Dokumentacja przetwarzania danych i implementacji Airflow DAGs

DAG 1: Podział danych na zbiory treningowy i testowy

Opis ogólny

Pierwszy DAG (dag_data_split) pobiera dane, dzieli je na zbiór treningowy (70%) i testowy (30%), a następnie zapisuje te zbiory do dwóch osobnych arkuszy Google Sheets.

Kroki:

1. **Pobranie danych** – Funkcja download_data wczytuje dane z lokalnego pliku CSV.
   - Używa funkcji pd.read_csv() do wczytania danych z pliku (C:/Users/Michal/Desktop/train.csv).
   - Wartości danych są przesyłane do innych zadań przy użyciu kwargs['ti'].xcom_push().
2. **Podział danych na zbiory** – Funkcja split_data dzieli dane na dwa zbiory:
   - Zbiór treningowy (70%) i zbiór testowy (30%) są tworzone przy pomocy funkcji train_test_split z biblioteki sklearn.
   - Zbiory danych są również przesyłane za pomocą kwargs['ti'].xcom_push().
3. **Zapis danych do Google Sheets** – Funkcja save_to_gsheets zapisuje oba zbiory (treningowy i testowy) do dwóch osobnych arkuszy Google Sheets.
   - Do zapisu wykorzystywana jest biblioteka gspread i klucz autoryzacji OAuth.
   - Zbiór treningowy trafia do arkusza o nazwie Zbior_Modelowy, a zbiór testowy do arkusza Zbior_Douczeniowy.

Workflow (DAG):

- **t1 (download_data)** – Pobiera dane z pliku CSV.
- **t2 (split_data)** – Dzieli dane na zbiór treningowy i testowy.
- **t3 (save_train_to_gsheets)** – Zapisuje zbiór treningowy do Google Sheets.
- **t4 (save_test_to_gsheets)** – Zapisuje zbiór testowy do Google Sheets.

Workflow jest zależny od siebie:
download_data → split_data → save_train_to_gsheets i save_test_to_gsheets.

DAG 2: Przetwarzanie danych – czyszczenie, skalowanie i zapis do Google Sheets

Opis ogólny

Drugi DAG (dag_data_processing) działa na danych, które zostały zapisane wcześniej do Google Sheets. Zadaniem tego DAG-a jest przetworzenie danych poprzez czyszczenie (usuwanie duplikatów i wypełnianie brakujących wartości), a także ich skalowanie (standaryzacja i normalizacja). Po przetworzeniu dane są zapisywane z powrotem do Google Sheets.

Kroki:

1. **Pobranie danych z Google Sheets** – Funkcja fetch_data_from_gsheets pobiera dane z arkusza Zbior_Modelowy znajdującego się w Google Sheets.
    - Używa do tego biblioteki gspread oraz konta usługi (Service Account).
    - Funkcja zwraca dane w formacie Pandas DataFrame.
2. **Czyszczenie danych** – Funkcja clean_data wykonuje dwie operacje:
    - Usuwa duplikaty w danych za pomocą drop_duplicates().
    - Wypełnia brakujące wartości średnią kolumny przy użyciu fillna(data.mean()).
3. **Skalowanie danych** – Funkcja scale_data wykonuje dwa etapy:
    - **Standaryzacja**: Dane są skalowane do rozkładu normalnego (średnia = 0, odchylenie standardowe = 1) przy użyciu StandardScaler z sklearn.
    - **Normalizacja**: Dane są skalowane w zakresie [0, 1] za pomocą MinMaxScaler z sklearn.
4. **Zapis przetworzonych danych do Google Sheets** – Funkcja save_to_gsheets zapisuje przetworzone dane do nowego arkusza w Google Sheets o nazwie Processed_Data.

Workflow (DAG):

- **t1 (fetch_data)** – Pobiera dane z arkusza Zbior_Modelowy w Google Sheets.
- **t2 (clean_data)** – Czyści dane (usuwa duplikaty, wypełnia brakujące wartości).
- **t3 (scale_data)** – Skalowanie i normalizacja danych.
- **t4 (save_processed_data)** – Zapisuje przetworzone dane do nowego arkusza Processed_Data.

Workflow w tym DAG-u jest również zależny od siebie:
fetch_data → clean_data → scale_data → save_processed_data.