

Lab3 : Versionnement des données et pipelines ML avec DVC

Étape 1 : Initialisation de DVC dans le projet

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> pip install dvc
Downloading filelock-3.20.1-py3-none-any.whl (16 kB)
Downloading tzlocal-5.3.1-py3-none-any.whl (18 kB)
Building wheels for collected packages: antlr4-python3-runtime
  Building wheel for antlr4-python3-runtime (pyproject.toml) ... done
  Created wheel for antlr4-python3-runtime: filename=antlr4_python3_runtime-4.9.3-py3-none-any.whl size=146615 sha256=7bad2bce38acc5daa022c1c7f0e1d5ae04454f9c85f5e624f783a23dfca02246
  Stored in directory: c:\users\pc\appdata\local\pip\cache\wheels\5\5\3\74\c35b65048c30e631c674c9c5475e6f6b69a467981446bd8
Successfully built antlr4-python3-runtime
Installing collected packages: pygit2, funcy, dictdiffer, appdirs, antlr4-python3-runtime, zc.lockfile, voluptuous, vine, tzlocal, tqdm, toolkit, tabulate, smmap, shtab, shortuuid, shellingham, sewer, ruamel.yaml.clib, python-dotenv, pydot, propcache, pathspec, orjson, omegaconf, networks, multidict, mdurl, grandalf, fsspec, frozenlist, flatten_dict, filelock, exceptiongroup, entrypoints, dvc-render, dulwich, dpath, distro, diskcache, configobj, cffi, billiard, atpublic, aiohappyeyeballs, yarl, sqttrie, ruamel.yaml, pygit2, markdown-it-py, iterative-telemetry, hydra-core, gitdb, fluff.lock, dvc-studio-client, dvc-objects, cryptography, click-repl, click-plugins, click-didyoumean, amqp, aiosignal, rich, pydantic-settings, kombu, gitpython, dvc-data, asyncssh, aiohttp, typer, celery, aiohttp-retry, scorepo, dvc-task, dvc-http, gto, DVC
Attempting uninstall: cffi
  Found existing installation: cffi 1.17.1
  Uninstalling cffi-1.17.1:
    Successfully uninstalled cffi-1.17.1
Successfully installed DVC-3.65.0 aiohappyeyeballs-2.6.1 aiohttp-3.13.2 aiohttp-retry-2.9.1 aiosignal-1.4.0 amqp-5.3.1 antlr4-python3-runtime-4.9.3 appdirs-1.4.4 asyncssh-2.22.0 atpublic-7.0.0 billiard-4.2.0 celery-5.6.0 cffi-2.0.0 click-didyoumean-0.3.1 click-plugins-1.1.1.2 click-repl-0.3.0 configobj-5.0.9 cryptography-46.0.3 dictdiffer-0.9.0 diskcache-5.6.3 distro-1.9.0 dpath-2.2.0 dulwich-0.25.0 dvc-data-3.17.0 dvc-http-2.32.0 dvc-objects-5.2.0 dvc-render-1.0.2 dvc-studio-client-0.22.0 dvc-task-0.08.2 entrypoints-0.4 exceptiongroup-1.3.1 filelock-3.20.1 flatten_dict-0.4.2 fluff.lock-8.2.0 frozenlist-1.8.0 fsspec-2025.12.0 funcy-2.0 gitdb-4.0.12 gitpython-3.1.45 grandalf-1f-0.8 gto-1.9.0 hydra-core-1.3.2 iterative-telemetry-0.0.10 kombu-5.6.1 markdown-it-py-4.0.0 mdurl-0.1.2 multidict-6.7.0 networks-1.6.1 omegaconf-2.3.0 orjson-3.11.5 pat-hspec-0.12.1 propcache-0.4.1 pydantic-settings-2.12.0 pydot-4.0.1 pygit2-1.19.0 pygit2-1.19.0 pygit2-1.19.0 python-dotenv-1.2.1 rich-14.2.0 ruamel.yaml-0.18.17 ruamel.yaml.clib-0.2.15 scorepo-3.6.0 sewer-3.0.4 shellingham-1.5.4 shortuuid-1.0.13 shtab-1.8.0 smmap-5.0.2 sqttrie-0.11.2 tabulate-0.9.0 toolkit-0.13.1 tqdm-4.67.1 typer-0.21.0 tzlocal-5.3.1 vine-5.1.0 voluptuous-0.16.0 yarl-1.22.0 zc.lockfile-4.0
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>
```

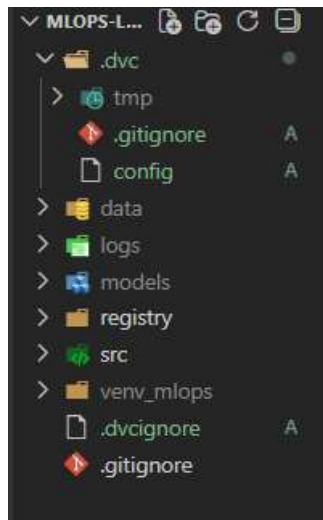
```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc --version
3.65.0
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>
```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc init
Initialized DVC repository.

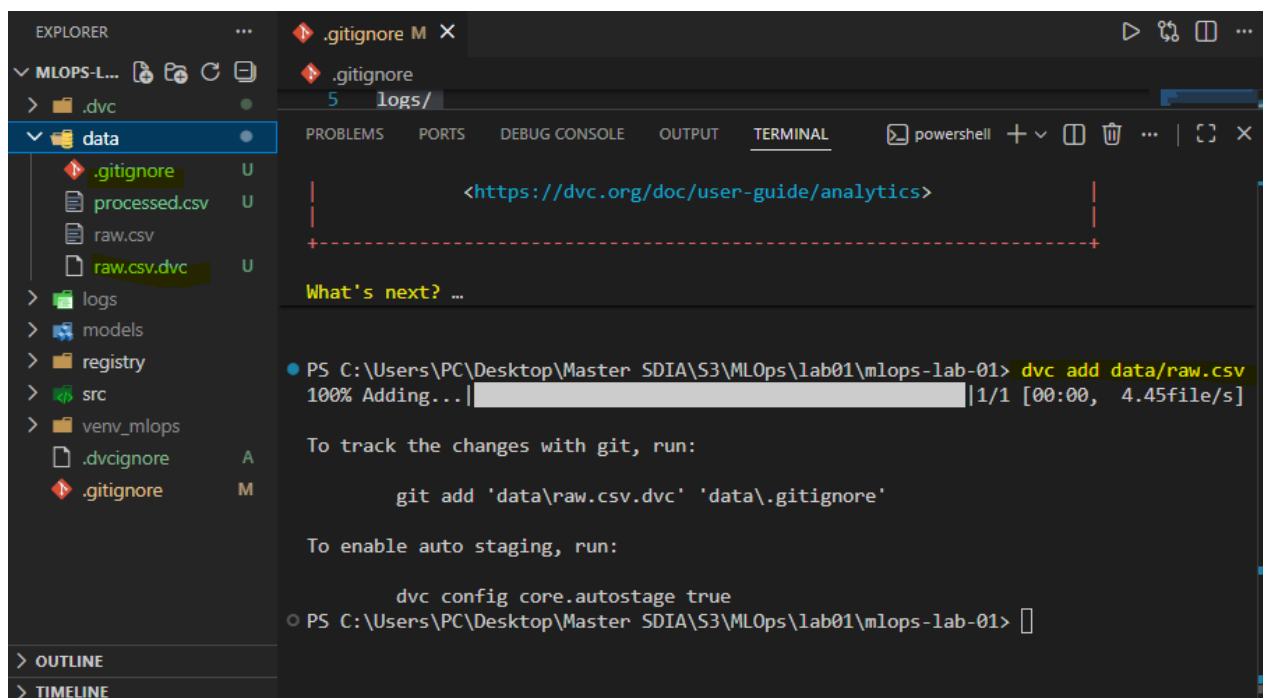
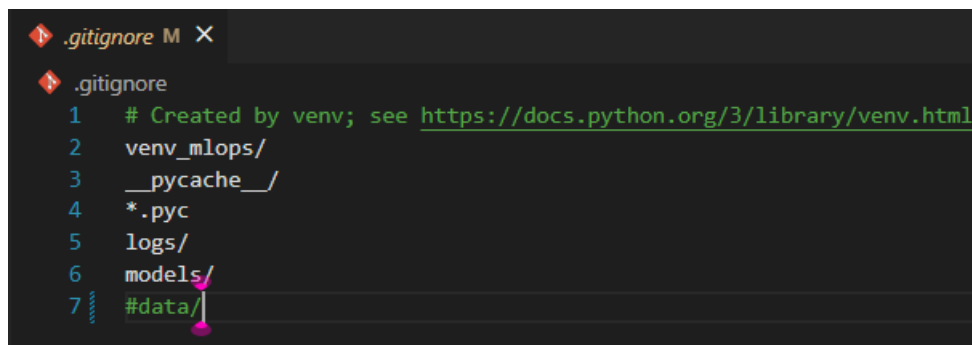
You can now commit the changes to git.

+-----+
|                                               |
| 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>
- Star us on GitHub: <https://github.com/treeverse/dvc>
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>
```



Étape 2 : Versionner les données brutes avec DVC



```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git add data/*.dvc .gitignore
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git commit -m "data: suivi du dataset brut via DVC"
[feature/drift-last-n e6ae3c3] data: suivi du dataset brut via DVC
5 files changed, 12 insertions(+), 1 deletion(-)
create mode 100644 .dvc/.gitignore
create mode 100644 .dvc/config
create mode 100644 .dvcignore
create mode 100644 data/raw.csv.dvc
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>

```

Étape 3: Configuration d'un remote DVC

```

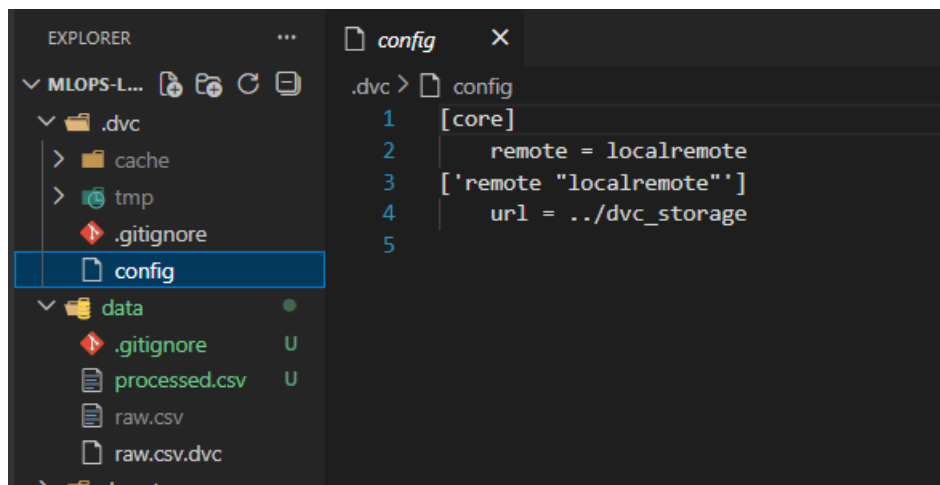
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> mkdir dvc_storage

Répertoire : C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01

Mode                LastWriteTime         Length Name
----                -
d-----          27/12/2025   10:31             dvc_storage

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc remote add -d localremote dvc_storage
Setting 'localremote' as a default remote.

```



```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git add .dvc/config
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git commit -m "dvc: configuration du remote local"
[feature/drift-last-n c5c6f50] dvc: configuration du remote local
1 file changed, 4 insertions(+)

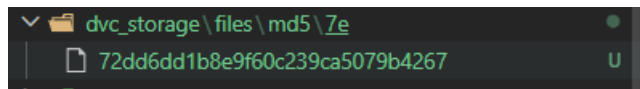
```

Étape 4 : Push des données dans le remote DVC

```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc push
Collecting |1.00 [00:00, 362entry/s]
Pushing
1 file pushed

```



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

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> del data\raw.csv
```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> ls data/

Répertoire : C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\data

Mode                LastWriteTime         Length Name
----                -
-a----           27/12/2025   10:15             10 .gitignore
-a----           14/12/2025   15:47          28240 processed.csv
-a----           27/12/2025   10:15             93 raw.csv.dvc
```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc pull
Collecting |1.00 [00:00, 301entry/s]
Fetching
Building workspace index |1.00 [00:00, 326entry/s]
Comparing indexes |3.00 [00:00, 1.30kentry/s]
Applying changes |1.00 [00:00, 164file/s]
A      data\raw.csv
1 file added
```

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

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc add data/processed.csv
100% Adding... |1/1 [00:00, 4.55file/s]

To track the changes with git, run:

    git add 'data\processed.csv.dvc' 'data\.gitignore'

To enable auto staging, run:

    dvc config core.autostage true
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc add registry/train_stats.json
Adding...
ERROR: output 'registry/train_stats.json' is already tracked by SCM (e.g. Git).
You can remove it from Git, then add to DVC.
To stop tracking from Git:
    git rm -r --cached 'registry/train_stats.json'
    git commit -m "stop tracking registry/train_stats.json"
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>
```

```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git rm --cached registry/train_stats.json
rm 'registry/train_stats.json'
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git commit -m "stop tracking registry/train_stats.json"
[feature/drift-last-n 50b5834] stop tracking registry/train_stats.json
1 file changed, 14 deletions(-)
delete mode 100644 registry/train_stats.json
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc add registry/train_stats.json
100% Adding... | 1/1 [00:00, 9.75file/s]

To track the changes with git, run:

    git add 'registry/train_stats.json.dvc' 'registry/.gitignore'

To enable auto staging, run:

    dvc config core.autostage true
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>

```

```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git add data/processed.csv.dvc registry/train_stats.json.dvc
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git commit -m "data: versionnement des données préparées et statistiques"
[feature/drift-last-n 0b4514e] data: versionnement des données préparées et statistiques
2 files changed, 10 insertions(+)
create mode 100644 data/processed.csv.dvc
create mode 100644 registry/train_stats.json.dvc
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>

```

```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc push
Collecting                                     [3.00 [00:00, 314entry/s]
Pushing
2 files pushed

```

```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc stage add -n prepare `
>> -d src/prepare_data.py `
>> -d data/raw.csv `
>> -o data/processed.csv `
>> -o registry/train_stats.json `
>> python src/prepare_data.py
Added stage 'prepare' in 'dvc.yaml'

To track the changes with git, run:

    git add 'data/.gitignore' dvc.yaml 'registry/.gitignore'

To enable auto staging, run:

    dvc config core.autostage true
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>

```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc stage add -n train `
>> -d src/train.py -d data/processed.csv `
>> -o models `
>> python src/train.py
Added stage 'train' in 'dvc.yaml'
```

To track the changes with git, run:

```
git add dvc.yaml .gitignore
```

To enable auto staging, run:

```
dvc config core.autostage true
```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>
```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc stage add -n evaluate `
>> -d src/evaluate.py `
>> -d models/model.joblib `
>> -d data/processed.csv `
>> -o reports/metrics.json `
>> python src/evaluate.py
Could not create .gitignore entry in C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\reports\.gitignore. DVC will attempt to create .gitignore entry again when the stage is run.
Added stage 'evaluate' in 'dvc.yaml'
```

To track the changes with git, run:

```
git add dvc.yaml
```

To enable auto staging, run:

```
dvc config core.autostage true
```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>
```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git add dvc.yaml
```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git commit -m "pipeline: ajout d
es étapes prepare et train"
```

```
[feature/drift-last-n d56834a] pipeline: ajout des étapes prepare et train
1 file changed, 24 insertions(+)
create mode 100644 dvc.yaml
```

```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>
```

```
config  train.py M X  evaluate.py
> train.py > ...
64 def main(version: str = "v1", seed: int = 42, gate_f1: float = 0.60)
65
66     # Log des métriques
67
68     print("[METRICS]", json.dumps(metrics, indent=2))
69
70     print(f"[OK] Modèle sauvegardé : {model_path}")
71
72
73     # Logique de "registry" minimal : mise à jour du modèle courant
74
75     if entry["passed_gate"]:
76         REGISTRY_DIR.mkdir(parents=True, exist_ok=True)
77         CURRENT_MODEL_PATH.write_text(
78             model_filename,
79             encoding="utf-8",
80         )
81         print(f"[DEPLOY] Modèle activé (current): {model_filename}")
82     else:
83         print(
84             "[DEPLOY] Refusé par le gate : F1 insuffisante "
85             "ou baseline non battue."
86         )
87
88
89 if __name__ == "__main__":
90     main()
```

```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc dag
● WARNING: Unable to find `less` in the PATH. Check out <https://man.dvc.org/pipeline/show> for more info.

+-----+
| data\raw.csv.dvc |
+-----+

      *
      *
      *

+-----+
| prepare |
+-----+
**      **

**      *
*      **
+-----+
| train |
+-----+
**      **
**      **
*      *

+-----+
| evaluate |
+-----+

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>

```

Étape 7 : Reproduire automatiquement tout le pipeline

```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc repro
'data\raw.csv.dvc' didn't change, skipping
Stage 'prepare' didn't change, skipping
Running stage 'train':
> python src/train.py
C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\src\train.py:713: DeprecationWarning: datetime.datetime.utcnow() is deprecated and scheduled for removal in a future version. Use timezone-aware objects to represent datetimes in UTC: datetime.datetime.now(datetime.UTC).
    timestamp = datetime.utcnow().strftime("YYYYMMDD%H%M%S")
[METRICS] {
  "accuracy": 0.6433333333333333,
  "precision": 0.668706080172974,
  "recall": 0.65625,
  "f1": 0.6624605678233438,
  "loss_line_f1": 0.0
}
Running stage 'evaluate':
> python src/evaluate.py
C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\src\evaluate.py:689: DeprecationWarning: datetime.datetime.utcnow() is deprecated and scheduled for removal in a future version. Use timezone-aware objects to represent datetimes in UTC: datetime.datetime.now(datetime.UTC).
    timestamp = datetime.utcnow().strftime("YYYYMMDD%H%M%S")
[METRICS] {
  "accuracy": 0.6433333333333333,
  "precision": 0.668706080172974,
  "recall": 0.65625,
  "f1_threshold_01": 0.6624605678233438,
  "f1": 0.7164179184477612,
  "test_threshold": 0.36,
  "loss_line_f1": 0.0
}
[OK] Model saved: C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\models\churn_model_v1_20251227_104801.joblib
[OK] Alias stable : C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\models\model1.joblib
[OK] Métriques sauvegardées : C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\reports\metrics.json
[DEPLOY] Model acté : churn_model_v1_20251227_104801.joblib
Updating lock file 'dvc.lock'

To track the changes with git, run:

    git add dvc.lock "reports\*.gitignore"

To enable auto staging, run:

    dvc config core.autostage true
Use 'dvc push' to send your updates to remote storage.
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01>

```