

Lab3 : Versionnement des données et pipelines ML avec DVC

Étape 1 : Initialisation de DVC dans le projet

```
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> pip install dvc
Collecting dvc
  Using cached dvc-3.65.0-py3-none-any.whl.metadata (17 kB)
Collecting attrs>=22.2.0 (from dvc)
```

```
• (venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> 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>>

Étape 2 : Versionner les données brutes avec DVC

```
❖ .gitignore
1  venv_mlops/
2  __pycache__/
3  *.pyc
4  logs/
5  models/
6  #data/
```

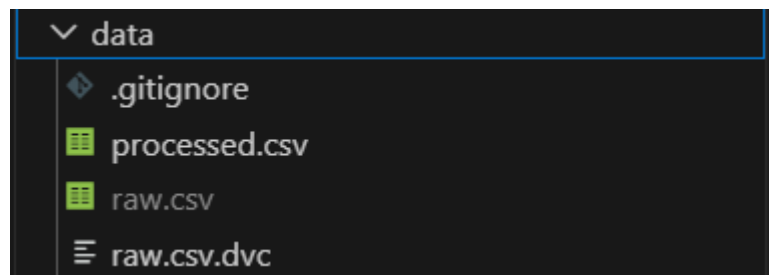
```
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> dvc add data/raw.csv
100% Adding...| 1/1 [00:00, 5.46file/s]

To track the changes with git, run:

    git add 'data/raw.csv.dvc' 'data/.gitignore'

To enable auto staging, run:

    dvc config core.autostage true
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01>
```



```
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> git add data/raw.csv.dvc data/.gitignore .gitignore
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> git commit -m "data: suivi du dataset brut via DVC"
[feature/drift-last-n 22bfab1] data: suivi du dataset brut via DVC
3 files changed, 7 insertions(+), 1 deletion(-)
create mode 100644 data/.gitignore
create mode 100644 data/raw.csv.dvc
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01>
```

Étape 3 : Configuration d'un remote DVC

```
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> mkdir dvc_storage

Directory: C:\Users\anoua\projects\MLOPS\mlops-lab-01

Mode                LastWriteTime         Length Name
----                -
d-----            1/4/2026  12:11 PM             dvc_storage

(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> dvc remote add -d localremote dvc_storage
Setting 'localremote' as a default remote.
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01>
```

```
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> git add .dvc/config
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> git commit -m "dvc: configuration du remote local"
[feature/drift-last-n ed87e84] dvc: configuration du remote local
1 file changed, 4 insertions(+)
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01>
```

Étape 4 : Push des données dans le remote DVC

```
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> dvc push
Collecting |1.00 [00:00, 200entry/s]
Pushing
1 file pushed
```

```
72dd6dd1b8e9f60c239ca5079b4267 x api.py monitor_drift.py rollback.py .gitignore .
dvc_storage > files > md5 > 7e > 72dd6dd1b8e9f60c239ca5079b4267
1 tenure_months,num_complaints,avg_session_minutes,plan_type,region,churn
2 6,3,21.5,basic,AF,1
3 46,3,50.19,basic,EU,0
4 39,2,47.79,basic,AF,0
5 26,3,41.45,premium,AF,1
6 26,2,58.67,premium,EU,0
7 51,0,32.73,basic,AF,0
8 6,0,71.38,basic,NA,0
9 42,3,32.27,basic,AF,1
10 12,3,38.06,basic,NA,1
11 6,4,30.23,basic,EU,1
12 32,1,39.47,basic,NA,0
13 58,0,45.03,premium,AS,1
14 44,4,33.63,premium,EU,1
15 45,1,56.32,basic,AF,1
16 43,0,37.87,basic,EU,0
17 47,2,38.25,basic,EU,1
18 31,2,24.09,basic,NA,1
19 8,2,40.02,basic,AS,1
```

Étape 5 : imulation d'une collaboration : supprimer localement et récupérer depuis DVC

```
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> del data\raw.csv
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> ls data/

Directory: C:\Users\anoua\projects\MLOPS\mlops-lab-01\data

Mode                LastWriteTime         Length Name
----                -
-a----             1/4/2026  12:05 PM             10 .gitignore
-a----             1/3/2026  10:11 PM          28240 processed.csv
-a----             1/4/2026  12:05 PM             93 raw.csv.dvc
```

```
(venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> dvc pull
Collecting |1.00 [00:00, 219entry/s]
Fetching
Building workspace index |1.00 [00:00, 184entry/s]
Comparing indexes |3.00 [00:00, 844entry/s]
Applying changes |1.00 [00:00, 115file/s]
A data\raw.csv
1 file added
```

Mode	LastWriteTime	Length	Name
----	-----	-----	----
-a----	1/4/2026 12:05 PM	10	.gitignore
-a----	1/3/2026 10:11 PM	28240	processed.csv
-a----	1/4/2026 12:23 PM	27933	raw.csv
-a----	1/4/2026 12:05 PM	93	raw.csv.dvc

Étape 6 : Création d'un pipeline reproductible dvc.yaml

```

! dvc.yaml
1  stages:
2    prepare:
3      cmd: python src/prepare_data.py
4      deps:
5        - data/raw.csv
6        - src/prepare_data.py
7      outs:
8        - data/processed.csv
9        - registry/train_stats.json
10   train:
11     cmd: python src/train.py
12     deps:
13       - data/processed.csv
14       - src/train.py
15     outs:
16       - models/model.joblib
17   evaluate:
18     cmd: python src/evaluate.py
19     deps:
20       - data/processed.csv
21       - models/model.joblib
22       - src/evaluate.py
23     outs:
24       - reports/metrics.json
25

```

```

• (venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> git add dvc.yaml registry/ reports/
• (venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> git commit -m "pipeline: ajout des étapes prepare/train/evaluate"
[feature/drift-last-n 55d09a0] pipeline: ajout des étapes prepare/train/evaluate
3 files changed, 47 insertions(+)
create mode 100644 dvc.yaml
create mode 100644 registry/train_stats.json
create mode 100644 reports/metrics.json
○ (venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01>

```

Étape 7 : Reproduire automatiquement tout le pipeline

```

• (venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> dvc repro
'data\raw.csv.dvc' didn't change, skipping
Running stage 'prepare':
> python src/prepare_data.py
[OK] Fichier prétraité généré : C:\Users\anoua\projects\MLOPS\mlops-lab-01\data\processed.csv
[OK] Statistiques d'entraînement générées : C:\Users\anoua\projects\MLOPS\mlops-lab-01\registry\train_stats.json
Hello
cheeck
Updating lock file 'dvc.lock'

Stage 'train' didn't change, skipping
Stage 'evaluate' didn't change, skipping

To track the changes with git, run:

    git add 'registry\.gitignore' dvc.lock 'data\.gitignore'

```

```

• (venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> git add dvc.lock
• (venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01> git commit -m "pipeline: lock after repro"
[feature/drift-last-n ef88a24] pipeline: lock after repro
1 file changed, 58 insertions(+)
create mode 100644 dvc.lock
○ (venv_mlops) PS C:\Users\anoua\projects\MLOPS\mlops-lab-01>

```