Steps to be followed to utilize dvc for data versioning:

pip install dvc

Initialize git and dvc using below commands

git init

dvc init

Now check the status using

git status

Commit your repo using

git commit -m 'initialize repo'

Configure the remote storage using dvc. # We can also store it in cloud

dvc remote add -d dvc-remote /tmp/dvc-storage

Content of config dvc file can be checked using

cat .dvc/config

Now commit the changes to the repo using

git commit .dvc/config -m 'configure remote storage'

create a new directory using

mkdir data

Then copy the data to this directory manually or using

cp /Volumes/GL/TECH/IIITH/TA\_Session\_Material/DVC\_Handson/Iris.csv data

You can check whether the file is copied using

ls -l data

# It will show the Iris.csv file and it size

Now you can add the dataset to dvc using

dvc add data/Iris.csv

Now we can see again what is there in the data folder using

ls -l data

# we can find Iris.csv and Iris.csv.dvc datasets

Now we can check what is there inside Iris.csv.dvc dataset using

cat data/Iris.csv.dvc

# It contains the information of the dataset and the name of the file as path

Now we can check the content of gitignore file using

cat data/.gitignore

# We can see only the original dataset Iris.csv, we can’t see .dvc file. This show sepreation between git and dvc files

Now we need to add both the files to git using

git add data/.gitignore data/Iris.csv.dvc

Then we need to commit using

git commit -m 'data: track'

Then we can tag the dataset using

git tag -a 'v1' -m 'raw-data'

To push the data from local to remote storage use

dvc push

Now we can check whether the file is pushed to the storage or not using

ls -lR /tmp/dvc-storage

Now we can remove the dataset from our remote folder using

rm -rf data/Iris.csv

# Only .csv file to be removed and not the .dvc file. .dvc file is link, so it should not be deleted. Else we have to clone it back.

To check the files in cache

ls -l .dvc/cache

If we decide to use the dataset again then we can use

dvc pull

Now we can reduce the size of the dataset using

sed -i '' '1,0001d' data/Iris.csv

# Here 10001d represents the size of the data in bytes

We can check the size using

ls -l data

Now we can add/copy the reduced file to dvc using

dvc add data/Iris.csv

To add it to git use

git add data/Iris.csv.dvc

We can now commit again using

git commit -m 'data: remove 15 lines'

Create a tag again using

git tag -a 'v2' -m 'removed 15 lines'

Copy to remote storage using

dvc push

Delete the data and cache

rm -rf data/Iris.csv

rm -rf .dvc/cache