Weather Forecasting & Climate Analysis Project

Project Overview

This project is part of the PM Accelerator mission to analyze and forecast global weather patterns using a real-world dataset. I applied data cleaning, exploratory analysis, multiple forecasting models, and advanced analytical techniques to uncover patterns in climate and environmental impact.

Dataset

- **Source**: Provided CSV file with weather data across countries
- **Fields**: Timestamp, temperature, humidity, wind speed, air quality (PM2.5, PM10, Ozone, NO_2, SO_2, CO), cloud coverage, UV index, visibility, pressure, precipitation, and more
- **Target**: `temperature_celsius`

---

1. Data Cleaning & Preprocessing

- Verified and confirmed **no missing values** in final modeling features.
- Detected and flagged outliers.
- Normalized/standardized key numerical variables where needed.
- Converted timestamps to datetime and sorted chronologically.
- Consolidated and grouped sub-features under categories (e.g., wind, air quality, temperature)

---

2. Exploratory Data Analysis (EDA)

- **Temperature and Precipitation**:
  - Most locations recorded negligible precipitation.
  - Temperatures mostly range from 5°C to 30°C.
- **Time Trends**:
  - Time series plots revealed local fluctuations.
- **Correlations**:
  - PM2.5 and PM10 highly correlated
  - Humidity inversely related to temperature

- **Anomaly Detection**:
  - Outliers in PM2.5, temperature, wind speed identified

---

3. Forecasting Models

- **Univariate Models**:
  - Linear Regression
  - ARIMA
- **Multivariate Models**:
  - XGBoost using full weather-related features
  - LSTM trained on temporal sequences of all available numerical inputs
- **Ensemble Learning (Stacking)**:
  - Combined predictions from ARIMA, XGBoost, and Prophet-style forecast
  - Final model trained using Linear Regression on prediction outputs

Evaluation Metrics:

Model	MAE	RMSE	R^2
Linear Reg	Medium	Low	Poor
ARIMA	~10	~13	-1.47
XGBoost	Lower MAE & RMSE	Good R^2	
LSTM	Sequence-aware, better temporal fit		
Ensemble	Best overall accuracy		

---

4. Advanced Analyses

Climate Analysis

- Country-level grouping of mean, min, max, std of temperature
- Demonstrated latitude dependence of climate

Environmental Impact

- Correlation analysis between air quality indices and meteorological parameters:
  - PM2.5 and PM10 levels increase with lower humidity and weak winds

    - UV index correlated with ozone and clear skies

Feature Importance

- Applied **XGBoost** to assess importance of all 25+ numerical weather features
- Most influential features:
    - Feels-like temperature
    - Humidity
    - Air Quality PM2.5
    - Wind degree
    - Visibility

Spatial & Geographical Patterns

- Aggregated country-wise averages for temperature, humidity, wind
- (Optional) Latitude-temperature scatter plot (may be sampled due to memory)
- Countries grouped by climate zones using average temperature/humidity

---

Deliverables

- Jupyter Notebook (`EDA.ipynb`) with data prep, EDA, modeling, evaluation
- Outlier and anomaly analysis table
- Feature importance charts and correlation heatmaps
- Final ensemble prediction results