📘 WEEK 2 Documentation: Pollution Drift Predictor

🧠 Objective

The goal for Week 2 was to implement a machine learning model that predicts pollution drift patterns using environmental data. The focus was on selecting a regression algorithm, preprocessing the data, training the model, evaluating its performance, and visualizing the results.

⚙️ Model Implementation

🔍 Algorithm Used

Linear Regression from scikit-learn was chosen for its simplicity and interpretability as a baseline model.

🧪 Features and Target

- so2: Sulfur Dioxide concentration
- no2: Nitrogen Dioxide concentration
- spm: Suspended Particulate Matter (target variable)

🧹 Preprocessing

- Dropped rows with missing values in so2, no2, and spm
- Selected so2 and no2 as input features
- Applied StandardScaler to normalize the features
- Used fit_transform() on training data and transform() on test data
- Saved both the trained model and scaler using joblib for Week 3 deployment

🧠 Training Logic

The dataset was split into training and test sets using an 80/20 ratio. The features were scaled using StandardScaler, and the model was trained on the scaled data. Predictions were made on the test set and evaluated using standard regression metrics.

📊 Model Evaluation

📈 Metrics Used

- R^2 Score: ~0.10 — indicates low explanatory power for this baseline model
- MAE: ~110.01 — average prediction error in SPM units
- MSE: ~21546.16 — penalizes larger errors more heavily

The model shows limited predictive power, suggesting that SO_2 and NO_2 alone may not fully explain SPM variability. This sets the stage for feature engineering and model refinement in Week 3.

📈 Visualizations

1. Actual vs Predicted SPM
   A scatter plot comparing predicted SPM values against actual observations. Most points cluster below the ideal line, indicating underprediction.
2. Residuals Distribution
   A histogram of prediction errors. Residuals are centered around zero but show a left-skewed tail, suggesting the model misses high SPM values.
3. SO_2 vs SPM (colored by NO_2)
   A scatter plot showing the relationship between SO_2 and SPM, with NO_2 levels represented by color. Clustering patterns suggest potential pollutant interactions worth exploring further.

✅ Week 2 Checklist

Task	Status
Implement ML model	✅ Done
Show model structure	✅ Done
Evaluate model accuracy	✅ Done
Visualize predictions	✅ Done
Document findings	✅ Done
Save model and scaler	✅ Done

📦 Artifacts Saved

- linear_regression_model.pkl — trained model

  [Not Uploaded due to File Size Restriction]
- forest_regressor_model.pkl — trained model

  [Not Uploaded due to File Size Restriction]
- scaler.pkl — fitted scaler
- model_metrics.md — evaluation summary
- X_test.csv — test features
- y_test_vs_pred.csv — actual vs predicted values
- actual_vs_predicted.png, residuals.png, scatter_so2_spm.png — visualizations

---