**📘 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. Specifically, the task focused on:

* Selecting and applying a regression algorithm
* Training the model on cleaned data
* Evaluating model performance using standard metrics
* Visualizing predictions and residuals

**⚙️ Model Implementation**

**🔍 Algorithm Used**

* **Linear Regression** from scikit-learn was chosen due to its simplicity and interpretability for baseline modeling.

**🧪 Features and Target**

| **Feature** | **Description** |
| --- | --- |
| 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
* Used spm as the target for prediction

**🧠 Training Logic**

X = df[['so2', 'no2']]

y = df['spm']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LinearRegression()

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

**📊 Model Evaluation**

**📈 Metrics Used**

| **Metric** | **Value** | **Interpretation** |
| --- | --- | --- |
| R² Score | ~0.10 | Low explanatory power — 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 SO₂ and NO₂ 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**

This plot compares predicted SPM values against actual observations. The red line (predicted) shows a smoother trend, while the blue line (actual) reveals more variability.

**2. Residuals Distribution**

The residuals are centered around zero, but the left-skewed tail indicates underprediction in some cases. This suggests the model may be missing key features or nonlinear patterns.

**3. SO₂ vs SPM (colored by NO₂)**

This scatter plot visualizes the relationship between SO₂ and SPM, with NO₂ 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 |