**Assignment-1 [Total Marks - 25]**

**M1: MLOps Foundations**

**Objective**: Understand the basics of MLOps and implement a simple CI/CD pipeline.

**Tasks**:

1. **Set Up a CI/CD Pipeline**:

• Use a CI/CD tool like GitHub Actions or GitLab CI to set up a pipeline for a sample machine learning project.

• Include stages for linting, testing, and deploying a simple machine learning model.

2. **Version Control**:

• Implement version control for your project using Git.

Here are the contributors for the git repository:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

• Demonstrate branching, merging, and pull requests.

Here you can see that there are 3 branches:

1. feature/v0.1
2. main
3. origin/feature/v0.1.1

Team members are contributing to the branches as per their scope of work

A screenshot of a computer

Description automatically generated

For merging and pull request <Gaurav> to raise PR and ask someone to approve.

**Deliverables**:

• A report detailing the CI/CD pipeline stages.

• Screenshots or logs showing successful runs of the pipeline.

• A Git repository link with branches and merge history.

**M2: Process and Tooling**

**Objective**: Gain hands-on experience with popular MLOps tools and understand the processes they support.

**Tasks**:

1. **Experiment Tracking**:

• Use MLflow to track experiments for a machine learning project.

• Record metrics, parameters, and results of at least three different model training runs.

2. **Data Versioning**:

• Use DVC (Data Version Control) to version control a dataset used in your project.

Please see the screen shot for DVC

Following commands were executed:

Step 1: brew install dvc<br>

Step 2: dvc init <br>

Step 3: git status <br>

Step 4: git commit -m "Initialize DVC" <br>

Step 5: dvc get https://github.com/iterative/dataset-registry \

get-started/data.xml -o data/data.xml <br>

Step 6: dvc add data/data.xml <br>

Step 7: git add data/data.xml.dvc data/.gitignore<br>

Step 8: git commit -m "Add raw data"<br>

Step 9: git add data/data.xml.dvc data/.gitignore<br>

Step 10: mkdir /tmp/dvcstore<br>

Step 11: dvc remote add -d myremote /tmp/dvcstore<br>

Step 12: dvc push<br>

Step 13: dvc pull<br>

Step 14: cp data/data.xml /tmp/data.xml<br>

Step 15: cat /tmp/data.xml >> data/data.xml<br>

Step 16: dvc add data/data.xml

Step 17: dvc push

Step 18: git commit data/data.xml.dvc -m "Dataset updates"

Here is the outcome of these commands:

(myenv) gauravshukla@XMRK9TWQW4WJ python % git init

Reinitialized existing Git repository in /Users/gauravshukla/bits/python/.git/

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc init

Initialized DVC repository.

You can now commit the changes to git.

+---------------------------------------------------------------------+

| |

| DVC has enabled anonymous aggregate usage analytics. |

| Read the analytics documentation (and how to opt-out) here: |

| <https://dvc.org/doc/user-guide/analytics> |

| |

+---------------------------------------------------------------------+

What's next?

------------

- Check out the documentation: <https://dvc.org/doc>

- Get help and share ideas: <https://dvc.org/chat>

- Star us on GitHub: <https://github.com/iterative/dvc>

(myenv) gauravshukla@XMRK9TWQW4WJ python % git status

On branch feature/v0.1

Your branch is up to date with 'origin/feature/v0.1'.

Untracked files:

(use "git add <file>..." to include in what will be committed)

.DS\_Store

data/get-started/

mlopsgroup44/

myenv/

~$signment\_Document.docx

nothing added to commit but untracked files present (use "git add" to track)

(myenv) gauravshukla@XMRK9TWQW4WJ python % git commit -m "Initialize DVC"

On branch feature/v0.1

Your branch is up to date with 'origin/feature/v0.1'.

Untracked files:

(use "git add <file>..." to include in what will be committed)

.DS\_Store

data/get-started/

mlopsgroup44/

myenv/

~$signment\_Document.docx

nothing added to commit but untracked files present (use "git add" to track)

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc get https://github.com/iterative/dataset-registry \

get-started/data.xml -o data/data.xml

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc add data/data.xml

100% Adding...|█████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████|1/1 [00:00, 8.03file/s]

To track the changes with git, run:

git add data/.gitignore data/data.xml.dvc

To enable auto staging, run:

dvc config core.autostage true

(myenv) gauravshukla@XMRK9TWQW4WJ python % git add data/data.xml.dvc data/.gitignore

(myenv) gauravshukla@XMRK9TWQW4WJ python % git commit -m "Add raw data"

Auto packing the repository in background for optimum performance.

See "git help gc" for manual housekeeping.

[feature/v0.1 c616cc5] Add raw data

2 files changed, 6 insertions(+)

create mode 100644 data/.gitignore

create mode 100644 data/data.xml.dvc

(myenv) gauravshukla@XMRK9TWQW4WJ python % mkdir /tmp/dvcstore

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc remote add -d myremote /tmp/dvcstore

Setting 'myremote' as a default remote.

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc remote add -d storage s3://mybucket/dvcstore

Setting 'storage' as a default remote.

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc push

Collecting |1.00 [00:00, 530entry/s]

Pushing

ERROR: unexpected error - Unable to locate credentials

Having any troubles? Hit us up at https://dvc.org/support, we are always happy to help!

(myenv) gauravshukla@XMRK9TWQW4WJ python % ls

Assignment\_Document.docx ci-cd-pipeline.yml models requirements.txt

Dockerfile data myenv src

README.md mlopsgroup44 notebooks ~$signment\_Document.docx

(myenv) gauravshukla@XMRK9TWQW4WJ python % ls -all

total 7056

drwxr-xr-x 19 gauravshukla staff 608 Aug 5 12:59 .

drwxr-xr-x 6 gauravshukla staff 192 Aug 5 11:40 ..

-rw-r--r--@ 1 gauravshukla staff 8196 Aug 5 11:17 .DS\_Store

drwxr-xr-x@ 6 gauravshukla staff 192 Aug 5 13:00 .dvc

-rw-r--r--@ 1 gauravshukla staff 139 Aug 5 12:59 .dvcignore

drwxr-xr-x@ 15 gauravshukla staff 480 Aug 5 13:01 .git

drwxr-xr-x@ 3 gauravshukla staff 96 Aug 1 17:28 .github

-rw-r--r--@ 1 gauravshukla staff 3574304 Aug 4 11:26 Assignment\_Document.docx

-rw-r--r--@ 1 gauravshukla staff 494 Aug 3 18:03 Dockerfile

-rw-r--r--@ 1 gauravshukla staff 2568 Aug 4 11:28 README.md

-rw-r--r--@ 1 gauravshukla staff 878 Aug 3 16:56 ci-cd-pipeline.yml

drwxr-xr-x 7 gauravshukla staff 224 Aug 5 13:00 data

drwxr-xr-x@ 5 gauravshukla staff 160 Jul 31 16:37 mlopsgroup44

drwxr-xr-x 2 gauravshukla staff 64 Jul 31 19:53 models

drwxr-xr-x 8 gauravshukla staff 256 Jul 31 18:50 myenv

drwxr-xr-x 2 gauravshukla staff 64 Jul 31 19:53 notebooks

-rw-r--r--@ 1 gauravshukla staff 36 Jul 31 18:42 requirements.txt

drwxr-xr-x 3 gauravshukla staff 96 Jul 31 19:53 src

-rw-r--r--@ 1 gauravshukla staff 162 Aug 4 21:32 ~$signment\_Document.docx

(myenv) gauravshukla@XMRK9TWQW4WJ python % mkdir tmp

(myenv) gauravshukla@XMRK9TWQW4WJ python % ls -all

total 7056

drwxr-xr-x 20 gauravshukla staff 640 Aug 5 14:14 .

drwxr-xr-x 6 gauravshukla staff 192 Aug 5 11:40 ..

-rw-r--r--@ 1 gauravshukla staff 8196 Aug 5 11:17 .DS\_Store

drwxr-xr-x@ 6 gauravshukla staff 192 Aug 5 13:00 .dvc

-rw-r--r--@ 1 gauravshukla staff 139 Aug 5 12:59 .dvcignore

drwxr-xr-x@ 15 gauravshukla staff 480 Aug 5 13:01 .git

drwxr-xr-x@ 3 gauravshukla staff 96 Aug 1 17:28 .github

-rw-r--r--@ 1 gauravshukla staff 3574304 Aug 4 11:26 Assignment\_Document.docx

-rw-r--r--@ 1 gauravshukla staff 494 Aug 3 18:03 Dockerfile

-rw-r--r--@ 1 gauravshukla staff 2568 Aug 4 11:28 README.md

-rw-r--r--@ 1 gauravshukla staff 878 Aug 3 16:56 ci-cd-pipeline.yml

drwxr-xr-x 7 gauravshukla staff 224 Aug 5 13:00 data

drwxr-xr-x@ 5 gauravshukla staff 160 Jul 31 16:37 mlopsgroup44

drwxr-xr-x 2 gauravshukla staff 64 Jul 31 19:53 models

drwxr-xr-x 8 gauravshukla staff 256 Jul 31 18:50 myenv

drwxr-xr-x 2 gauravshukla staff 64 Jul 31 19:53 notebooks

-rw-r--r--@ 1 gauravshukla staff 36 Jul 31 18:42 requirements.txt

drwxr-xr-x 3 gauravshukla staff 96 Jul 31 19:53 src

drwxr-xr-x@ 2 gauravshukla staff 64 Aug 5 14:14 tmp

-rw-r--r--@ 1 gauravshukla staff 162 Aug 4 21:32 ~$signment\_Document.docx

(myenv) gauravshukla@XMRK9TWQW4WJ python % cd tmp

(myenv) gauravshukla@XMRK9TWQW4WJ tmp % mkdir dvcstore

(myenv) gauravshukla@XMRK9TWQW4WJ tmp % ls

dvcstore

(myenv) gauravshukla@XMRK9TWQW4WJ tmp % cd ..

(myenv) gauravshukla@XMRK9TWQW4WJ python % ls -all

total 7056

drwxr-xr-x 20 gauravshukla staff 640 Aug 5 14:14 .

drwxr-xr-x 6 gauravshukla staff 192 Aug 5 11:40 ..

-rw-r--r--@ 1 gauravshukla staff 8196 Aug 5 11:17 .DS\_Store

drwxr-xr-x@ 6 gauravshukla staff 192 Aug 5 13:00 .dvc

-rw-r--r--@ 1 gauravshukla staff 139 Aug 5 12:59 .dvcignore

drwxr-xr-x@ 15 gauravshukla staff 480 Aug 5 13:01 .git

drwxr-xr-x@ 3 gauravshukla staff 96 Aug 1 17:28 .github

-rw-r--r--@ 1 gauravshukla staff 3574304 Aug 4 11:26 Assignment\_Document.docx

-rw-r--r--@ 1 gauravshukla staff 494 Aug 3 18:03 Dockerfile

-rw-r--r--@ 1 gauravshukla staff 2568 Aug 4 11:28 README.md

-rw-r--r--@ 1 gauravshukla staff 878 Aug 3 16:56 ci-cd-pipeline.yml

drwxr-xr-x 7 gauravshukla staff 224 Aug 5 13:00 data

drwxr-xr-x@ 5 gauravshukla staff 160 Jul 31 16:37 mlopsgroup44

drwxr-xr-x 2 gauravshukla staff 64 Jul 31 19:53 models

drwxr-xr-x 8 gauravshukla staff 256 Jul 31 18:50 myenv

drwxr-xr-x 2 gauravshukla staff 64 Jul 31 19:53 notebooks

-rw-r--r--@ 1 gauravshukla staff 36 Jul 31 18:42 requirements.txt

drwxr-xr-x 3 gauravshukla staff 96 Jul 31 19:53 src

drwxr-xr-x@ 3 gauravshukla staff 96 Aug 5 14:15 tmp

-rw-r--r--@ 1 gauravshukla staff 162 Aug 4 21:32 ~$signment\_Document.docx

(myenv) gauravshukla@XMRK9TWQW4WJ python % git add data/data.xml.dvc data/.gitignore

(myenv) gauravshukla@XMRK9TWQW4WJ python % git commit -m "Add raw data"

On branch feature/v0.1

Your branch is ahead of 'origin/feature/v0.1' by 1 commit.

(use "git push" to publish your local commits)

Changes not staged for commit:

(use "git add <file>..." to update what will be committed)

(use "git restore <file>..." to discard changes in working directory)

modified: .dvc/config

Untracked files:

(use "git add <file>..." to include in what will be committed)

.DS\_Store

data/get-started/

mlopsgroup44/

myenv/

~$signment\_Document.docx

no changes added to commit (use "git add" and/or "git commit -a")

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc remote add -d myremote /tmp/dvcstore

Setting 'myremote' as a default remote.

ERROR: configuration error - config file error: remote 'myremote' already exists. Use `-f|--force` to overwrite it.

(myenv) gauravshukla@XMRK9TWQW4WJ python % pwd

/Users/gauravshukla/bits/python

(myenv) gauravshukla@XMRK9TWQW4WJ python % vi .dvc/config

(myenv) gauravshukla@XMRK9TWQW4WJ python % ls -all

total 7056

drwxr-xr-x 20 gauravshukla staff 640 Aug 5 14:14 .

drwxr-xr-x 6 gauravshukla staff 192 Aug 5 11:40 ..

-rw-r--r--@ 1 gauravshukla staff 8196 Aug 5 11:17 .DS\_Store

drwxr-xr-x@ 6 gauravshukla staff 192 Aug 5 14:28 .dvc

-rw-r--r--@ 1 gauravshukla staff 139 Aug 5 12:59 .dvcignore

drwxr-xr-x@ 15 gauravshukla staff 480 Aug 5 14:16 .git

drwxr-xr-x@ 3 gauravshukla staff 96 Aug 1 17:28 .github

-rw-r--r--@ 1 gauravshukla staff 3574304 Aug 4 11:26 Assignment\_Document.docx

-rw-r--r--@ 1 gauravshukla staff 494 Aug 3 18:03 Dockerfile

-rw-r--r--@ 1 gauravshukla staff 2568 Aug 4 11:28 README.md

-rw-r--r--@ 1 gauravshukla staff 878 Aug 3 16:56 ci-cd-pipeline.yml

drwxr-xr-x 7 gauravshukla staff 224 Aug 5 13:00 data

drwxr-xr-x@ 5 gauravshukla staff 160 Jul 31 16:37 mlopsgroup44

drwxr-xr-x 2 gauravshukla staff 64 Jul 31 19:53 models

drwxr-xr-x 8 gauravshukla staff 256 Jul 31 18:50 myenv

drwxr-xr-x 2 gauravshukla staff 64 Jul 31 19:53 notebooks

-rw-r--r--@ 1 gauravshukla staff 36 Jul 31 18:42 requirements.txt

drwxr-xr-x 3 gauravshukla staff 96 Jul 31 19:53 src

drwxr-xr-x@ 3 gauravshukla staff 96 Aug 5 14:15 tmp

-rw-r--r--@ 1 gauravshukla staff 162 Aug 4 21:32 ~$signment\_Document.docx

(myenv) gauravshukla@XMRK9TWQW4WJ python % cd tmp

(myenv) gauravshukla@XMRK9TWQW4WJ tmp % ls

dvcstore

(myenv) gauravshukla@XMRK9TWQW4WJ tmp % cd ..

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc push

Collecting |0.00 [00:00, ?entry/s]

Pushing

ERROR: unexpected error - Unable to locate credentials

Having any troubles? Hit us up at https://dvc.org/support, we are always happy to help!

(myenv) gauravshukla@XMRK9TWQW4WJ python % vi .dvc/config

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc push

ERROR: failed to push data to the cloud - config file error: remote 'storage' doesn't exist

(myenv) gauravshukla@XMRK9TWQW4WJ python % vi .dvc/config

(myenv) gauravshukla@XMRK9TWQW4WJ python % vi .dvc/config

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc push

Collecting |0.00 [00:00, ?entry/s]

Pushing

ERROR: failed to push data to the cloud - config file error: no remote specified in /Users/gauravshukla/bits/python. Setup default remote with

dvc remote default <remote name>

or use:

dvc <command> -r <remote name>

(myenv) gauravshukla@XMRK9TWQW4WJ python % vi .dvc/config

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc remote default myremote

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc push

Collecting |1.00 [00:00, 537entry/s]

Pushing

1 file pushed

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc pull

Collecting |1.00 [00:00, 724entry/s]

Fetching

Building workspace index |2.00 [00:00, 238entry/s]

Comparing indexes |3.00 [00:00, 2.83kentry/s]

Applying changes |0.00 [00:00, ?file/s]

Everything is up to date.

(myenv) gauravshukla@XMRK9TWQW4WJ python % cp data/data.xml /tmp/data.xml

(myenv) gauravshukla@XMRK9TWQW4WJ python % cat /tmp/data.xml >> data/data.xml

(myenv) gauravshukla@XMRK9TWQW4WJ python % vi /tmp/data.xml

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc add data/data.xml

100% Adding...|█████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████|1/1 [00:00, 8.01file/s]

To track the changes with git, run:

git add data/data.xml.dvc

To enable auto staging, run:

dvc config core.autostage true

(myenv) gauravshukla@XMRK9TWQW4WJ python % dvc push

Collecting |1.00 [00:00, 777entry/s]

Pushing

1 file pushed

(myenv) gauravshukla@XMRK9TWQW4WJ python % git commit data/data.xml.dvc -m "Dataset updates"

Auto packing the repository in background for optimum performance.

See "git help gc" for manual housekeeping.

warning: The last gc run reported the following. Please correct the root cause

and remove .git/gc.log

Automatic cleanup will not be performed until the file is removed.

warning: There are too many unreachable loose objects; run 'git prune' to remove them.

[feature/v0.1 2fc60f9] Dataset updates

1 file changed, 2 insertions(+), 2 deletions(-)

(myenv) gauravshukla@XMRK9TWQW4WJ python % git status

On branch feature/v0.1

Your branch is ahead of 'origin/feature/v0.1' by 2 commits.

(use "git push" to publish your local commits)

Changes not staged for commit:

(use "git add <file>..." to update what will be committed)

(use "git restore <file>..." to discard changes in working directory)

modified: .dvc/config

modified: README.md

Untracked files:

(use "git add <file>..." to include in what will be committed)

.DS\_Store

data/get-started/

mlopsgroup44/

myenv/

tmp/

~$signment\_Document.docx

no changes added to commit (use "git add" and/or "git commit -a")

(myenv) gauravshukla@XMRK9TWQW4WJ python % git add .dvc/config README.md data/get-started/ tmp/

(myenv) gauravshukla@XMRK9TWQW4WJ python % git commit -m "dvc changes"

Auto packing the repository in background for optimum performance.

See "git help gc" for manual housekeeping.

warning: The last gc run reported the following. Please correct the root cause

and remove .git/gc.log

Automatic cleanup will not be performed until the file is removed.

warning: There are too many unreachable loose objects; run 'git prune' to remove them.

[feature/v0.1 c642a2b] dvc changes

4 files changed, 24 insertions(+), 7 deletions(-)

create mode 100644 data/get-started/.gitignore

create mode 100644 tmp/.DS\_Store

(myenv) gauravshukla@XMRK9TWQW4WJ python % git push

Enumerating objects: 23, done.

Counting objects: 100% (23/23), done.

Delta compression using up to 10 threads

Compressing objects: 100% (15/15), done.

Writing objects: 100% (18/18), 1.99 KiB | 203.00 KiB/s, done.

Total 18 (delta 6), reused 3 (delta 0), pack-reused 0

remote: Resolving deltas: 100% (6/6), completed with 2 local objects.

To github.com:shuklagauravjn/mlopsgroup44

ad517da..c642a2b feature/v0.1 -> feature/v0.1

(myenv) gauravshukla@XMRK9TWQW4WJ python %

• Show how to revert to a previous version of the dataset.

Switching between versions

git checkout <...>

$ **dvc checkout**

Return to a previous version of the dataset

git checkout HEAD~1 data/data.xml.dvc

$ **dvc checkout**

git commit data/data.xml.dvc -m "Revert dataset updates"

**Deliverables**:

• MLflow experiment logs with different runs and their results.

• A DVC repository showing different versions of the dataset.

**M3: Model Experimentation and Packaging**

**Objective**: Train a machine learning model, perform hyperparameter tuning, and package the model for deployment.

**Tasks**:

1. **Hyperparameter Tuning**:

• Use a library like Optuna or Scikit-learn’s GridSearchCV to perform hyperparameter tuning on a chosen model.

• Document the tuning process and the best parameters found.

2. **Model Packaging**:

• Package the best-performing model using tools like Docker and Flask.

• Create a Dockerfile and a simple Flask application to serve the model.

These are the screen shot which shows docker build. In the screen shot below docker build -t shuklagauravjn/ml-ops-group44-v-01:latest .

Leads to building of docker container.

A screenshot of a computer program

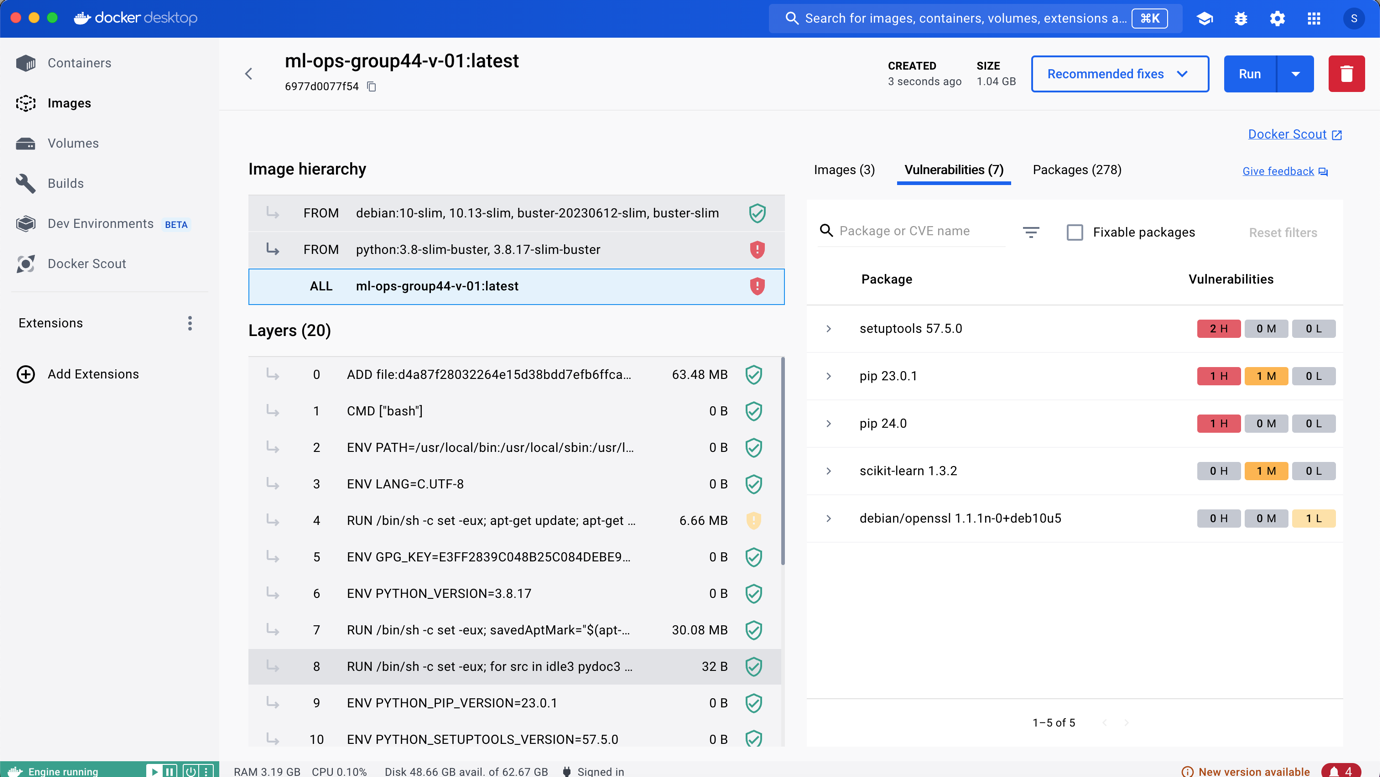
Description automatically generated

Once the docker container is built, it’s visible in docker desktop on mac.

A screenshot of a computer

Description automatically generated

The screen shot below shows the docker image and possible vulnerabilities.



Once the docker container was run from docker desktop, here is the log:A screenshot of a computer

Description automatically generated

Later this docker image was pushed to docker hub using the following command:

docker login -u "shuklagauravjn" -p "XSDSDwewe#" docker.io

docker push shuklagauravjn/ml-ops-group44-v-01:latest

A screenshot of a computer program

Description automatically generated

So it can be seen in hub.docker.com

A screenshot of a computer

Description automatically generated

**Deliverables**:

• A report on hyperparameter tuning results.

• A Dockerfile and Flask application code.

• Screenshots of the model running in a Docker container.

**M4: Model Deployment & Orchestration (Optional)**

**Objective**: Deploy a machine learning model and orchestrate its operations using Kubernetes.

**Tasks**:

1. **Model Deployment**:

• Deploy the Dockerized model from M3 to a cloud platform like AWS, Azure, or GCP.

• Use a platform service like AWS ECS, Azure AKS, or Google Kubernetes Engine (GKE).

2. **Orchestration**:

• Set up a Kubernetes cluster.

For local mac following things were done:

brew install minikube

minikube start

kubectl create deployment ml-ops-group44-v-01 --image=shuklagauravjn/ml-ops-group44-v-01:latest

kubectl expose deployment ml-ops-group44-v-01 --type=NodePort --port=8080

kubectl get services ml-ops-group44-v-01

minikube service ml-ops-group44-v-01

A screenshot of a computer

Description automatically generated

• Deploy the model using Kubernetes and create a Helm chart for managing deployments.

**Deliverables**:

• A link to the deployed model endpoint.

• Kubernetes configuration files and Helm chart.

• A report detailing the deployment and orchestration process.

**M5: Final Deliverables**

• A zip file containing:

• Code

• Data

• Model

• A one-page summary that includes:

• Description of the work completed

• Justification for the choices made

• A screen recording (maximum 5 minutes) that:

* Explains the work done
* Shows the results