**Project Report: Weather Forecasting and Anomaly Detection using Machine Learning**

**1. Introduction**

This project aims to analyze historical weather data, forecast future trends using machine learning models (XGBoost and ARIMA), and detect anomalies through advanced exploratory analysis and statistical methods. The dataset was cleaned, engineered, and analyzed through multiple phases to uncover meaningful insights that could aid environmental monitoring and forecasting efforts.

**2. Data Cleaning & Preparation**

* **Source**: Cleaned dataset with timestamped weather features.
* **Null Values**: Removed or imputed using forward fill, backward fill, or mean substitution based on the context.
* **Date-Time Processing**: Converted to datetime objects and sorted for time-series analysis.
* **Categorical Features**: One-hot encoded (condition\_text, air\_quality\_\*).
* **Lag Features**: Created lagged variables for temporal dependencies (e.g., precip\_mm\_lag\_1, precip\_mm\_lag\_7).

**3. Exploratory Data Analysis (EDA)**

**Summary Statistics and Distribution Analysis**

**Temperature and Wind Extremes**

* **Temperature Range**: -24.9°C to 49.2°C
* **Wind Speed (mph)**: Max = 1841.2 (anomalous)
* **Wind Speed (kph)**: Max = 2963.2 (anomalous)

**Pressure & Humidity**

* **Pressure (mb)**: Range 947 to 3006 (abnormal; typical 980–1050)
* **Humidity (%)**: Range 2 to 100 (valid)

**Visibility and UV Index**

* **Visibility (km)**: 0 to 32 (expected)
* **UV Index**: Up to 16.3 (high; verify source)

**Air Quality Data Issues**

* air\_quality\_Carbon\_Monoxide, Sulphur\_dioxide, PM10 had extreme negative values (e.g., -9999) and were replaced or ignored.

**Correlation Analysis**

* Strong correlations found:
  + temperature\_celsius ↔ temperature\_fahrenheit: **0.9999**
  + wind\_mph ↔ wind\_kph: **0.9999**
  + gust\_mph ↔ gust\_kph: **0.9999**
  + feels\_like\_celsius ↔ feels\_like\_fahrenheit: **0.9999**
  + visibility\_km ↔ visibility\_miles: **0.9999**
* Decision: Drop redundant features to avoid multicollinearity.
* **Distributions**: Histograms, box plots to understand skewness and outliers.
* **Time Series Trends**: Plots of variables like temp\_c, humidity, wind\_kph, visibility\_km, air\_quality\_PM2.5, etc.
* **Correlations**: Heatmap among top numerical features showed strong relationships:
  + Positive: temp\_c ↔ humidity, uv\_index.
  + Negative: wind\_power ↔ air\_quality\_PM10.
* **Feature Interactions**: Detected seasonal and conditional patterns (e.g., rain, fog impact).

## ****4. Feature Engineering****

### ****Lag Features****

* Created lag features for:
  + Precipitation (1-day, 7-day)
  + Wind power
  + Air quality metrics

### ****Interaction Features****

* Derived:
  + temp\_diff = feels\_like\_celsius - temperature\_celsius
  + wind\_power = wind\_kph × gust\_kph

### ****Encoding****

* One-hot encoded:
  + condition\_text
  + air\_quality\_us-epa-index
  + air\_quality\_gb-defra-index

**5. Forecasting Models**

**🧠 XGBoost Regressor**

* **Target**: temp\_c, precip\_mm, etc.
* **Performance**:
  + MAE: ~0.83
  + RMSE: ~1.02
  + R²: ~0.89
* **Feature Importance**: humidity, uv\_index, dew\_point, wind\_kph.

**📉 ARIMA Model**

* **Target**: Primarily temp\_c
* **Stationarity**: Achieved after differencing.
* **Performance**:
  + RMSE: ~2.35
  + Not as accurate as XGBoost but interpretable.
* **Insight**: XGBoost outperforms ARIMA in both accuracy and adaptability.

**6. Advanced EDA & Anomaly Detection**

**Z-Score Method**

* Used to detect outliers in numerical features using a threshold of ±3.
* **Top Anomaly-Prone Features**:
  + visibility\_km, wind\_power, air\_quality\_PM2.5, precip\_mm\_lag\_1.

**Isolation Forest**

* Used to assign anomaly scores and flags.
* ~583 anomalies detected out of ~58,000 samples.
* **Overlay Plots**: Time series of wind\_power, visibility\_km with red points for anomalies.
* **Red Point Missing Case**: Some anomalies not visible due to NaN in specific features.

**7. Key Insights**

* Weather variables show distinct periodicity and clustering by condition.
* wind\_power and air\_quality\_PM2.5 are critical indicators for both forecasting and anomaly detection.
* XGBoost is robust and recommended for future weather prediction applications.
* Anomaly detection is vital for understanding environmental changes (e.g., sudden pollution or visibility drops).

**8. Files & Structure**

* 01\_data\_exploration.ipynb: Initial cleaning and EDA
* 02\_feature\_eng.ipynb: Feature creation and transformation
* 03\_model\_training.ipynb: ML models and evaluation
* 04\_Adv.EDA\_anomaly.ipynb: Advanced EDA and anomaly detection