

## Lab3 : Versionnement des données et pipelines ML avec DVC

### Étape 1 : Initialisation de DVC dans le projet

```
PROBLEMS  PRINT  DEBUG CONSOLE  OUTPUT  TERMINAL
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 trilocal-5.3.1-py3-none-any.whl (16 kB)
Building wheels for collected packages: antir4-python3-runtime
  Building wheel for antir4-python3-runtime (pyproject.toml) ... done
    Created wheel for antir4-python3-runtime: filename=antir4_python3_runtime-4.0.3-py3-none-any.whl.size=144615 sha256=7bad7bce38cc9daa022a1c9e1d4a94549c35f9d624f783a23defca02246
    Stored in directory: c:\users\pc\appdata\local\pip\cache\wheels\53\74\23\5b6848c9e631c744c9475mfcb3e69a467981446bd8
Successfully built antir4-python3-runtime
Installing collected packages: pytris, funcy, dictdiffer, appdirs, antir4-python3-runtime, xx-lockfile, voluptuous, vine, tzlocal, tomkit, tabulate, smap, shlib, shortuid, shellingham, soname, ruamel.yaml.lib, python-dotenv, pydot, propcache, pathspec, orjson, omegaconf, networks, multidict, enum, grandalf, fsspec, frozenlist, flatten_dict, filelink, exceptiongroup, entrypoints, dvc-render, dulwich, dpath, distro, diskcache, configobj, cffi, billiard, aspublic, asynappympathlib, yaml, multiset, ruamel.yaml, pygit2, markdown-it-py, iterative-telemetry, hydra-core, gitdb, tuflock, dvc-studio-client, dvc-objects, cryptography, click-repl, click-plugins, click-dydynamen, amqp, alosignal, rich, pydantic-settings, komma, pypython, dvc-data, asynccss, aiohttp, typer, celery, aiohttp-retry, screnpo, 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 asynappympathlib-2.6.1 aiohttp-3.13.2 aiohttp-retry-2.9.1 aiosignal-1.4.0 amqp-5.3.1 amled4-python3-runtime-4.0.3 appdirs-1.4.4 asynssh-2.22.0 asthetic-2.0.0 billiard-4.2.0 celery-5.6.0 ffi-2.9.0 click-didyoumean-0.3.3 click-plugins-1.3.1 click-repl-0.1.0 configobj-5.0.9 cryptography-46.8.3 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-objects-5.7.0 dvc-render-1.0.2 dvc-studio-client-0.22.0 dvc-task-0.40.2 entrypoints-0.4 exceptiongroup-1.3.1 filelock-3.20.1 flatten_dict-0.4.2 fuflock-0.2.0 frozenlist-1.0.0 fsspec-2025.12.0 gitdb-4.0.12 gitpython-3.1.045 grandalf-0.8 grn-1.9.0 hydra-core-0.9.18 komma-5.6.1 markdown-it-py-4.0.0 ruamel-0.1.7 multidict-6.7.0 networks-3.6.1 omegaconf-2.1.0 orjson-3.11.5 pathep-0.12.1 propcache-0.4.1 pydantic-settings-2.12.0 pyot-4.0.1 pypit2-1.19.0 pytrie-2.5.8 python-dotenv-1.2.1 rich-14.2.0 ruamel.yaml-0.18.17 ruamel.yaml.lib-0.2.15 corepo-3.6.0 sensor-3.0.4 shellingham-1.5.4 shortuid-1.0.13 shlib-1.8.0 smap-5.0.2 sqlite-0.11.2 tabulate-0.9.0 tomkit-0.13.1 typer-0.21.0 tzlocal-5.3.1 vine-5.1.0 voluptuous-0.16.0 yaml-1.22.0 xx-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>
```

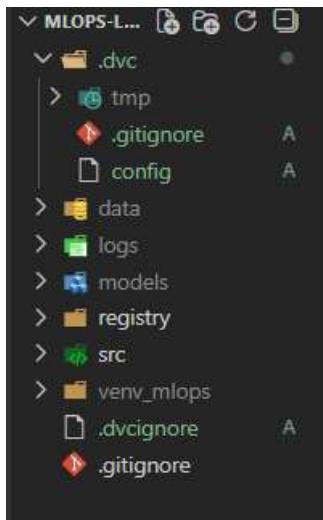
```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc init
Initialized DVC repository.
```

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>
+-----+
```

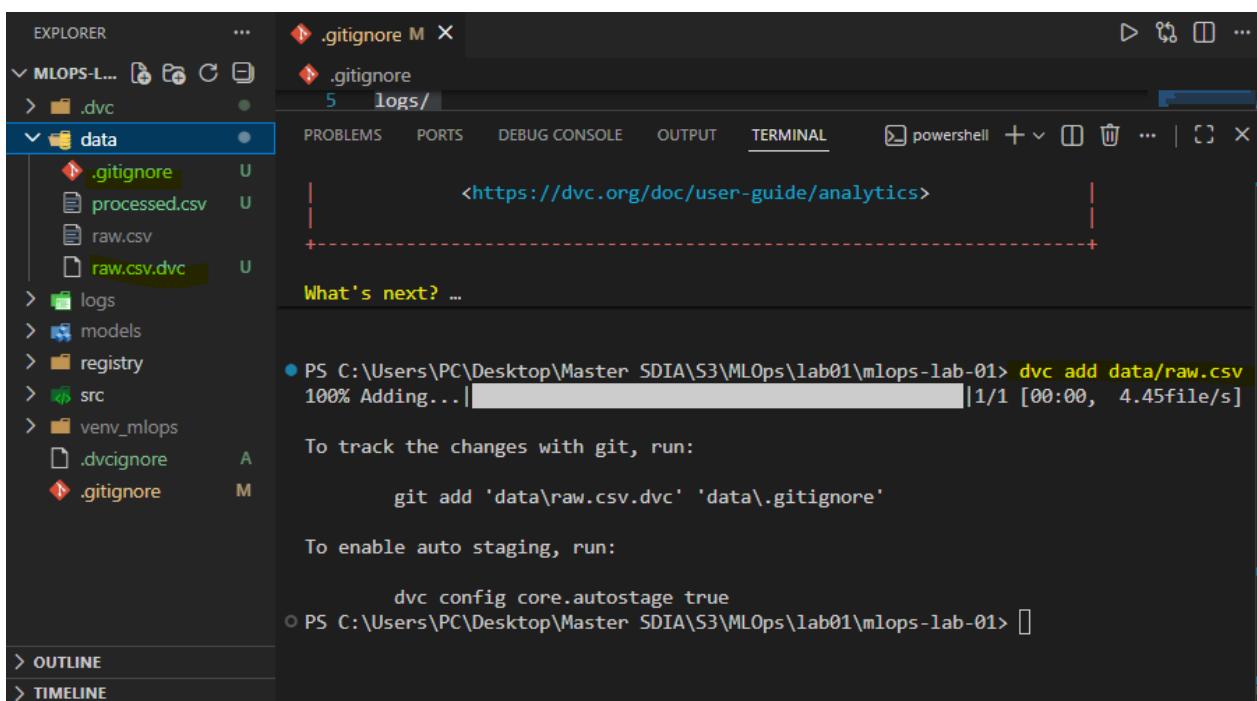
What's next?

- ```
-----
- 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

```
gitignore M X
.gitignore
1 # Created by venv; see https://docs.python.org/3/library/venv.html
2 venv_mlops/
3 __pycache__/
4 *.pyc
5 logs/
6 models/
7 #data/
```



```

PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-lab-01> git add data/*.dvc .gitignore
● PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-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\mllops-lab-01>

```

### Étape 3: Configuration d'un remote DVC

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

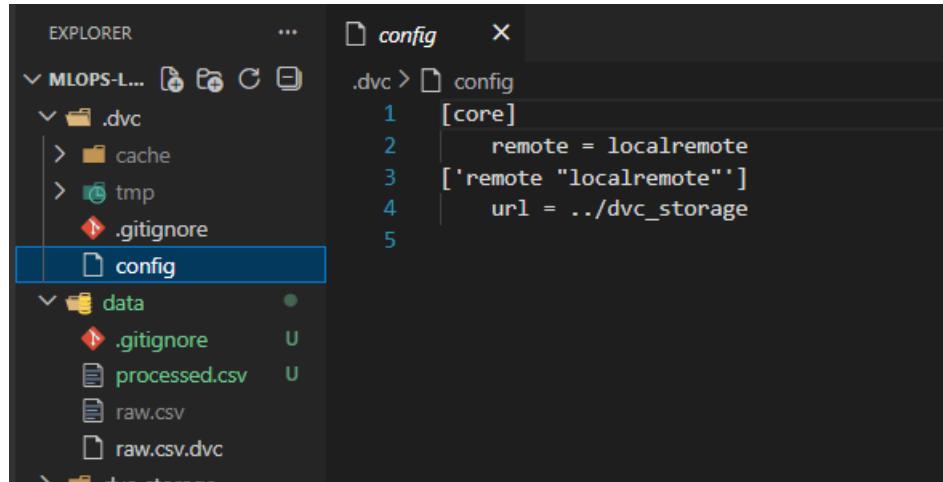
Répertoire : C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-lab-01

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

```

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

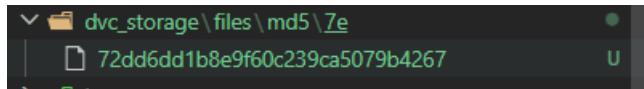
```

### Étape 4 : Push des données dans le remote DVC

```

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

```



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

```
+ FILE pushed
● PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-lab-01> del data\raw.csv
○ PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-lab-01>
```

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

```
● PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-lab-01> dvc pull
Collecting
Fetching
Building workspace index
Comparing indexes
Applying changes
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\mllops-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\mllops-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\mllops-lab-01>
```

```
● PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-lab-01> git rm --cached registry/train_stats.json  
rm 'registry/train_stats.json'  
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-lab-01> git commit -m "stop tracking registry/train_st  
● ts.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\mllops-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\mllops-lab-01>
```

```
○ PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-lab-01> git add data/processed.csv.dvc registry/train_s  
● tats.json.dvc  
● PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-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\mllops-lab-01> □
```

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

```
● PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mllops-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\mllops-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 des 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 > ...
54  def main(version: str = "v1", seed: int = 42, gate_f1: float = 0.60)
55
56  # Log des métriques
57
58  print("[METRICS]", json.dumps(metrics, indent=2))
59
60  print(f"[OK] Modèle sauvegardé : {model_path}")
61
62
63
64  # Logique de "registry" minimal : mise à jour du modèle courant
65
66  if entry["passed_gate"]:
67      REGISTRY_DIR.mkdir(parents=True, exist_ok=True)
68      CURRENT_MODEL_PATH.write_text(
69          model_filename,
70          encoding="utf-8",
71      )
72      print(f"[DEPLOY] Modèle activé (current): {model_filename}")
73  else:
74      print(
75          "[DEPLOY] Refusé par le gate : F1 insuffisante "
76          "ou baseline non battue."
77      )
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95  if __name__ == "__main__":
96
97      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\models\train.py:73: DeprecationWarning: datetime.datetime.utcnow() is deprecated and scheduled for removal in a future version. Use a timezone-aware objects to represent datetimes in UTC: datetime.datetime.now(datetime.UTC).
    timestamp = datetime.utcnow().strftime("%Y%m%d_%H%M%S")
[METRICS] {
  "accuracy": 0.643333333333333,
  "precision": 0.6687999999999997,
  "recall": 0.65625,
  "f1": 0.6246859578233438,
  "baseline_f1": 0.0
}

Running stage 'evaluate':
  python src/evaluate.py
C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\models\evaluate.py:49: DeprecationWarning: datetime.datetime.utcnow() is deprecated and scheduled for removal in a future version. Use a timezone-aware objects to represent datetimes in UTC: datetime.datetime.now(datetime.UTC).
    timestamp = datetime.utcnow().strftime("%Y%m%d_%H%M%S")
[METRICS] {
  "accuracy": 0.643333333333333,
  "precision": 0.6687999999999997,
  "recall": 0.65625,
  "f1_threshold_05": 0.6624685678233438,
  "f1": 0.7164379986477612,
  "best_threshold": 0.36,
  "baseline_f1": 0.0
}
[DOK] Module savegardes : C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\models\churn_model_v1_20251222_104800.joblib
[DOK] Alias stable : C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\models\stable.joblib
[DOK] Métriques savegardes : C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01\reports\metrics.json
[DEPLOY] Module activé : churn_model_v1_20251222_104800.joblib
Updating lock file 'dvc.lock'

In-track the changes with git, run:
  git add dvc.lock --reportsgitignore

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>

```

