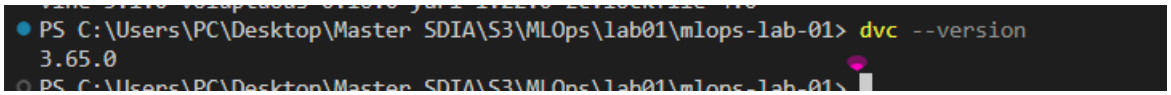
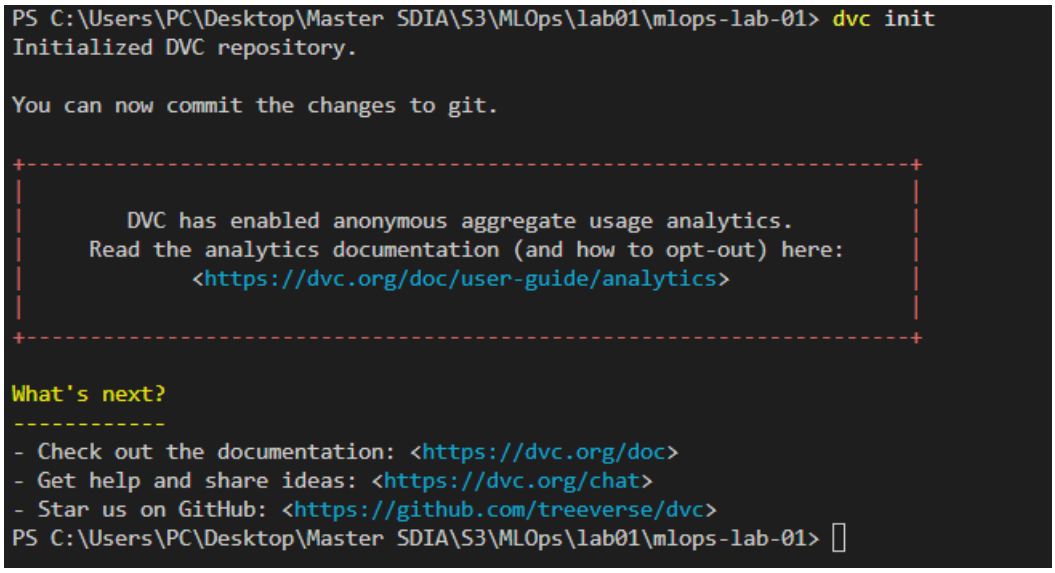
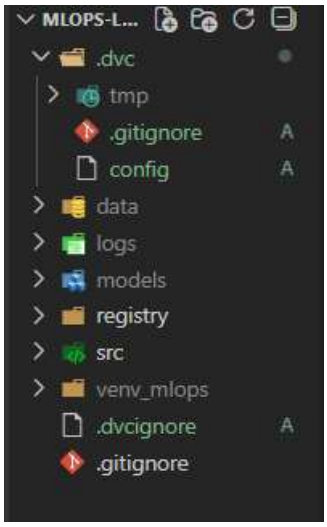
Lab3 : Versionnement des données et pipelines ML avec DVC

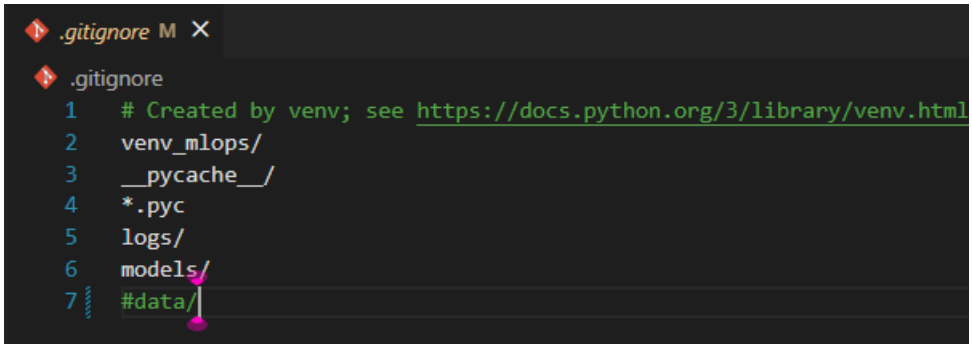
Étape 1 : Initialisation de DVC dans le projet

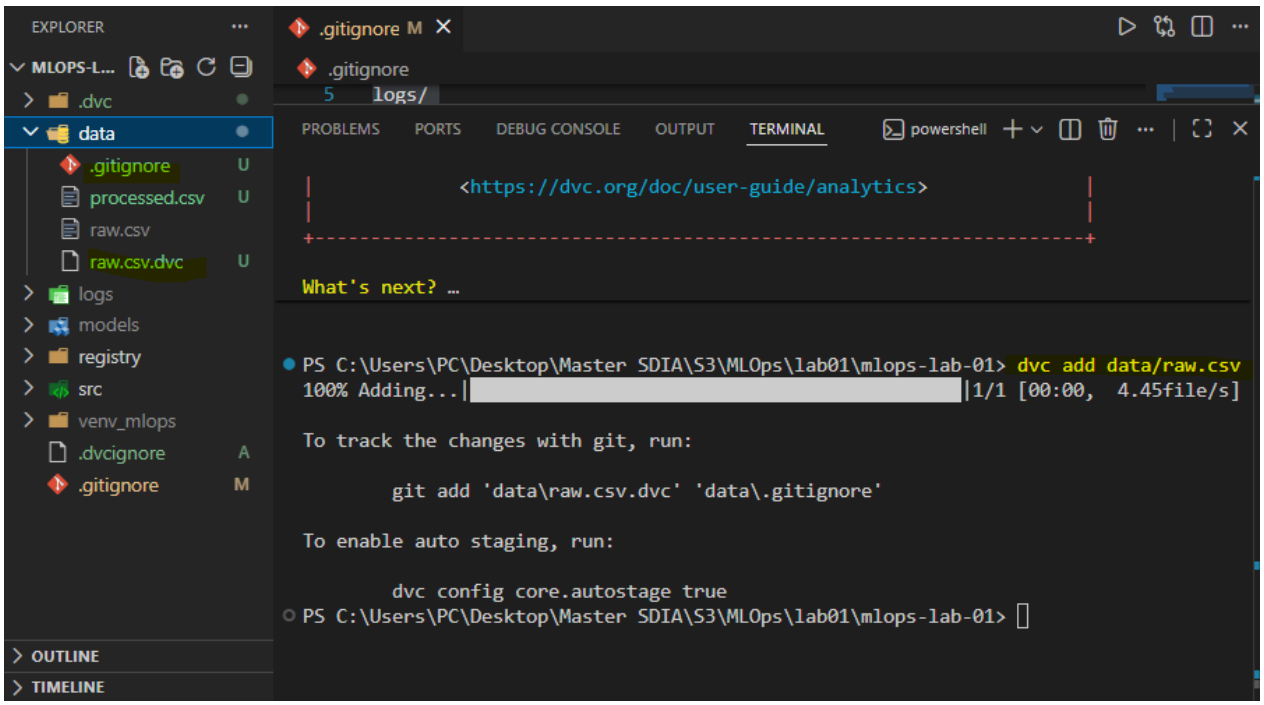






Étape 2 : Versionner les données brutes avec DVC



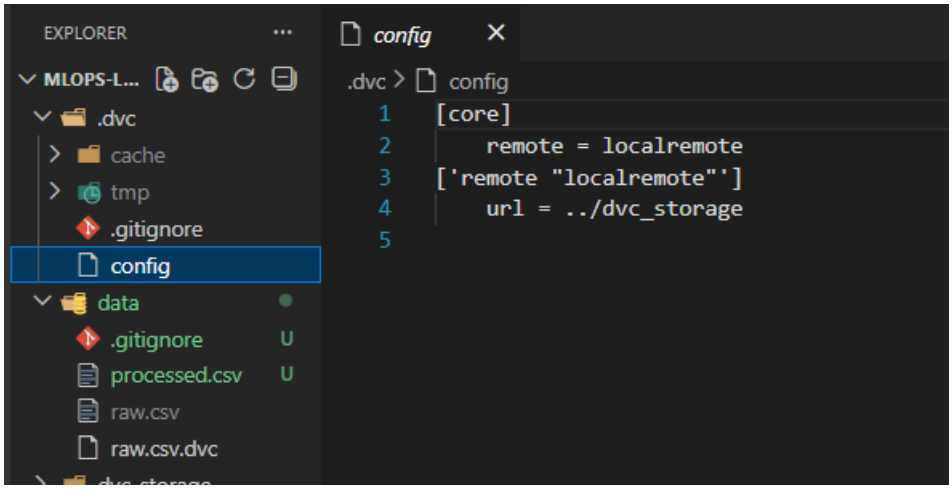


```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git add data/*.dvc .gitignore
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> git commit -m "data: suivi du dat
aset brut via DVC"
[feature/drift-last-n e6ae3c3] data: suivi du dataset brut via DVC
 5 files changed, 12 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
```

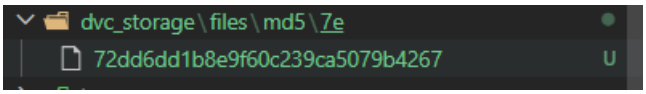
```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc remote add -d localremote dvc
_storage
Setting 'localremote' as a default remote.
```




```
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: configuratio
n 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>
```

```
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_sta
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\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_s
tats.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>
```

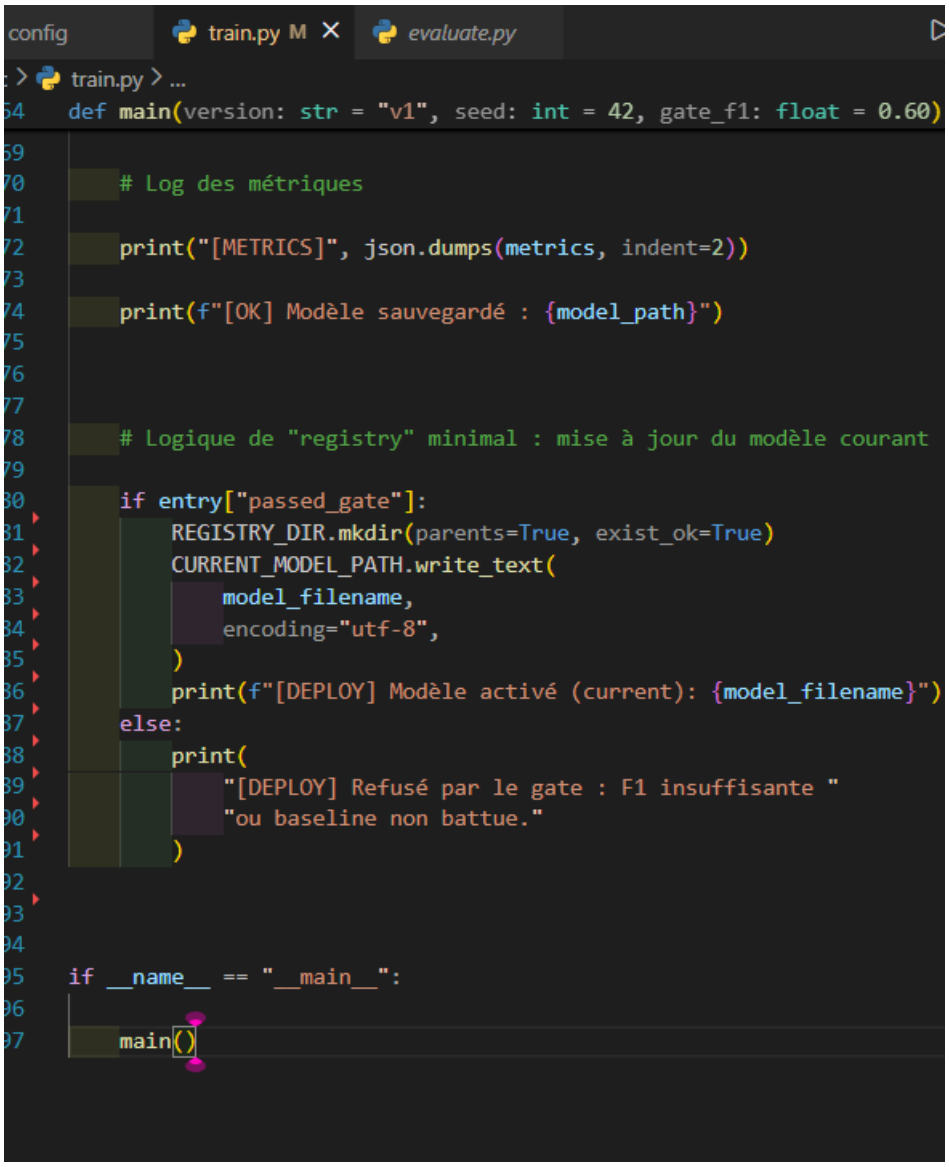
```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc push
Collecting                                                                     |3.00 [00:00,  314entry/s]
Pushing
2 files pushed
```

```
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>
```

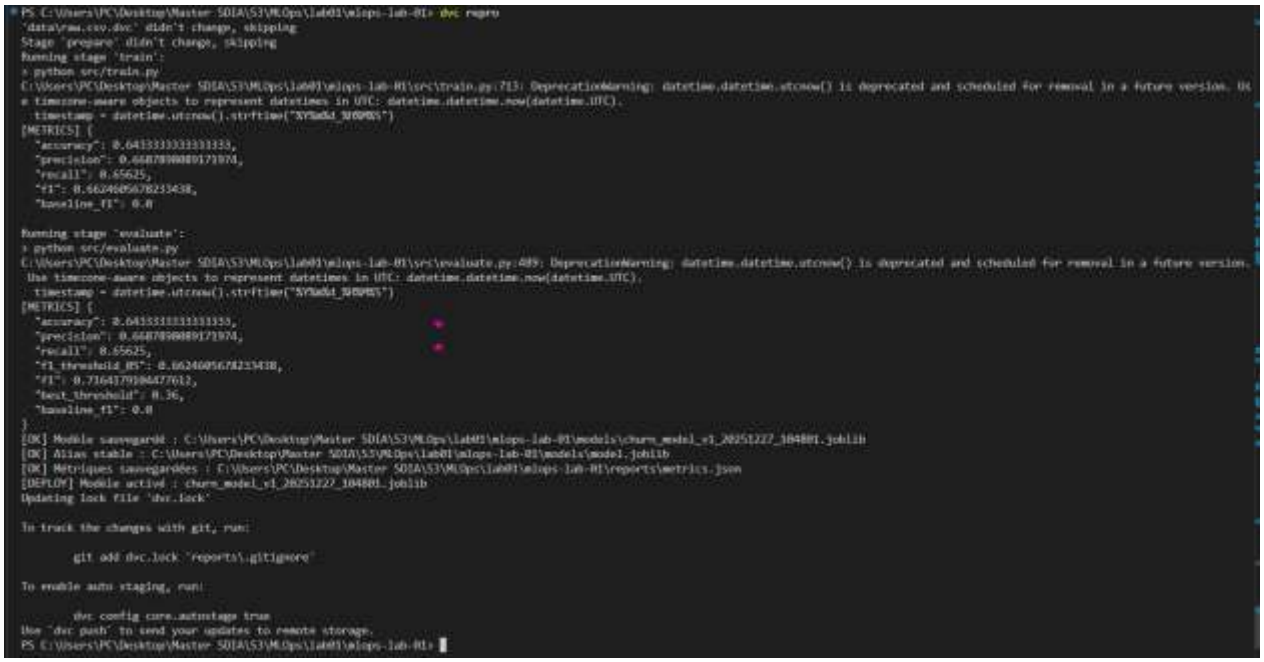
```
PS C:\Users\PC\Desktop\Master SDIA\S3\MLOps\lab01\mlops-lab-01> dvc stage add -n evaluate `
>>    -d src/evaluate.py `
>>    -d models/model.joblib `
>>    -d data/processed.csv `
>>    -o reports/metrics.json `
>>    python src/evaluate.py
Could not create .gitignore entry in C:\Users\PC\Desktop\Master 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
evaluate.py
train.py > ...
54 def main(version: str = "v1", seed: int = 42, gate_f1: float = 0.60)
59
70 # Log des métriques
71
72 print("[METRICS]", json.dumps(metrics, indent=2))
73
74 print(f"[OK] Modèle sauvegardé : {model_path}")
75
76
77
78 # Logique de "registry" minimal : mise à jour du modèle courant
79
80 if entry["passed_gate"]:
81 REGISTRY_DIR.mkdir(parents=True, exist_ok=True)
82 CURRENT_MODEL_PATH.write_text(
83 model_filename,
84 encoding="utf-8",
85)
86 print(f"[DEPLOY] Modèle activé (current): {model_filename}")
87 else:
88 print(
89 "[DEPLOY] Refusé par le gate : F1 insuffisante "
90 "ou baseline non battue."
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/sho
w> 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



```
```