00_Environment_setup

March 16, 2025

1 MLOps Project: Environment Setup

1.1 Introduction

In this notebook, the development environment will be set up and all necessary tools for the project Binary Classification of Income (over/under 50,000) using the "Adult income dataset" will be installed. This project is built and based on Jupyter Notebooks of the course materials Machine Learning Operations. The provided Jupyter Notebooks serve as a foundational guide, offering structured insights into data processing. By leveraging these notebooks, we ensure that the project follows best practices in data exploration and preprocessing, aligning with the principles taught in the course.

1.2 Lernziele

After completing this notebook, the following will be achieved: - Have a working Python environment with all required packages - Understand the basic project structure - Have initialized a Git repository - Have verified the functionality of all MLOps tools

1.3 1. Environment setup

1.3.1 1.1 Create a virtual environment

These commands will be executed in the terminal:

```
python -m venv mlops-venv
# Unter Windows
.\mlops-venv\Scripts\activate
# Unter Unix/MacOS
source mlops-venv/bin/activate
```

1.3.2 1.2 Install dependencies

The following packages are installed:

[49]:
```
# cell 1: Install the required packages
!pip install numpy pandas scikit-learn mlflow pytest fastapi uvicorn␣
 ↪great-expectations docker python-dotenv matplotlib seaborn
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: numpy in c:\programdata\anaconda3\lib\site-
packages (1.26.4)
```

```
Requirement already satisfied: pandas in
c:\users\tanli\appdata\roaming\python\python312\site-packages (2.1.4)
Requirement already satisfied: scikit-learn in
c:\programdata\anaconda3\lib\site-packages (1.5.1)
Requirement already satisfied: mlflow in
c:\users\tanli\appdata\roaming\python\python312\site-packages (2.20.2)
Requirement already satisfied: pytest in c:\programdata\anaconda3\lib\site-
packages (7.4.4)
Requirement already satisfied: fastapi in
c:\users\tanli\appdata\roaming\python\python312\site-packages (0.115.8)
Requirement already satisfied: uvicorn in
c:\users\tanli\appdata\roaming\python\python312\site-packages (0.34.0)
Requirement already satisfied: great-expectations in
c:\users\tanli\appdata\roaming\python\python312\site-packages (1.3.6)
Requirement already satisfied: docker in
c:\users\tanli\appdata\roaming\python\python312\site-packages (7.1.0)
Requirement already satisfied: python-dotenv in
c:\programdata\anaconda3\lib\site-packages (0.21.0)
Requirement already satisfied: matplotlib in c:\programdata\anaconda3\lib\site-
packages (3.9.2)
Requirement already satisfied: seaborn in c:\programdata\anaconda3\lib\site-
packages (0.13.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from pandas)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in
c:\programdata\anaconda3\lib\site-packages (from pandas) (2024.1)
Requirement already satisfied: tzdata>=2022.1 in
c:\programdata\anaconda3\lib\site-packages (from pandas) (2023.3)
Requirement already satisfied: scipy>=1.6.0 in
c:\programdata\anaconda3\lib\site-packages (from scikit-learn) (1.13.1)
Requirement already satisfied: joblib>=1.2.0 in
c:\programdata\anaconda3\lib\site-packages (from scikit-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=3.1.0 in
c:\programdata\anaconda3\lib\site-packages (from scikit-learn) (3.5.0)
Requirement already satisfied: mlflow-skinny==2.20.2 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from mlflow)
(2.20.2)
Requirement already satisfied: Flask<4 in c:\programdata\anaconda3\lib\site-
packages (from mlflow) (3.0.3)
Requirement already satisfied: Jinja2<4,>=3.0 in
c:\programdata\anaconda3\lib\site-packages (from mlflow) (3.1.4)
Requirement already satisfied: alembic!=1.10.0,<2 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from mlflow)
(1.14.1)
Requirement already satisfied: graphene<4 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from mlflow)
(3.4.3)
```

```
Requirement already satisfied: markdown<4,>=3.3 in
c:\programdata\anaconda3\lib\site-packages (from mlflow) (3.4.1)
Requirement already satisfied: pyarrow<19,>=4.0.0 in
c:\programdata\anaconda3\lib\site-packages (from mlflow) (16.1.0)
Requirement already satisfied: sqlalchemy<3,>=1.4.0 in
c:\programdata\anaconda3\lib\site-packages (from mlflow) (2.0.34)
Requirement already satisfied: waitress<4 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from mlflow)
(3.0.2)
Requirement already satisfied: cachetools<6,>=5.0.0 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(5.3.3)
Requirement already satisfied: click<9,>=7.0 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(8.1.7)
Requirement already satisfied: cloudpickle<4 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(3.0.0)
Requirement already satisfied: databricks-sdk<1,>=0.20.0 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from mlflow-
skinny==2.20.2->mlflow) (0.44.1)
Requirement already satisfied: gitpython<4,>=3.1.9 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(3.1.43)
Requirement already satisfied: importlib_metadata!=4.7.0,<9,>=3.7.0 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(7.0.1)
Requirement already satisfied: opentelemetry-api<3,>=1.9.0 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from mlflow-
skinny==2.20.2->mlflow) (1.30.0)
Requirement already satisfied: opentelemetry-sdk<3,>=1.9.0 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from mlflow-
skinny==2.20.2->mlflow) (1.30.0)
Requirement already satisfied: packaging<25 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from mlflow-
skinny==2.20.2->mlflow) (23.2)
Requirement already satisfied: protobuf<6,>=3.12.0 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(4.25.3)
Requirement already satisfied: pydantic<3,>=1.10.8 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(2.8.2)
Requirement already satisfied: pyyaml<7,>=5.1 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(6.0.1)
Requirement already satisfied: requests<3,>=2.17.3 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(2.32.3)
```

```
Requirement already satisfied: sqlparse<1,>=0.4.0 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from mlflow-
skinny==2.20.2->mlflow) (0.5.3)
Requirement already satisfied: typing-extensions<5,>=4.0.0 in
c:\programdata\anaconda3\lib\site-packages (from mlflow-skinny==2.20.2->mlflow)
(4.11.0)
Requirement already satisfied: iniconfig in c:\programdata\anaconda3\lib\site-
packages (from pytest) (1.1.1)
Requirement already satisfied: pluggy<2.0,>=0.12 in
c:\programdata\anaconda3\lib\site-packages (from pytest) (1.0.0)
Requirement already satisfied: colorama in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from pytest)
(0.4.6)
Requirement already satisfied: starlette<0.46.0,>=0.40.0 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from fastapi)
(0.45.3)
Requirement already satisfied: h11>=0.8 in c:\programdata\anaconda3\lib\site-
packages (from uvicorn) (0.14.0)
Requirement already satisfied: altair<5.0.0,>=4.2.1 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from great-
expectations) (4.2.2)
Requirement already satisfied: cryptography>=3.2 in
c:\programdata\anaconda3\lib\site-packages (from great-expectations) (43.0.0)
Requirement already satisfied: jsonschema>=2.5.1 in
c:\programdata\anaconda3\lib\site-packages (from great-expectations) (4.23.0)
Requirement already satisfied: marshmallow<4.0.0,>=3.7.1 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from great-
expectations) (3.26.1)
Requirement already satisfied: mistune>=0.8.4 in
c:\programdata\anaconda3\lib\site-packages (from great-expectations) (2.0.4)
Requirement already satisfied: posthog<4,>3 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from great-
expectations) (3.13.0)
Requirement already satisfied: pyparsing>=2.4 in
c:\programdata\anaconda3\lib\site-packages (from great-expectations) (3.1.2)
Requirement already satisfied: ruamel.yaml>=0.16 in
c:\programdata\anaconda3\lib\site-packages (from great-expectations) (0.18.6)
Requirement already satisfied: tqdm>=4.59.0 in
c:\programdata\anaconda3\lib\site-packages (from great-expectations) (4.66.5)
Requirement already satisfied: tzlocal>=1.2 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from great-
expectations) (5.3)
Requirement already satisfied: pywin32>=304 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from docker)
(306)
Requirement already satisfied: urllib3>=1.26.0 in
c:\programdata\anaconda3\lib\site-packages (from docker) (2.2.3)
Requirement already satisfied: contourpy>=1.0.1 in
```

```
c:\programdata\anaconda3\lib\site-packages (from matplotlib) (1.2.0)
Requirement already satisfied: cycler>=0.10 in
c:\programdata\anaconda3\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in
c:\programdata\anaconda3\lib\site-packages (from matplotlib) (4.51.0)
Requirement already satisfied: kiwisolver>=1.3.1 in
c:\programdata\anaconda3\lib\site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: pillow>=8 in c:\programdata\anaconda3\lib\site-
packages (from matplotlib) (10.4.0)
Requirement already satisfied: Mako in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from
alembic!=1.10.0,<2->mlflow) (1.3.9)
Requirement already satisfied: entrypoints in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from
altair<5.0.0,>=4.2.1->great-expectations) (0.4)
Requirement already satisfied: toolz in c:\programdata\anaconda3\lib\site-
packages (from altair<5.0.0,>=4.2.1->great-expectations) (0.12.0)
Requirement already satisfied: cffi>=1.12 in c:\programdata\anaconda3\lib\site-
packages (from cryptography>=3.2->great-expectations) (1.17.1)
Requirement already satisfied: Werkzeug>=3.0.0 in
c:\programdata\anaconda3\lib\site-packages (from Flask<4->mlflow) (3.0.3)
Requirement already satisfied: itsdangerous>=2.1.2 in
c:\programdata\anaconda3\lib\site-packages (from Flask<4->mlflow) (2.2.0)
Requirement already satisfied: blinker>=1.6.2 in
c:\programdata\anaconda3\lib\site-packages (from Flask<4->mlflow) (1.6.2)
Requirement already satisfied: graphql-core<3.3,>=3.1 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from
graphene<4->mlflow) (3.2.6)
Requirement already satisfied: graphql-relay<3.3,>=3.1 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from
graphene<4->mlflow) (3.2.0)
Requirement already satisfied: MarkupSafe>=2.0 in
c:\programdata\anaconda3\lib\site-packages (from Jinja2<4,>=3.0->mlflow) (2.1.3)
Requirement already satisfied: attrs>=22.2.0 in
c:\programdata\anaconda3\lib\site-packages (from jsonschema>=2.5.1->great-
expectations) (23.1.0)
Requirement already satisfied: jsonschema-specifications>=2023.03.6 in
c:\programdata\anaconda3\lib\site-packages (from jsonschema>=2.5.1->great-
expectations) (2023.7.1)
Requirement already satisfied: referencing>=0.28.4 in
c:\programdata\anaconda3\lib\site-packages (from jsonschema>=2.5.1->great-
expectations) (0.30.2)
Requirement already satisfied: rpds-py>=0.7.1 in
c:\programdata\anaconda3\lib\site-packages (from jsonschema>=2.5.1->great-
expectations) (0.10.6)
Requirement already satisfied: six>=1.5 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from
posthog<4,>3->great-expectations) (1.16.0)
```

```
Requirement already satisfied: monotonic>=1.5 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from
posthog<4,>3->great-expectations) (1.6)
Requirement already satisfied: backoff>=1.10.0 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from
posthog<4,>3->great-expectations) (2.2.1)
Requirement already satisfied: annotated-types>=0.4.0 in
c:\programdata\anaconda3\lib\site-packages (from pydantic<3,>=1.10.8->mlflow-
skinny==2.20.2->mlflow) (0.6.0)
Requirement already satisfied: pydantic-core==2.20.1 in
c:\programdata\anaconda3\lib\site-packages (from pydantic<3,>=1.10.8->mlflow-
skinny==2.20.2->mlflow) (2.20.1)
Requirement already satisfied: charset-normalizer<4,>=2 in
c:\programdata\anaconda3\lib\site-packages (from requests<3,>=2.17.3->mlflow-
skinny==2.20.2->mlflow) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in
c:\programdata\anaconda3\lib\site-packages (from requests<3,>=2.17.3->mlflow-
skinny==2.20.2->mlflow) (3.7)
Requirement already satisfied: certifi>=2017.4.17 in
c:\programdata\anaconda3\lib\site-packages (from requests<3,>=2.17.3->mlflow-
skinny==2.20.2->mlflow) (2024.8.30)
Requirement already satisfied: ruamel.yaml.clib>=0.2.7 in
c:\programdata\anaconda3\lib\site-packages (from ruamel.yaml>=0.16->great-
expectations) (0.2.8)
Requirement already satisfied: greenlet!=0.4.17 in
c:\programdata\anaconda3\lib\site-packages (from sqlalchemy<3,>=1.4.0->mlflow)
(3.0.1)
Requirement already satisfied: anyio<5,>=3.6.2 in
c:\programdata\anaconda3\lib\site-packages (from
starlette<0.46.0,>=0.40.0->fastapi) (4.2.0)
Requirement already satisfied: sniffio>=1.1 in
c:\programdata\anaconda3\lib\site-packages (from
anyio<5,>=3.6.2->starlette<0.46.0,>=0.40.0->fastapi) (1.3.0)
Requirement already satisfied: pycparser in c:\programdata\anaconda3\lib\site-
packages (from cffi>=1.12->cryptography>=3.2->great-expectations) (2.21)
Requirement already satisfied: google-auth~=2.0 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from databricks-
sdk<1,>=0.20.0->mlflow-skinny==2.20.2->mlflow) (2.38.0)
Requirement already satisfied: gitdb<5,>=4.0.1 in
c:\programdata\anaconda3\lib\site-packages (from gitpython<4,>=3.1.9->mlflow-
skinny==2.20.2->mlflow) (4.0.7)
Requirement already satisfied: zipp>=0.5 in c:\programdata\anaconda3\lib\site-
packages (from importlib_metadata!=4.7.0,<9,>=3.7.0->mlflow-
skinny==2.20.2->mlflow) (3.17.0)
Requirement already satisfied: deprecated>=1.2.6 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from
opentelemetry-api<3,>=1.9.0->mlflow-skinny==2.20.2->mlflow) (1.2.18)
Requirement already satisfied: opentelemetry-semantic-conventions==0.51b0 in
```

```
c:\users\tanli\appdata\roaming\python\python312\site-packages (from
opentelemetry-sdk<3,>=1.9.0->mlflow-skinny==2.20.2->mlflow) (0.51b0)
Requirement already satisfied: wrapt<2,>=1.10 in
c:\programdata\anaconda3\lib\site-packages (from
deprecated>=1.2.6->opentelemetry-api<3,>=1.9.0->mlflow-skinny==2.20.2->mlflow)
(1.14.1)
Requirement already satisfied: smmap<5,>=3.0.1 in
c:\programdata\anaconda3\lib\site-packages (from
gitdb<5,>=4.0.1->gitpython<4,>=3.1.9->mlflow-skinny==2.20.2->mlflow) (4.0.0)
Requirement already satisfied: pyasn1-modules>=0.2.1 in
c:\programdata\anaconda3\lib\site-packages (from google-auth~=2.0->databricks-
sdk<1,>=0.20.0->mlflow-skinny==2.20.2->mlflow) (0.2.8)
Requirement already satisfied: rsa<5,>=3.1.4 in
c:\users\tanli\appdata\roaming\python\python312\site-packages (from google-
auth~=2.0->databricks-sdk<1,>=0.20.0->mlflow-skinny==2.20.2->mlflow) (4.9)
Requirement already satisfied: pyasn1<0.5.0,>=0.4.6 in
c:\programdata\anaconda3\lib\site-packages (from pyasn1-modules>=0.2.1->google-
auth~=2.0->databricks-sdk<1,>=0.20.0->mlflow-skinny==2.20.2->mlflow) (0.4.8)
```

1.3.3 1.3 Checking installation

To verify if all packages are correctly installed, you can use the following code snippet in your Jupyter notebook. This code will attempt to import the necessary packages and print a success message if all imports are successful:

[3]:
```python
# Zelle 2: Import und Versionscheck
import sys
import numpy as np
import pandas as pd
import mlflow
import great_expectations as ge
from fastapi import FastAPI
import pytest
from sklearn.impute import SimpleImputer

# Versionen ausgeben
print(f"Python Version: {sys.version}")
print(f"NumPy Version: {np.__version__}")
print(f"Pandas Version: {pd.__version__}")
print(f"MLflow Version: {mlflow.__version__}")
print(f"Great Expectations Version: {ge.__version__}")
```

```
Python Version: 3.12.7 | packaged by Anaconda, Inc. | (main, Oct  4 2024,
13:17:27) [MSC v.1929 64 bit (AMD64)]
NumPy Version: 1.26.4
Pandas Version: 2.1.4
MLflow Version: 2.20.2
Great Expectations Version: 1.3.6
```

1.4 2. Projekt structure

The project structure will be as follows:

```
adult_income/
  data/
      raw/
      processed/
  notebooks/
      00_Umgebung_Einrichtung.ipynb
      01_Daten_Exploration.ipynb
      02_Daten_Vorverarbeitung.ipynb
      03_Modell_Engineering.ipynb
      04_Modell_Deployment.ipynb
  src/
      data/
      features/
      models/
      api/
  tests/
  .gitignore
  README.md
  requirements.txt
```

The structure is created:

[4]:
```python
# Zelle 3: Projektstruktur erstellen
import os

def create_project_structure():
    # Verzeichnisstruktur definieren
    directories = [
        'data/raw',
        'data/processed',
        'notebooks',
        'src/data',
        'src/features',
        'src/models',
        'src/api',
        'tests'
    ]

    # Verzeichnisse erstellen
    for dir_path in directories:
        os.makedirs(dir_path, exist_ok=True)
        print(f"Verzeichnis erstellt: {dir_path}")

create_project_structure()
```

```
Verzeichnis erstellt: data/raw
```

```
Verzeichnis erstellt: data/processed
Verzeichnis erstellt: notebooks
Verzeichnis erstellt: src/data
Verzeichnis erstellt: src/features
Verzeichnis erstellt: src/models
Verzeichnis erstellt: src/api
Verzeichnis erstellt: tests
```

1.5 3. Git setup

1.5.1 3.1 Git Repository initialising

The commands are executed in terminal:

```
git init
```

1.5.2 3.2 create .gitignore

.gitignore file is created:

[7]:
```
# Zelle 4: .gitignore erstellen
gitignore_content = """
# Python
__pycache__/
*.py[cod]
*$py.class
*.so
.Python
env/
build/
develop-eggs/
dist/
downloads/
eggs/
.eggs/
lib/
lib64/
parts/
sdist/
var/
*.egg-info/
.installed.cfg
*.egg

# Virtuelle Umgebung
mlops-venv/
venv/
ENV/
```

```
# Jupyter Notebook
.ipynb_checkpoints

# MLflow
mlruns/

# Daten
data/raw/*
data/processed/*
!data/raw/.gitkeep
!data/processed/.gitkeep

# IDE
.idea/
.vscode/
"""

with open('.gitignore', 'w') as f:
    f.write(gitignore_content)
print(".gitignore Datei wurde erstellt")
```

```
.gitignore Datei wurde erstellt
```

1.6 4. Download dataset

The dataset Adult Income is downloaded from here: https://www.kaggle.com/datasets/wenruliu/adult-income-dataset/data

[1]:
```
#https://raw.githubusercontent.com/vladislabv/fhswf-mlops-project/refs/heads/
 ↪1_Datenverarbeitung_lip/adult.csv

# Zelle 5: Datensatz herunterladen
import pandas as pd

url = "https://raw.githubusercontent.com/fhswf-study-projects/
 ↪mlops-data-processor/refs/heads/Datenexploration/adult.csv"
df = pd.read_csv(url)
df.to_csv('data/raw/adult-income.csv', index=False)
print("The dataset was downloaded and saved in data/raw/adult-income.csv")
```

```
---------------------------------------------------------------------------
OSError                                   Traceback (most recent call last)
Cell In[1], line 8
      6 url = "https://raw.githubusercontent.com/fhswf-study-projects/
 ↪mlops-data-processor/refs/heads/Datenexploration/adult.csv"
      7 df = pd.read_csv(url)
----> 8 df.to_csv('data/raw/adult-income.csv', index=False)
```

```
      9 print("The dataset was downloaded and saved in data/raw/adult-income.
 ↪csv")


File ~\AppData\Roaming\Python\Python312\site-packages\pandas\core\generic.py:
 ↪3902, in NDFrame.to_csv(self, path_or_buf, sep, na_rep, float_format, columns␣
 ↪header, index, index_label, mode, encoding, compression, quoting, quotechar,␣
 ↪lineterminator, chunksize, date_format, doublequote, escapechar, decimal,␣
 ↪errors, storage_options)
   3891 df = self if isinstance(self, ABCDataFrame) else self.to_frame()
   3893 formatter = DataFrameFormatter(
   3894     frame=df,
   3895     header=header,
   (…)
   3899     decimal=decimal,
   3900 )
-> 3902 return DataFrameRenderer(formatter).to_csv(
   3903     path_or_buf,
   3904     lineterminator=lineterminator,
   3905     sep=sep,
   3906     encoding=encoding,
   3907     errors=errors,
   3908     compression=compression,
   3909     quoting=quoting,
   3910     columns=columns,
   3911     index_label=index_label,
   3912     mode=mode,
   3913     chunksize=chunksize,
   3914     quotechar=quotechar,
   3915     date_format=date_format,
   3916     doublequote=doublequote,
   3917     escapechar=escapechar,
   3918     storage_options=storage_options,
   3919 )

File ~\AppData\Roaming\Python\Python312\site-packages\pandas\io\formats\format.
 ↪py:1152, in DataFrameRenderer.to_csv(self, path_or_buf, encoding, sep,␣
 ↪columns, index_label, mode, compression, quoting, quotechar, lineterminator,␣
 ↪chunksize, date_format, doublequote, escapechar, errors, storage_options)
   1131     created_buffer = False
   1133 csv_formatter = CSVFormatter(
   1134     path_or_buf=path_or_buf,
   1135     lineterminator=lineterminator,
   (…)
   1150     formatter=self.fmt,
   1151 )
-> 1152 csv_formatter.save()
   1154 if created_buffer:
   1155     assert isinstance(path_or_buf, StringIO)
```

```
File ~\AppData\Roaming\Python\Python312\site-packages\pandas\io\formats\csvs.py
 ↪247, in CSVFormatter.save(self)
    243 """
    244 Create the writer & save.
    245 """
    246 # apply compression and byte/text conversion
--> 247 with get_handle(
    248     self.filepath_or_buffer,
    249     self.mode,
    250     encoding=self.encoding,
    251     errors=self.errors,
    252     compression=self.compression,
    253     storage_options=self.storage_options,
    254 ) as handles:
    255     # Note: self.encoding is irrelevant here
    256     self.writer = csvlib.writer(
    257         handles.handle,
    258         lineterminator=self.lineterminator,
   (…)
    263         quotechar=self.quotechar,
    264     )
    266     self._save()

File ~\AppData\Roaming\Python\Python312\site-packages\pandas\io\common.py:739,␣
 ↪in get_handle(path_or_buf, mode, encoding, compression, memory_map, is_text,␣
 ↪errors, storage_options)
    737 # Only for write methods
    738 if "r" not in mode and is_path:
--> 739     check_parent_directory(str(handle))
    741 if compression:
    742     if compression != "zstd":
    743         # compression libraries do not like an explicit text-mode

File ~\AppData\Roaming\Python\Python312\site-packages\pandas\io\common.py:604,␣
 ↪in check_parent_directory(path)
    602 parent = Path(path).parent
    603 if not parent.is_dir():
--> 604     raise OSError(rf"Cannot save file into a non-existent directory:␣
 ↪'{parent}'")

OSError: Cannot save file into a non-existent directory: 'data\raw'
```

The following columns are given in the dataset:

Age, workclass, fnlwgt, education, educational-num, marital-status, occupation, relationship, race, gender, capital-gain, capital-loss, hours-per-week, native-country, income

[6]:
```
# First, we will get an overview about the dataset

print(df.info())
print(df.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   age              48842 non-null  int64
 1   workclass        48842 non-null  object
 2   fnlwgt           48842 non-null  int64
 3   education        48842 non-null  object
 4   educational-num  48842 non-null  int64
 5   marital-status   48842 non-null  object
 6   occupation       48842 non-null  object
 7   relationship     48842 non-null  object
 8   race             48842 non-null  object
 9   gender           48842 non-null  object
 10  capital-gain     48842 non-null  int64
 11  capital-loss     48842 non-null  int64
 12  hours-per-week   48842 non-null  int64
 13  native-country   48842 non-null  object
 14  income           48842 non-null  object
dtypes: int64(6), object(9)
memory usage: 5.6+ MB
None
   age  workclass  fnlwgt     education  educational-num      marital-status  \
0   25    Private  226802          11th                7       Never-married
1   38    Private   89814       HS-grad                9  Married-civ-spouse
2   28  Local-gov  336951    Assoc-acdm               12  Married-civ-spouse
3   44    Private  160323  Some-college               10  Married-civ-spouse
4   18          ?  103497  Some-college               10       Never-married

          occupation relationship   race  gender  capital-gain  capital-loss  \
0  Machine-op-inspct    Own-child  Black    Male             0             0
1    Farming-fishing      Husband  White    Male             0             0
2    Protective-serv      Husband  White    Male             0             0
3  Machine-op-inspct      Husband  Black    Male          7688             0
4                  ?    Own-child  White  Female             0             0

   hours-per-week native-country income
0              40  United-States  <=50K
1              50  United-States  <=50K
2              40  United-States   >50K
3              40  United-States   >50K
4              30  United-States  <=50K
```

There are values missing. The missing values are marked with '?' in the dataset. The parameters used in the dataset for adult income prediction are:

- age: the age of an individual
- workclass: a general term to represent the employment status of an individual
- fnlwgt: final weight. This is the number of people the census believes the entry represents..
- education: the highest level of education achieved by an individual.
- education-num: the highest level of education achieved in numerical form.
- marital-status: marital status of an individual.
- occupation: the general type of occupation of an individual
- relationship: represents what this individual is relative to others.
- race: Descriptions of an individual's race
- sex: the sex of the individual
- capital-gain: capital gains for an individual
- capital-loss: capital loss for an individual
- hours per week: the hours an individual has reported to work per week
- native country: country of origin for an individual

In the next step, we will start with data exploration.

[ ]: