**📘 WEEK 1 Documentation: Pollution Drift Predictor**

**🧠 Problem Statement**

Airborne pollution, especially particulate matter (PM), poses serious health and environmental risks. Construction zones, industrial areas, and urban corridors often experience unpredictable pollution drift due to changing environmental conditions. This project aims to build an AI-powered system that forecasts pollution drift using environmental data such as wind speed, wind direction, humidity, and timestamp.

“Can we predict how pollution spreads across a region based on environmental factors?”

**📂 Dataset Overview**

**📍 Source**

The dataset used is data.csv, located in the /data/ folder of the repository. It contains historical pollution readings from various monitoring stations across Hyderabad, India.

**📊 Key Columns**

| **Column Name** | **Description** |
| --- | --- |
| stn\_code | Station identifier |
| sampling\_date | Month and year of sampling |
| state, location, type | Geographical and zone metadata |
| so2, no2, spm, pm2\_5 | Pollutant concentrations (some missing values) |
| date | Standardized timestamp |

**🧹 Preprocessing Steps**

* Dropped rows with missing values in key pollutant columns (so2, no2, spm)
* Selected so2 and no2 as features, and spm as the target variable
* Converted date to datetime format for future time-based analysis

**🎯 ML Objective**

To build a regression model that predicts **SPM (Suspended Particulate Matter)** levels based on **SO₂** and **NO₂** concentrations.

This aligns with the Week 1 goal of:

* Defining a clear ML problem
* Preparing a dataset suitable for modeling
* Identifying relevant features and target variables

**🧰 Tools & Technologies**

| **Tool** | **Purpose** |
| --- | --- |
| Python 3.11 | Core programming language |
| Pandas, NumPy | Data manipulation and cleaning |
| Scikit-learn | ML model training and evaluation |
| Matplotlib, Seaborn | Visualization and plotting |
| Git & GitHub | Version control and collaboration |

**🧪 Sample Data Snapshot**

stn\_code,sampling\_date,state,location,type,so2,no2,spm,date

150,February - M021990,Andhra Pradesh,Hyderabad,Residential,4.8,17.4,NA,1990-02-01

151,March - M031990,Andhra Pradesh,Hyderabad,Industrial,4.7,7.5,82,1990-03-01

152,June - M061990,Andhra Pradesh,Hyderabad,Residential,3.3,19.3,111,1990-06-01

more available… (check repository)

**📈 Prepared Features**

After cleaning and selection:

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

target = df['spm']

This structure ensures the model receives numeric inputs and a continuous target variable for regression.

**✅ Week 1 Checklist**

| **Task** | **Status** |
| --- | --- |
| Search dataset related to theme | ✅ Done |
| Define problem statement | ✅ Done |
| Prepare dataset for ML model | ✅ Done |