">Cluster name</a>	<a href="#">▼</a>	<a href="#">Status</a>
<b>No Clusters</b> No Clusters to display.						

5. Give any name you want to the cluster. Additionally, ensure that you select the latest version of EMR available and choose 'Spark Interactive' under the Application bundle options.

Amazon EMR > EMR on EC2: Clusters > Create cluster

### Create cluster [Info](#)

**Name and applications - required** [Info](#)

Name your cluster and choose the applications that you want to install to your cluster.

Name:

Amazon EMR release [Info](#)

A release contains a set of applications which can be installed on your cluster.

emr-7.1.0

Application bundle

Spark Interactive	Core Hadoop	Flink	HBase	Presto	Trino	Custom
-------------------	-------------	-------	-------	--------	-------	--------

AmazonCloudWatchAgent 1.300032.2  
 HCatalog 3.1.3  
 Hue 4.11.0  
 Livy 0.8.0  
 Phoenix 5.1.3  
 Spark 3.5.0  
 Tez 0.10.2  
 ZooKeeper 3.9.1

AWS Glue Data Catalog settings

**Summary** [Info](#)

**Name and applications - required**

Name: winequality\_prediction

Amazon EMR release: emr-7.1.0

Application bundle: Spark Interactive (Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)

**Cluster configuration - required**

Uniform instance groups: Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)

**Cluster scaling and provisioning - required**

Provisioning configuration

**Configure IAM roles**

You must choose a service role and instance profile.

**6.I selected the instance type m5.xlarge for the Primary, Core, and Task instance groups. You have the flexibility to choose any instance type that suits your needs.**

The screenshot shows the 'Cluster configuration - required' step of the AWS EMR wizard. It includes sections for 'Primary', 'Core', and 'Task' instance groups, each with an 'Actions' button. Below these are optional node configuration sections. To the right, a summary panel shows the cluster name as 'winequality\_prediction', application bundle as 'emr-7.1.0', and uniform instance groups as 'Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)'. A 'Cluster scaling and provisioning - required' section is also present.

**7. To ensure cluster scalability and provisioning, configure Core's instance size to 1 and Task-1's instance size to 2.**

The screenshot shows the 'Cluster scaling and provisioning - required' configuration screen. It features three options: 'Set cluster size manually' (selected), 'Use EMR-managed scaling', and 'Use custom automatic scaling'. Under 'Provisioning configuration', it specifies 'Core' with 1 instance and 'Task - 1' with 2 instances. To the right, a summary panel shows the same cluster details as the previous screenshot, including the selected provisioning configuration.

- 8.** In order to pick the security groups for EC2's Primary node, Core, and task nodes, select the security groups that are displayed below.

The screenshot shows the 'Name and applications' section on the right, which includes the cluster name 'winequality\_prediction', application bundle 'Amazon EMR release emr-7.1.0', and JupyterEnterpriseGateway 2.3.5.... The main area displays the 'EC2 security groups (firewall)' configuration. It shows two sections: 'Primary node' and 'Core and task nodes'. Under 'Primary node', there is an 'EMR-managed security group' dropdown containing 'ElasticMapReduce-Primary' (sg-0d6a41217a7a3a91b) and an optional 'Additional security groups' dropdown labeled 'Choose additional security groups'. Similarly, under 'Core and task nodes', there is an 'EMR-managed security group' dropdown containing 'ElasticMapReduce-Core' (sg-0f3d5c33b4411ee5e) and an optional 'Additional security groups' dropdown labeled 'Choose additional security groups'.

- 9.** To prevent the cluster from being terminated automatically, make sure that the option to manually end the cluster is selected.

The screenshot shows the 'Cluster termination and node replacement' section. It includes a note about using commands and scripts to tell the cluster where to find and how to process data. The 'Termination option' section has three radio button options: 'Manually terminate cluster' (selected), 'Automatically terminate cluster after last step ends', and 'Automatically terminate cluster after idle time (Recommended)'. Below this is a checked checkbox for 'Use termination protection', with a note explaining it protects the cluster from accidental termination. There is also a link to 'Unhealthy node replacement - new'.

- 10.** create a new key pair

## ▼ Security configuration and EC2 key pair [Info](#)

Choose a security configuration or create a new one that you can reuse with other clusters.

### Security configuration

Select your cluster encryption, authentication, authorization, and instance metadata service settings.

Choose a security configuration



Browse

Create security configuration

### Amazon EC2 key pair for SSH to the cluster [Info](#)

Enter a key name or choose Browse to select an Amazon EC2 key

Browse

Create key pair

**⚠** You haven't entered an EC2 key. If you're outside a VPN and want to enable SSH or use Hue SQL assistant with this cluster, you must enter an EC2 key.

- 11.** Provide a custom name for the key pair and select 'ppk' as the file format when creating the key pair

EC2 > [Key pairs](#) > Create key pair

Create key pair [Info](#)

**Key pair**  
A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name  The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

Key pair type [Info](#)  RSA  ED25519

Private key file format  .pem For use with OpenSSH  .ppk For use with PuTTY

Tags - optional  
No tags associated with the resource.  Add new tag  
You can add up to 50 more tags.

Cancel

- 12.** The key pair named 'programming-2' has been created, and the key has been downloaded and stored on the local system for connecting to the cluster using PuTTY as the SSH server.

Name	Type	Created	Fingerprint	ID
programming-2	rsa	2024/04/27 13:25 GMT-4	04:12:61:5e:6a:ab:c0:4e:91:50:66:c9...	key-027f6a51068aaaf57e
vockey	rsa	2024/02/26 20:05 GMT-5	3a:3d:39:e3:49:f1:82:27:a5:30:8e...	key-0c6dc6a49a14533b
Object	rsa	2024/02/27 16:54 GMT-5	18:11:79:fe:67:12:4c:77:8:ded:ae:4f...	key-0609ad02da45af865
wineprediction	rsa	2024/04/21 15:41 GMT-4	25:55:b3:e0:02:fe:3a:ac:55:ae:b8:a5:...	key-0d69cf73abcebbe7c
prediction	rsa	2024/04/21 15:43 GMT-4	fa:48:fe:66:36:ba:60:55:ae:91:43:98:...	key-06f994487d3b3c28e

13. The key pair named 'programming-2' has been created, and you can now browse the key .

▼ Security configuration and EC2 key pair [Info](#)

Choose a security configuration or create a new one that you can reuse with other clusters.

Security configuration

Select your cluster encryption, authentication, authorization, and instance metadata service settings.

Choose a security configuration

Amazon EC2 key pair for SSH to the cluster [Info](#)

programming-2

▼ Identity and Access Management (IAM) roles - required [Info](#)

14. Go to the IAM roles and choose them as indicated below.

**Amazon EMR service role** Info

The service role is an IAM role that Amazon EMR assumes to provision resources and perform service-level actions with other AWS services.

Choose an existing service role  
Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.

Create a service role  
Let Amazon EMR create a new service role so that you can grant and restrict access to resources in other AWS services.

Service role

▼

---

**EC2 instance profile for Amazon EMR**

The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

Choose an existing instance profile  
Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.

Create an instance profile  
Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

Instance profile

▼

---

**Custom automatic scaling role - optional**

**Summary** Info

---

**Name and applications - required**

Name  
winequality\_preditction

Amazon EMR release  
emr-7.1.0

Application bundle  
Spark Interactive (Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spar 3.5....)

**Cluster configuration - required**

Uniform instance groups  
Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)

**Cluster scaling and provisioning - required**

Provisioning configuration

15. Click the button labeled "Create Cluster" to initiate the process of creating the clusters.

Select a default service role or a custom role with IAM policies attached so that your cluster can interact with other AWS services.

Service role

▼

---

**EC2 instance profile for Amazon EMR**

The instance profile assigns a role to every EC2 instance in a cluster. The instance profile must specify a role that can access the resources for your steps and bootstrap actions.

Choose an existing instance profile  
Select a default role or a custom instance profile with IAM policies attached so that your cluster can interact with your resources in Amazon S3.

Create an instance profile  
Let Amazon EMR create a new instance profile so that you can specify a custom set of resources for it to access in Amazon S3.

Instance profile

▼

---

**Custom automatic scaling role - optional**

When a custom automatic scaling rule triggers, Amazon EMR assumes this role to add and terminate EC2 instances. Learn more

Custom automatic scaling role

Name  
winequality\_preditction

Amazon EMR release  
emr-7.1.0

Application bundle  
Spark Interactive (Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5....)

**Cluster configuration - required**

Uniform instance groups  
Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)

**Cluster scaling and provisioning - required**

Provisioning configuration

Cancel
 Create cluster

16. Cluster created successfully.

Your cluster "winequality\_prediction" has been successfully created.

Amazon EMR > EMR on EC2: Clusters > winequality\_prediction

## winequality\_prediction

Updated less than a minute ago

[Edit](#) [Terminate](#) [Clone in AWS CLI](#) [Clone](#)

Summary		Status and time	
Cluster info	Applications	Cluster management	Status
Cluster ID j-P2PSYL3YJBVF	Amazon EMR version emr-7.1.0	Log destination in Amazon S3 aws-logs-654654144067-us-east-1/elasticmapreduce	Starting
Cluster configuration	Installed applications	Primary node public DNS	Creation time
Instance groups	Hadoop 3.3.6, Hive 3.1.3, JupyterEnterpriseGateway 2.6.0, Livy 0.8.0, Spark 3.5.0	-	April 27, 2024, 13:43 (UTC-04:00)
Capacity	1 Primary 1 Core 2 Task		Elapsed time
			0 seconds

[Properties](#) [Bootstrap actions](#) [Instances \(Hardware\)](#) [Steps](#) [Applications](#) [Configurations](#) [Monitoring](#) [Events](#) [Tags \(0\)](#)

**Cluster logs** [Info](#)

Archive log files to Amazon S3  
Turned on

Amazon S3 location  
<s3://aws-logs-654654144067-us-east-1/elasticmapreduce/>

**Cluster termination and node replacement** [Info](#) [Edit](#)

Termination option  
Manually terminate cluster

Idle time  
-

Termination protection  
On

Unhealthy node replacement  
On

**Network and security** [Info](#)

CloudShell [Feedback](#) © 2024, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

17. Navigate to the EC2 instance page. As seen below, there are four EC2 instances launched, one of which is a Master node and the other three are Slave nodes.

**Instances** [Info](#)

Find Instance by attribute or tag (case-sensitive) [Clear filters](#) All states ▾

Instance state = running [X](#)

<input type="checkbox"/>	Name	Instance ID	Instance state	Instance type	Status check	Alarm status	Availability Zone	Public IPv4 DNS
<input type="checkbox"/>	i-0f7ca45c3a5f4dd63	i-0f7ca45c3a5f4dd63	Running	m5.xlarge	Initializing	<a href="#">View alarms</a> +	us-east-1f	ec2-44-211-27-0
<input type="checkbox"/>	i-0e63c06c8bc25788f	i-0e63c06c8bc25788f	Running	m5.xlarge	Initializing	<a href="#">View alarms</a> +	us-east-1f	ec2-35-170-74-0
<input type="checkbox"/>	i-07723f20b4e5321f4	i-07723f20b4e5321f4	Running	m5.xlarge	Initializing	<a href="#">View alarms</a> +	us-east-1f	ec2-18-232-153-
<input type="checkbox"/>	i-0c1a32e92b5b911bc	i-0c1a32e92b5b911bc	Running	m5.xlarge	Initializing	<a href="#">View alarms</a> +	us-east-1f	ec2-3-235-79-90

**Instances (4) Info**

Find Instance by attribute or tag (case-sensitive) [Clear filters](#) All states ▾

Instance state = running [X](#)

4 ...	Elastic IP	IPv6 IPs	Monitoring	Security group name	Key name	Launch time	Platform
.0	-	-	disabled	ElasticMapReduce-master	programming-2	2024/04/27 13:43 GMT-4	Linux/UNIX
.0	-	-	disabled	ElasticMapReduce-slave	programming-2	2024/04/27 13:43 GMT-4	Linux/UNIX
3.122	-	-	disabled	ElasticMapReduce-slave	programming-2	2024/04/27 13:43 GMT-4	Linux/UNIX
0	-	-	disabled	ElasticMapReduce-slave	programming-2	2024/04/27 13:43 GMT-4	Linux/UNIX

18. To access the "ElasticMapReduce-Master" security group, navigate to the EC2 service and select the appropriate Security ID.

Instance summary for i-0f7ca45c3a5f4dd63		
<a href="#">EC2</a> > <a href="#">Instances</a> > i-0f7ca45c3a5f4dd63		
Updated less than a minute ago		
Instance ID <a href="#">i-0f7ca45c3a5f4dd63</a>	Public IPv4 address <a href="#">44.211.27.0</a>   <a href="#">open address</a>	Private IPv4 addresses <a href="#">172.31.76.1</a>
IPv6 address -	Instance state <a href="#">Running</a>	Public IPv4 DNS <a href="#">ec2-44-211-27-0.compute-1.amazonaws.com</a>   <a href="#">open address</a>
Hostname type IP name: ip-172-31-76-1.ec2.internal	Private IP DNS name (IPv4 only) <a href="#">ip-172-31-76-1.ec2.internal</a>	Elastic IP addresses -
Answer private resource DNS name -	Instance type m5.xlarge	AWS Compute Optimizer finding <a href="#">Opt-in to AWS Compute Optimizer for recommendations.</a> <a href="#">Learn more</a>
Auto-assigned IP address <a href="#">44.211.27.0 [Public IP]</a>	VPC ID <a href="#">vpc-0512e1100d889881a</a>	Auto Scaling Group name -
IAM Role <a href="#">EMR_EC2_DefaultRole</a>	Subnet ID <a href="#">subnet-02fd81fc1fbf276ad</a>	
IMDSv2 Required		
<a href="#">Details</a> <a href="#">Status and alarms</a> <a href="#">New</a> <a href="#">Monitoring</a> <a href="#">Security</a> <a href="#">Networking</a> <a href="#">Storage</a> <a href="#">Tags</a>		
<b>Instance details</b> <a href="#">Info</a>		
Platform <a href="#">Linux/UNIX (Inferred)</a>	AMI ID <a href="#">ami-0f06d16b26c594254</a>	Monitoring disabled

Details		
<a href="#">Status and alarms</a> <a href="#">New</a> <a href="#">Monitoring</a> <a href="#">Security</a> <a href="#">Networking</a> <a href="#">Storage</a> <a href="#">Tags</a>		
<b>Instance details</b> <a href="#">Info</a>		
Platform <a href="#">Linux/UNIX (Inferred)</a>	AMI ID <a href="#">ami-0f06d16b26c594254</a>	Monitoring disabled
Platform details <a href="#">Linux/UNIX</a>	AMI name <a href="#">emr-7_1_0-x86_64-2023_4_20240416_0-Hadoop_Hive_Spark-2024-04-24T14-20-59.949Z</a>	Termination protection Enabled
Stop protection Disabled	Launch time <a href="#">Sat Apr 27 2024 13:43:53 GMT-0400 (Eastern Daylight Time) (10 minutes)</a>	AMI location <a href="#">amazon/emr-7_1_0-x86_64-2023_4_20240416_0-Hadoop_Hive_Spark-2024-04-24T14-20-59.949Z</a>
Instance auto-recovery Default	Lifecycle normal	Stop-hibernate behavior Disabled
AMI Launch index 0	Key pair assigned at launch <a href="#">programming-2</a>	State transition reason -
Credit specification Not supported by instance type	Kernel ID -	State transition message -
Usage operation <a href="#">RunInstances</a>	RAM disk ID -	Owner <a href="#">654654144067</a>

19. Click on the security group to edit its inbound rules

Screenshot of the AWS CloudWatch Metrics console showing the Security tab for a specific metric. The IAM Role is EMR\_EC2\_DefaultRole, Owner ID is 654654144067, and Launch time is Sat Apr 27 2024 13:43:53 GMT-0400 (Eastern Daylight Time). The security group is sg-0d6a41217a7a3a91b (ElasticMapReduce-master).

**Inbound rules:**

Name	Security group rule ID	Port range	Protocol	Source	Security group
-	sgr-0137b10533f9bc30c	22	TCP	0.0.0.0/0	ElasticMapR
-	sgr-02659355d49fc783f	All	ICMP	sg-0d6a41217a7a3a91b	ElasticMapR
-	sgr-05e82fccb0111431b	0 - 65535	TCP	sg-0d6a41217a7a3a91b	ElasticMapR
-	sgr-00d305f7443e21437	0 - 65535	UDP	sg-0d6a41217a7a3a91b	ElasticMapR
-	sgr-04eb4d7dce1514c74	8443	TCP	pl-f8bd5e91	ElasticMapR
-	sgr-04879c57d4c3d5b78	8081	TCP	::/48	ElasticMapR
-	sgr-076ded12faa7a00e4	0 - 65535	UDP	sg-0f3d5c33b4411ee5e	ElasticMapR

20. In the Inbound Rules section, select Edit Inbound Rules.

Screenshot of the AWS VPC Security Groups console showing the details for the security group sg-0d6a41217a7a3a91b (ElasticMapReduce-master). The security group name is ElasticMapReduce-master, owner is 654654144067, and it has 10 inbound rules and 3 outbound rules.

**Inbound rules (10):**

Name	Security group rule ID	Type	Protocol	Port range
-	sgr-0137b10533f9bc30c	SSH	TCP	22
-	sgr-02659355d49fc783f	All ICMP - IPv4	ICMP	All

EC2 > Security Groups > sg-0d6a41217a7a3a91b - ElasticMapReduce-master > Edit inbound rules

### Edit inbound rules Info

Inbound rules control the incoming traffic that's allowed to reach the instance.

Security group rule ID	Type <small>Info</small>	Protocol <small>Info</small>	Port range	Source <small>Info</small>	Description - optional <small>Info</small>
sgr-0137b10533f9bc30c	SSH	TCP	22	Custom	<input type="text" value="Q"/> <input type="text" value="0.0.0/0"/> <input type="button" value="Delete"/>
sgr-02659355d49fc783f	All ICMP - IPv4	ICMP	All	Custom	<input type="text" value="Q"/> <input type="text" value="sg-0d6a41217a7a3a91b"/> <input type="button" value="Delete"/>
sgr-05e82fccb0111431b	All TCP	TCP	0 - 65535	Custom	<input type="text" value="Q"/> <input type="text" value="sg-0d6a41217a7a3a91b"/> <input type="button" value="Delete"/>
sgr-00d305f7443e21437	All UDP	UDP	0 - 65535	Custom	<input type="text" value="Q"/> <input type="text" value="sg-0d6a41217a7a3a91b"/> <input type="button" value="Delete"/>
sgr-04eb4d7dce1514c74	Custom TCP	TCP	8443	Custom	<input type="text" value="Q"/> <input type="text" value="pl-f8bd5e91"/> <input type="button" value="Delete"/>

CloudShell Feedback © 2024, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences 61°F FNG 2:04 PM

21. Enter the port numbers 22 and 4040 with the settings indicated below, then click the Add rule button and save the rules.

sgr-0a8de41c515212a45	All ICMP - IPv4	ICMP	All	Custom	<input type="text" value="Q"/> <input type="text" value="sg-0f3d5c33b4411ee5e"/> <input type="button" value="Delete"/>
-	SSH	TCP	22	My IP	<input type="text" value="Q"/> <input type="text" value="allowing access from my home"/> <input type="text" value="73.112.127.74/32"/> <input type="button" value="Delete"/>
-	Custom TCP	TCP	4040	Custom	<input type="text" value="Q"/> <input type="text" value="0.0.0/16"/> <input type="text" value="for web interface of spark"/> <input type="text" value="0.0.0/16"/> <input type="button" value="Delete"/>
<input type="button" value="Add rule"/>					
⚠ Rules with source of 0.0.0.0/0 or ::/0 allow all IP addresses to access your instance. We recommend setting security group rules to allow access from known IP addresses only. <input type="button" value="X"/>					

The screenshot shows a green notification bar at the top stating: "Inbound security group rules successfully modified on security group (sg-0d6a41217a7a3a91b | ElasticMapReduce-master)". Below this, a breadcrumb navigation shows "EC2 > Security Groups > sg-0d6a41217a7a3a91b - ElasticMapReduce-master". The main title is "sg-0d6a41217a7a3a91b - ElasticMapReduce-master". On the right, there is an "Actions" dropdown menu. The "Details" tab is selected, displaying the following information:

Security group name ElasticMapReduce-master	Security group ID sg-0d6a41217a7a3a91b	Description Master group for Elastic MapReduce created on 2024-04-19T19:57:14.911Z	VPC ID vpc-0512e1100d889881a
Owner 654654144067	Inbound rules count 12 Permission entries	Outbound rules count 3 Permission entries	

Below the table, there are three tabs: "Inbound rules" (selected), "Outbound rules", and "Tags".

22. Create an S3 bucket in AWS services to store the dataset.

23. Choose "Create Bucket".

The screenshot shows the Amazon S3 landing page with the heading "Amazon S3" and the subtext "Store and retrieve any amount of data from anywhere". Below this, a description states: "Amazon S3 is an object storage service that offers industry-leading scalability, data availability, security, and performance." To the right, a "Create a bucket" section contains the text: "Every object in S3 is stored in a bucket. To upload files and folders to S3, you'll need to create a bucket where the objects will be stored." A prominent orange "Create bucket" button is located at the bottom of this section.

24. Label your bucket as "dataset-programming-assignment-2". Scroll down and click on the "Create bucket" button.

The screenshot shows the AWS S3 'Create bucket' wizard. The 'General configuration' tab is selected, displaying fields for AWS Region (US East (N. Virginia) us-east-1), Bucket type (set to 'General purpose'), Bucket name (dataset-programming-assignment-2), and Object Ownership. The 'Default encryption' tab is also visible, showing options for Encryption type (Server-side encryption with Amazon S3 managed keys (SSE-S3) is selected), Bucket Key (Bucket Keys are disabled), and Advanced settings. A note at the bottom indicates that after creating the bucket, files can be uploaded and additional settings configured.

Amazon S3 > Buckets > Create bucket

## Create bucket Info

Buckets are containers for data stored in S3.

### General configuration

AWS Region  
US East (N. Virginia) us-east-1

Bucket type Info

General purpose  
Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.

Directory - New  
Recommended for low-latency use cases. These buckets use only the S3 Express One Zone storage class, which provides faster processing of data within a single Availability Zone.

Bucket name Info  
dataset-programming-assignment-2

Bucket name must be unique within the global namespace and follow the bucket naming rules. See [rules for bucket naming](#).

Copy settings from existing bucket - optional  
Only the bucket settings in the following configuration are copied.

Choose bucket

Format: s3://bucket/prefix

### Object Ownership Info

Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.

Add tag

### Default encryption Info

Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption type Info

Server-side encryption with Amazon S3 managed keys (SSE-S3)

Server-side encryption with AWS Key Management Service keys (SSE-KMS)

Dual-layer server-side encryption with AWS Key Management Service keys (DSSE-KMS)  
Secure your objects with two separate layers of encryption. For details on pricing, see [DSSE-KMS pricing](#) on the [Storage](#) tab of the [Amazon S3 pricing page](#).

Bucket Key  
Using an S3 Bucket Key for SSE-KMS reduces encryption costs by lowering calls to AWS KMS. S3 Bucket Keys aren't supported for DSSE-KMS. [Learn more](#)

Disable

Enable

### Advanced settings

After creating the bucket, you can upload files and folders to the bucket, and configure additional bucket settings.

Cancel Create bucket

25. Access the buckets page to view your newly created bucket.
26. Click on the bucket name you created.

Total storage: 6.0 MB Object count: 619 Average object size: 10.0 KB You can enable advanced metrics in the "default-account-dashboard" configuration.

General purpose buckets | Directory buckets

General purpose buckets (2) [Info](#) All AWS Regions

Buckets are containers for data stored in S3.

Find buckets by name

Name	AWS Region	IAM Access Analyzer	Creation date
aws-logs-654654144067-us-east-1	US East (N. Virginia) us-east-1	<a href="#">View analyzer for us-east-1</a>	April 27, 2024, 13:43:46 (UTC-04:00)
dataset-programming-assignment-2	US East (N. Virginia) us-east-1	<a href="#">View analyzer for us-east-1</a>	April 27, 2024, 14:23:38 (UTC-04:00)

27. Click the upload button.

Amazon S3 > Buckets > dataset-programming-assignment-2

dataset-programming-assignment-2 [Info](#)

Objects | Properties | Permissions | Metrics | Management | Access Points

Objects (0) [Info](#)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Find objects by prefix

Name	Type	Last modified	Size	Storage class
No objects You don't have any objects in this bucket.				

Upload

28. Select the.csv files from your local system by clicking on "Add Files." Now, click the "Upload button" to place the dataset in the S3 bucket. You now have two.csv files in your S3 bucket: ValidationDataset.csv and TrainingDataset.csv.

☰

**Upload succeeded**  
View details below.

**Upload: status**

The information below will no longer be available after you navigate away from this page.

**Summary**

Destination s3://dataset-programming-assignment-2	Succeeded 2 files, 75.7 KB (100.00%)	Failed 0 files, 0 B (0%)
--	---	-----------------------------

**Files and folders** Configuration

**Files and folders (2 Total, 75.7 KB)**

Name	Folder	Type	Size	Status	Error
ValidationD...	-	text/csv	8.6 KB	<span style="color: green;">Succeeded</span>	-
TrainingDat...	-	text/csv	67.2 KB	<span style="color: green;">Succeeded</span>	-

**Close**

29. Connect to the server using PuTTY by specifying the PPK file for authentication.

[EC2](#) > [Instances](#) > i-0f7ca45c3a5f4dd63

**Instance summary for i-0f7ca45c3a5f4dd63** [Info](#)

Updated less than a minute ago

**Actions**

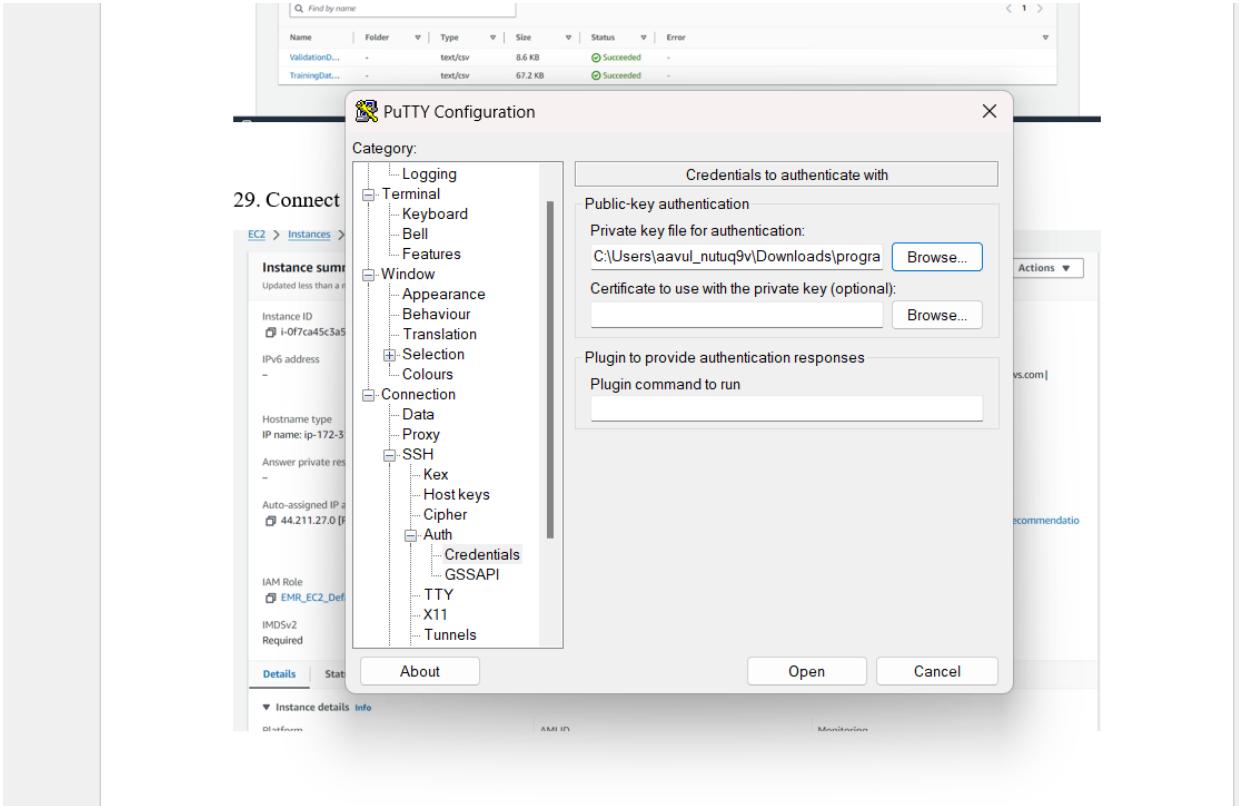
Instance ID <a href="#">i-0f7ca45c3a5f4dd63</a>	<b>PuTTY Configuration</b>	Private IPv4 addresses <a href="#">172.31.76.1</a>
IPv6 address -	Category: Session	Public IPv4 DNS <a href="#">ec2-44-211-27-0.compute-1.amazonaws.com</a>   <a href="#">open address</a>
Hostname type IP name: ip-172-31-76-1.ec2.internal	Specify the destination you want to connect to Host Name (or IP address) Port 44.211270	Elastic IP addresses -
Answer private resource DNS name -	Connection type: <input checked="" type="radio"/> SSH <input type="radio"/> Serial <input type="radio"/> Other Telnet	AWS Compute Optimizer finding <a href="#">Opt-in to AWS Compute Optimizer for recommendations.</a>
Auto-assigned IP address <a href="#">44.211.27.0 [Public IP]</a>	Load, save or delete a stored session Saved Sessions Default Settings emergency server server2	<a href="#">Learn more</a>
IAM Role <a href="#">EMR_EC2_DefaultRole</a>	About	Auto Scaling Group name -
IMDSv2 Required	<a href="#">Open</a> <a href="#">Cancel</a>	

**Details** Status and alarms [New](#) Monitoring Security Networking Storage Tags

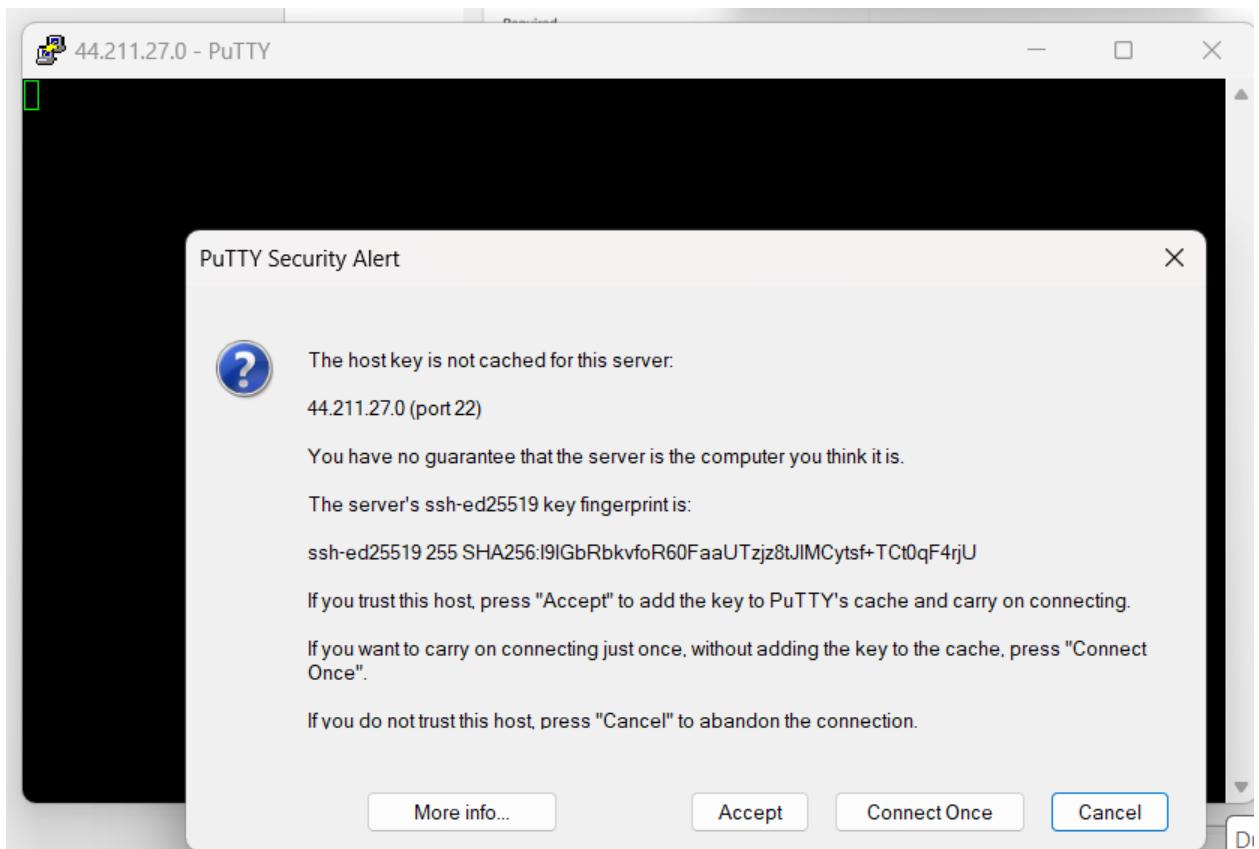
▼ Instance details [Info](#)

Platform AMI ID Monitoring

30. Click on "SSH" under "Auth", select "Credentials", provide the path to the location of the PPK file, and then click "Open".



31.click on accept



32. provide ec2-user

```
ec2-user@ip-172-31-76-1:~$ login as: ec2-user
Authenticating with public key "programming-2"
      _#
     ~\ _###_
    ~~ \_#####
   ~~ \###|
  ~~ \#/ __->
  ~~ V~' /-
  ~~~ /
  ~~~ / \
  ~~~ / \
  ~m' / \
[ec2-user@ip-172-31-76-1 ~]$
```

33. To set up credentials, configure them in the Master node EC2 instance.  
To configure credentials for the master node, type the following instructions in your terminal:

```
# mkdir .aws  
# touch .aws/credentials  
# vi .aws/credentials
```

34. Copy and paste the credentials from the AWS Academy page and AWS information.

**AWS CLI:**  
Copy and paste the following into  
.aws/credentials

```
[default]
aws_access_key_id=ASIAZQ3DNPZB6BB4OR3V
aws_secret_access_key=tBakn87qp0aG3DEZuQbwA
ztYyqE+nsxSKNg05CmM
aws_session_token=IQoJb3JpZ2l1uX2VjEJ
D//////////wEaCXVzLXd1c3QtMiJHMEUCIQD1LrDt1
2Pe7X+yq4qocuPTKZrk99LhhjirjsYdc2KEcQIgYR+8
jKUnJ5LKUYHJmRI8hsX10hbhs0ij8IMIw8bvo78qsQI
I2f//////////ARAAAGgw2NTQ2NTQxNDQwNjcidiDJi0Dc
EsoZPX4UUUiRyqFAjDLpepf0M1tzKH+2g6My29KtGSPY
Hfm1Pmd091c00XbhQqyH4tt0xdJchxjf9qjDaYKji+r
Ts5pIX4IHGXGygXHDt3wG6H91dafwkoZsoTj7L1XcDWG
n4FERPjmuZnsZcWT2JMu+eLpMkhgDOs+TKdpey/1or6
PzrykOIBFCbBBEdbCXzga6JJkw/biKTqDkJsx0xc54S
Zr0Cy6KGRcv1pq7mm6qimMZryT0uUMJX9n+roTDn/81
RLhLwaexFsQqX01WEh10KMLPOB8R1mk0TYt316S3GT/
Y+X/mM5aT/Tjg5H/n20+uGoEiW0nGrgBzMvzqypu8PF
1DSJ6k6RDoc2SxYAPfuDCDv7SxBjqdAbIUCws9is91W
wuV5hbII7xXjr14/wgOEc9X+R4G8WWtZ7pPQmKqc5FL
DkXA9qCnp0R1DXcIi04EBwB4cucgzXyFRXfYaUfLGRn
eIpmpfTzX69hYLZKdYX3K5TLFYvznpEODVaMrw4ayhDQ
3xLF2qMeZ+A9OLGOzUBj+U67MsLDbWSKMzQod/L1bmC
```

35.To execute our Spark application, we must first install the required packages.

Run the commands below:

```
# sudo yum update.
# sudo yum install git.
# pip install pyspark findspark boto3 numpy pandas scikit-learn datetime
```

36.To clone the GitHub repository, enter the following command:

```
#git clone https://github.com/gayatriaavula/winepred_quality.git
```

37.To get started with the Spark Application, perform the following commands:

```
# spark-submit --master yarn CS643_Programming_assignment_2/WineTraining.py
# spark-submit --master yarn CS643_Programming_assignment_2/WineTesting.py >
output.txt
```

```
[hadoop@ip-172-31-69-133 ~]$ spark-submit --master yarn CS643_Programming_assignment_2/WineTraining.py
Apr 26, 2024 8:56:28 PM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
24/04/26 20:56:32 INFO SparkContext: Running Spark version 3.5.0-amzn-1
24/04/26 20:56:32 INFO SparkContext: OS info Linux, 6.1.84-99.169.amzn2023.x86_64, amd64
24/04/26 20:56:32 INFO SparkContext: Java version 17.0.10
24/04/26 20:56:32 INFO ResourceUtils: =====
24/04/26 20:56:32 INFO ResourceUtils: No custom resources configured for spark.driver.
24/04/26 20:56:32 INFO ResourceUtils: =====
24/04/26 20:56:32 INFO SparkConf: Submitted application: WineQuality_Training
24/04/26 20:56:32 INFO ResourceProfile: DefaultResourceProfile created, executor resources: Map(cores -> name: cores, amount: 4, script: , vendor: , memory -> name: memory, amount: 9486, s
cript: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ) task resources: Map(cpu -> name: cpus, amount: 1.0)
24/04/26 20:56:32 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/04/26 20:56:32 INFO SecurityManager: Changing view acls to: hadoop
24/04/26 20:56:32 INFO SecurityManager: Changing modify acls to: hadoop
24/04/26 20:56:32 INFO SecurityManager: Changing view acls groups to:
24/04/26 20:56:32 INFO SecurityManager: Changing modify acls groups to:
24/04/26 20:56:32 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: hadoop; groups with view permissions: EMPTY; users with modi
ty permissions: hadoop; groups with modify permissions: EMPTY
24/04/26 20:56:32 INFO Utils: Successfully started service 'sparkDriver' on port 40123.
24/04/26 20:56:32 INFO SparkEnv: Registering OutputTracker
24/04/26 20:56:32 INFO SparkEnv: Registering BlockManagerMaster
24/04/26 20:56:32 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/04/26 20:56:32 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/04/26 20:56:32 INFO BlockManager: Registered local directory at /tmp/blockmgrr-19a6d5b6-a51f-4d1b-b115-eb9b8f2239af
24/04/26 20:56:32 INFO MemoryStore: MemoryStore started with capacity 1048.6 Mib
24/04/26 20:56:32 INFO SparkEnv: Registering OutputCommitCoordinator
24/04/26 20:56:32 INFO SubResultCacheManager: Sub-result caches are disabled.
24/04/26 20:56:32 INFO JettyUtils: Start Jetty 0.0.0.0:4040 for SparkUI
24/04/26 20:56:33 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/26 20:56:33 INFO Utils: Using 50 preallocated executors [minExecutors: 0]. Set spark.dynamicAllocation.preallocateExecutors to 'false' disable executor preallocation.
24/04/26 20:56:33 INFO DefaultNoAMFalloverProxyProvider: Connecting to ResourceManager at ip-172-31-69-133.ec2.internal/172.31.69.133:8032
24/04/26 20:56:33 INFO Configuration: resource-types.xml not found
24/04/26 20:56:33 INFO ResourceUtils: Unable to find 'resource-types.xml'.
24/04/26 20:56:33 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
24/04/26 20:56:33 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
24/04/26 20:56:33 INFO Client: Setting up container launch context for our AM
24/04/26 20:56:33 INFO Client: Setting up the launch environment for our AM container
24/04/26 20:56:33 WARN Client: Neither spark.yarn.jars nor spark.archive is set, falling back to uploading libraries under SPARK HOME.
24/04/26 20:56:34 INFO Client: Uploading resource file:/mnt/tmp/spark-4db2f66a-0ddf-4955-bde0-c52a65fe4c83/_spark_libs_1278197152423226073.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020
/usr/hadoop/.sparkStaging/application_1714161677101_0003/_spark_libs_1278197152423226073.zip
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/spark/conf.dist/hive-site.xml -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_00
hive-site.xml
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/hudi/conf.dist/hudi-defaults.conf -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_0003/hudi-defaults.conf
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_0003/pyspark.zip
24/04/26 20:56:35 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_0003/py4j-0.10.9.7-src.zip
```

```
24/04/26 20:56:33 INFO Utils: Successfully started service 'SparkUI' on port 4040.
24/04/26 20:56:33 INFO Utils: Using 50 preallocated executors [minExecutors: 0]. Set spark.dynamicAllocation.preallocateExecutors to 'false' disable executor preallocation.
24/04/26 20:56:33 INFO DefaultNoAMFalloverProxyProvider: Connecting to ResourceManager at ip-172-31-69-133.ec2.internal/172.31.69.133:8032
24/04/26 20:56:33 INFO Configuration: resource-types.xml not found
24/04/26 20:56:33 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (12288 MB per container)
24/04/26 20:56:33 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
24/04/26 20:56:33 INFO Client: Setting up container launch context for our AM
24/04/26 20:56:33 INFO Client: Setting up the launch environment for our AM container
24/04/26 20:56:33 INFO Client: Preparing resources for our AM container
24/04/26 20:56:33 WARN Client: Neither spark.yarn.jars nor spark.archive is set, falling back to uploading libraries under SPARK HOME.
24/04/26 20:56:34 INFO Client: Uploading resource file:/mnt/tmp/spark-4db2f66a-0ddf-4955-bde0-c52a65fe4c83/_spark_libs_1278197152423226073.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020
/usr/hadoop/.sparkStaging/application_1714161677101_0003/_spark_libs_1278197152423226073.zip
24/04/26 20:56:35 INFO Client: Uploading resource file:/etc/spark/conf.dist/hive-site.xml -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_00
hive-site.xml
24/04/26 20:56:35 INFO Client: Upgrading resource file:/etc/hudi/conf.dist/hudi-defaults.conf -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_0003/hudi-defaults.conf
24/04/26 20:56:35 INFO Client: Upgrading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_0003/pyspark.zip
24/04/26 20:56:35 INFO Client: Upgrading resource file:/usr/lib/spark/python/lib/py4j-0.10.9.7-src.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020/user/hadoop/.sparkStaging/application_1714161677101_0003/py4j-0.10.9.7-src.zip
24/04/26 20:56:36 INFO Client: Upgrading resource file:/mnt/tmp/spark-4db2f66a-0ddf-4955-bde0-c52a65fe4c83/_spark_conf_104927068038660577.zip -> hdfs://ip-172-31-69-133.ec2.internal:8020
/usr/hadoop/.sparkStaging/application_1714161677101_0003/_spark_conf_.zip
24/04/26 20:56:36 INFO SecurityManager: Changing view acls to: hadoop
24/04/26 20:56:36 INFO SecurityManager: Changing modify acls to: hadoop
24/04/26 20:56:36 INFO SecurityManager: Changing view acls groups to:
24/04/26 20:56:36 INFO SecurityManager: Changing modify acls groups to:
24/04/26 20:56:36 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: hadoop; groups with view permissions: EMPTY; users with modi
ty permissions: hadoop; groups with modify permissions: EMPTY
24/04/26 20:56:37 INFO Client: Submitting application application_1714161677101_0003 to ResourceManager
24/04/26 20:56:37 INFO YarnClientImpl: Submitted application application_1714161677101_0003
24/04/26 20:56:37 INFO Client: Application report for application_1714161677101_0003 (state: ACCEPTED)
24/04/26 20:56:37 INFO Client:
  client token: N/A
  diagnostics: N/A
  ApplicationMaster host: N/A
  ApplicationMaster RPC port: -1
  queue: default
  start time: 1714164996129
  final status: UNDEFINED
  tracking URL: http://ip-172-31-69-133.ec2.internal:20888/proxy/application_1714161677101_0003/
  user: hadoop
24/04/26 20:56:40 INFO YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY_HOSTS -> ip-172-31-69-133.ec2.internal, PROXY_URI_
BASES -> http://ip-172-31-69-133.ec2.internal:20888/proxy/application_1714161677101_0003), /proxy/application_1714161677101_0003
24/04/26 20:56:41 INFO YarnSchedulerBackend: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
24/04/26 20:56:41 INFO Client: Application report for application_1714161677101_0003 (state: RUNNING)
24/04/26 20:56:41 INFO Client:
  client token: N/A
  diagnostics: N/A
  ApplicationMaster host: 172.31.79.179
  ApplicationMaster RPC port: -1
  queue: default
  start time: 1714164996129
  final status: UNDEFINED
  tracking URL: http://ip-172-31-69-133.ec2.internal:20888/proxy/application_1714161677101_0003/
```

```

user: hadoop
24/04/26 20:56:41 INFO YarnClientSchedulerBackend: Application application_1714161677101_0003 has started running.
24/04/26 20:56:41 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 40265.
24/04/26 20:56:41 INFO NettyBlockTransferService: Server created on ip-172-31-69-133.ec2.internal:40265
24/04/26 20:56:41 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/04/26 20:56:41 INFO BlockManager: external shuffle service port = 7337
24/04/26 20:56:41 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-69-133.ec2.internal:40265 with 1048.8 MB RAM, BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO BlockManager: Registered BlockManager BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-69-133.ec2.internal, 40265, None)
24/04/26 20:56:41 INFO SingleWeventLogFileWriter: Logging events to hdfs://var/log/spark/apps/application_1714161677101_0003.inprogress
24/04/26 20:56:41 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to 'false' disable executor preallocation.
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/job/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /jobs/job/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/stage/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/stage/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/pool/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/pool/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/rdd/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /storage/rdd/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /environment/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/threadDump: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/threadDump/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/heapHistogram: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /executors/heapHistogram/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /static: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /api/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /api/v1/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /stages/stage/kill/: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO ServerInfo: Adding filter to /metrics/json: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter
24/04/26 20:56:41 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
Importing: s3a://dataset-programming-assignment-2/TrainingDataset.csv
>>> Model Path set: 3a1/dataset-programming-assignment-2/models
SLF4J: Failed to load class "org.slf4j.impl.StaticLoggerBinder".
SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
root
|--- ****"fixed acidity": string (nullable = true)
|--- ****"volatile acidity": string (nullable = true)
|--- ****"citric acid": string (nullable = true)
|--- ****"residual sugar": string (nullable = true)
|--- ****"chlorides": string (nullable = true)
|--- ****"free sulfur dioxide": string (nullable = true)
|--- ****"total sulfur dioxide": string (nullable = true)

```

```

SLF4J: Defaulting to no-operation (NOP) logger implementation
SLF4J: See http://www.slf4j.org/codes.html#StaticLoggerBinder for further details.
root
|--- ****"fixed acidity": string (nullable = true)
|--- ****"volatile acidity": string (nullable = true)
|--- ****"citric acid": string (nullable = true)
|--- ****"residual sugar": string (nullable = true)
|--- ****"chlorides": string (nullable = true)
|--- ****"free sulfur dioxide": string (nullable = true)
|--- ****"total sulfur dioxide": string (nullable = true)
|--- ****"density": string (nullable = true)
|--- ****"pH": string (nullable = true)
|--- ****"sulphates": string (nullable = true)
|--- ****"alcohol": string (nullable = true)
|--- ****"quality": string (nullable = true)

+-----+-----+-----+-----+-----+-----+
|****"fixed acidity"****"volatile acidity"****"citric acid"****"residual sugar"****"chlorides"****"free sulfur dioxide"****"total sulfur dioxide"****"density"****"pH"
|****"sulphates"****"alcohol"****"quality"****
+-----+-----+-----+-----+-----+-----+
|       8.9|      0.22|     0.48|      1.8|     0.077|      29|      60|    0.9968| |
|       5.3|      9.4|      6|      0.31|     2.3|     0.082|      23|      71|    0.9982|
|       7.6|      0.39|     5|      0.31|     2.3|     0.082|      23|      71|    0.9982|
|       0.69|      9.7|      5|      0.21|     1.6|     0.106|      10|      37|    0.9966|
|       0.81|      9.5|      5|      0.21|     2.3|     0.084|      9|      67|    0.9968|
|       8.5|      0.49|     5|      0.11|     2.3|     0.084|      9|      67|    0.9968|
|       0.51|      9.4|      5|      0.11|     2.3|     0.084|      9|      67|    0.9968|
|       6.9|      0.41|     5|      0.14|     2.4|     0.085|      21|      40|    0.9968|
|       0.63|      9.7|      6|      0.14|     2.4|     0.085|      21|      40|    0.9968|
|       6.3|      0.39|     6|      0.16|     1.4|      0.08|      11|      23|    0.9955|
|       0.56|      9.3|      5|      0.24|     1.8|      0.08|      4|      11|    0.9962|
|       7.6|      0.41|     5|      0.21|     1.6|     0.106|      10|      37|    0.9966|
|       0.59|      9.5|      5|      0.21|     1.6|     0.106|      10|      37|    0.9966|
|       7.9|      0.43|     5|      0.21|     1.6|     0.106|      10|      37|    0.9966|
|       0.91|      9.5|      5|      0.21|     1.6|     0.106|      10|      37|    0.9966|
|       7.1|      0.71|     0|      0|     1.5|      0.08|      14|      35|    0.9972|
|       0.55|      9.4|      5|      0.1|     1.5|      0.08|      14|      35|    0.9972|
|       7.8|      0.645|     6|      0|     2|      0.082|      8|      16|    0.9964|
|       0.55|      9.8|      6|      0|     2|      0.082|      8|      16|    0.9964|
|       6.7|      0.675|     6|      0.07|     2.4|      0.089|      17|      82|    0.9958|
|       0.54|     10.1|      5|      0|     2.5|      0.105|      22|      37|    0.9966|
|       6.9|      0.685|     5|      0|     2.5|      0.105|      22|      37|    0.9966|
|       0.57|     10.6|      6|      0|     2.3|      0.083|      15|      113|    0.9966|
|       8.3|      0.655|     6|      0.12|     2.3|      0.083|      15|      113|    0.9966|
|       0.66|      9.8|      5|      0.12|     10.7|      0.073|      40|      83|    0.9993|
|       6.9|      0.605|     5|      0.12|     10.7|      0.073|      40|      83|    0.9993|
|       0.52|      9.4|      6|      0.25|     1.8|     0.103|      13|      50|    0.9957|
|       5.2|      0.32|     6|      0.25|     1.8|     0.103|      13|      50|    0.9957|
|       0.55|      9.2|      5|      0|     5.5|      0.086|      5|      18|    0.9986|
|       7.8|      0.645|     6|      0|     5.5|      0.086|      5|      18|    0.9986|
|       0.55|      9.6|      6|      0.14|     2.4|      0.086|      3|      15|    0.9975|
|       7.8|      0.6|      6|      0.14|     2.4|      0.086|      3|      15|    0.9975|

```

The results will be displayed here, along with the Accuracy and F1 scores of the Machine Learning methods used.

```

Training DataSet Metrics
Accuracy: 0.6114151681000782
F-measure: 0.5904050519731796
+-----+
|fixed_acidity|volatile_acidity|citric_acid|residual_sugar|chlorides|free_sulfur_dioxide|total_sulfur_dioxide|density| pH|sulphates|alcohol|label|
+-----+
|    7.4|      0.7|      0.0|      1.9|     0.076|           11|          34| 0.9978|3.51|   0.56|   9.4| 5.0|
|    7.8|      0.88|      0.0|      2.6|     0.098|           25|          67| 0.9968|3.2|   0.68|   9.8| 5.0|
|    7.8|      0.76|      0.04|      2.3|     0.092|           15|          54| 0.9971|3.26|   0.65|   9.8| 5.0|
|   11.2|      0.28|      0.56|      1.9|     0.075|           17|          60| 0.9981|3.16|   0.58|   9.8| 6.0|
|    7.4|      0.7|      0.0|      1.9|     0.076|           11|          34| 0.9978|3.51|   0.56|   9.4| 5.0|
+-----+
only showing top 5 rows

Validation Training Set Metrics
+-----+
|features|label|prediction|
+-----+
|[9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]| 15.0 | 15.0 |
|[9.8,0.68,3.2,0.9968,25.0,67.0,0.0,0.098,2.6,0.0,0.88,7.8]| 15.0 | 15.0 |
|[9.8,0.65,3.26,0.997,15.0,54.0,0.0092,2.3,0.04,0.76,7.8]| 15.0 | 15.0 |
|[9.8,0.58,3.16,0.998,17.0,60.0,0.075,1.9,0.56,0.28,11.2]| 16.0 | 15.0 |
|[9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4]| 15.0 | 15.0 |
+-----+
only showing top 5 rows

The accuracy of the model is 0.575
F1: 0.5619407071339173

```

## DOCKER IMPLEMENTATION –

1.Update System Packages:

```
sudo yum update -y
```

2.Install Docker:

```
sudo yum install -y docker
```

3.Start Docker Service:

```
sudo service docker start
```

4.Check Docker Service Status:

```
sudo service docker status
```

```

aws Services Search [Alt+S] N. Virginia vocabs/user3110677=ga348@njit.edu @ 6546-5414-4067
last metadata expiration check: 0:07:39 ago on Sat Apr 27 20:00:16 2024.
Package docker-25.0.3-1.amzn2023.0.1.x86_64 is already installed.
Dependencies resolved.
Nothing to do.
Complete!
[root@ip-172-31-68-235 ~]# sudo service docker start
Redirecting to /bin/systemctl start docker.service
[root@ip-172-31-68-235 ~]# sudo service docker status
Redirecting to /bin/systemctl status docker.service
● docker.service - Docker Application Container Engine
   Loaded: loaded (/usr/lib/systemd/system/docker.service; disabled; preset: disabled)
     Active: active (running) since Sat 2024-04-27 20:08:20 UTC; 9s ago
    TriggeredBy: • docker.socket
      Docs: https://docs.docker.com
   Process: 19905 ExecStartPre=/bin/mkdir -p /run/docker (code-exited, status=0/SUCCESS)
   Process: 19907 ExecStartPre=/usr/libexec/docker/docker-setup-runtimes.sh (code-exited, status=0/SUCCESS)
 Main PID: 19908 (dockerd)
   Tasks: 10
    Memory: 109.6M
      CPU: 307ms
     CGroup: /system.slice/docker.service
             └─19908 /usr/bin/dockerd -H fd:// --containerd=/run/containerd/containerd.sock --default-ulimit nofile=32768:65536

apr 27 20:08:19 ip-172-31-68-235.ec2.internal dockerd[19908]: time="2024-04-27T20:08:20.299703799Z" level=info msg="Starting up"
apr 27 20:08:20 ip-172-31-68-235.ec2.internal dockerd[19908]: time="2024-04-27T20:08:20.455851200Z" level=info msg="Loading containers: start."
apr 27 20:08:20 ip-172-31-68-235.ec2.internal dockerd[19908]: time="2024-04-27T20:08:20.804997187Z" level=info msg="Loading containers: done."
apr 27 20:08:20 ip-172-31-68-235.ec2.internal dockerd[19908]: time="2024-04-27T20:08:20.834877311Z" level=info msg="Docker daemon" commit=f417435 containerd-snapshotter=f
apr 27 20:08:20 ip-172-31-68-235.ec2.internal dockerd[19908]: time="2024-04-27T20:08:20.834998692Z" level=info msg="daemon has completed initialization"
apr 27 20:08:20 ip-172-31-68-235.ec2.internal dockerd[19908]: time="2024-04-27T20:08:20.873578391Z" level=info msg="API listen on /run/docker.sock"
apr 27 20:08:20 ip-172-31-68-235.ec2.internal systemd[1]: Started docker.service - Docker Application Container Engine.

```

## 5. Create a Dockerfile and build an image with the docker build command.

```
sudo docker build -t ga348/cs643-programming-assignment-2 .
```

## 6. To check if a Docker image was built, use the following command:

```
# sudo docker image ls
```

```
[root@ip-172-31-68-235 CS643_Programming_assignment_2]# docker image ls
REPOSITORY          TAG      IMAGE ID   CREATED        SIZE
ga348/cs643-programming-assignment-2 latest   7fc54dac10b9  19 seconds ago  2.42GB
[root@ip-172-31-68-235 CS643_Programming_assignment_2]#
```

You can see here that your docker image has been built.

## 7. To run the docker image, use the following command:

```
sudo docker run -it ga348/cs643-programming-assignment-2
```

Instead of using the image name, you can use the image ID: # sudo docker run -it

## 8. This results in the same return for Accuracy and F1 scores.

```

20/11/29 23:17:05 WARN NativeCodeLoader: Unable to load native Hadoop library for your platform... using built-in Java classes.
TestingDataSet Metrics

+-----+-----+
|features          |label|prediction|
+-----+-----+
|[9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4] |15.0 |15.0 |
|[9.8,0.68,3.2,0.9968,25.0,67.0,0.098,2.6,0.0,0.88,7.8] |15.0 |15.0 |
|[9.8,0.65,3.26,0.997,15.0,54.0,0.092,2.3,0.04,0.76,7.8] |15.0 |15.0 |
|[9.8,0.58,3.16,0.998,17.0,60.0,0.075,1.9,0.56,0.28,11.2] |16.0 |15.0 |
|[9.4,0.56,3.51,0.9978,11.0,34.0,0.076,1.9,0.0,0.7,7.4] |15.0 |15.0 |
+-----+-----+
only showing top 5 rows

The accuracy of the model is 0.6271186440677966
F1: 0.593151718932272

```

9. Run the following command to submit the produced Docker image to the DockerHub repository:

```
sudo docker push ga348/cs643-programming-assignment-2
```

```

Username: ga348
Password:
WARNING! Your password will be stored unencrypted in /root/.docker/config.json.
Configure a credential helper to remove this warning. See
https://docs.docker.com/engine/reference/commandline/login/#credentials-store

Login Succeeded
[root@ip-172-31-68-235 CS643_Programming_assignment_2]# sudo docker push ga348/cs643-programming-assignment-2
Using default tag: latest
The push refers to repository [docker.io/ga348/cs643-programming-assignment-2]
a30860ff413b: Pushed
f93ab35fe624: Pushed
119e1277745c: Pushed
bce9c947b0d3: Pushed
6026c907f1fd: Pushed
57c651240c9f: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
553c43e260d1: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
36ef902c4c66: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
ea1b8bc1ff8: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
632ccc24d10f: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
8933d669b084: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
367158596a5c: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
c9ac6abbc04d: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
41caa71c39b5: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
97393f8c8163: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
d7802b8508af: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
e3abdc2e9252: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
eafe6e032dbd: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
92a4e8a3140f: Mounted from guoxiaojun2/spark-3.2.2-bin-hadoop3.2
latest: digest: sha256:1e13e8341a051f20a4896f253dd7ac488f61b09d234abfbe4890f178ba486125 size: 4516

```

10. Download and execute the Docker image from the DockerHub repository, following the instructions provided on the website.

## Git Bash:

the steps for moving code from my local system to a GitHub repository using Git Bash.

1.git init

2.git add.

### 3.git status

4.git commit -m " updated code"

5.git remote add origin https://github.com/gayatriaavula/winepred\_quality.git

6. git push -u origin main

```
aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud
$ git init
initialized empty Git repository in c:/Users/aavul_nutuq9v/OneDrive/Documents/DevOps/cloud/.git/
aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud (master)
$ git status
On branch master
No commits yet
Untracked files:
  (use "git add <file>..." to include in what will be committed)
    1.png
    2.png
    3.png
    4.png
    CS643_homeworkset4#_Gayatri_Aavula.docx
    Internship offer letter.docx
    aws-keys.txt
    aws-programming1-notes.txt
    certificate_226002383_9fe5093e.pdf
    images.docx
    mfa.txt
    object1.pem
    object1.ppk
    presentation_cloud.docx
    programming2/
    recording-images.docx
    winepred_quality/
```

```
aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git status
On branch main
Your branch is up to date with 'origin/main'.
Untracked files:
  (use "git add <file>..." to include in what will be committed)
    Dockerfile
    Trainingdataset.csv
    Validationdataset.csv
    WineTesting.py
    WineTraining.py
nothing added to commit but untracked files present (use "git add" to track)
aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git add .
aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git status
On branch main
Your branch is up to date with 'origin/main'.
Changes to be committed:
  (use "git restore --staged <file>..." to unstage)
    new file:   Dockerfile
    new file:   Trainingdataset.csv
    new file:   Validationdataset.csv
    new file:   WineTesting.py
    new file:   WineTraining.py
aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git commit -m "updated code"
git: 'git' is not a git command. See 'git --help'.
The most similar command is
  init
aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git commit -m "updated code"
[main c041357] updated code
 5 files changed, 683 insertions(+)
create mode 100644 Dockerfile
create mode 100644 Trainingdataset.csv
create mode 100644 Validationdataset.csv
create mode 100644 WineTesting.py
create mode 100644 WineTraining.py
aavul_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$
```

```
aavu]_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git remote add origin https://github.com/gayatriaavula/winepred_quality.git
error: remote origin already exists.

aavu]_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ git push -u origin main
Enumerating objects: 8, done.
Counting objects: 100% (8/8), done.
Delta compression using up to 12 threads
Compressing objects: 100% (7/7), done.
Writing objects: 100% (7/7), 24.54 KiB | 2.73 MiB/s, done.
Total 7 (delta 0), reused 0 (delta 0), pack-reused 0
To https://github.com/gayatriaavula/winepred_quality.git
  f8dd3c3..c041357  main -> main
branch 'main' set up to track 'origin/main'.

aavu]_nutuq9v@AAVULADURGA MINGW64 ~/OneDrive/Documents/DevOps/cloud/winepred_quality (main)
$ |
```

