Weather Forecasting Project

1. Objective

To forecast future weather conditions using global weather data and demonstrate data science capabilities through comprehensive analysis and modeling. This includes:

- Data cleaning and preprocessing
- Exploratory Data Analysis (EDA)
- Forecasting using ARIMA and LSTM
- Advanced analyses like anomaly detection, spatial visualization, and feature importance

2. Data Cleaning And Pre Processing

1. Null Values: No missing values were found.
2. Outliers: Detected using IQR and visualized with boxplots. Handled in key features like temperature, humidity, and air quality.
3. Normalization: MinMaxScaler applied to scale numerical features.
4. Categorical Encoding:
    a. Target column condition_text grouped into 6 major weather types (e.g., Sunny, Rainy, Snowy).
    b. Categorical features like wind_direction, moon_phase encoded using LabelEncoder.

3. Exploratory Data Analysis

- Visualizations:
  Temperature and Precipitation trends using line and box plot
  Correlation heatmap of key features

Insights:

- Strong correlation between humidity, precipitation, and condition type
- Seasonal temperature variations visible by country and region

4. Fore Casting Models

📈 ARIMA (AutoRegressive Integrated Moving Average)

- Used for time series forecasting of temperature
- Parameters: (5,1,0) optimized based on AIC
- Forecasted temperature with decent accuracy

🤖 LSTM (Long Short-Term Memory)

- Neural network model for sequential data
- Trained on normalized temperature values
- Outperformed ARIMA in capturing nonlinear patterns

🔄 Ensemble Model

- Combined ARIMA and LSTM predictions (average-based ensemble)
- Provided smoother and more accurate temperature forecast

Advanced Analysis:

✅ Anomaly Detection

- Identified extreme or unusual weather days using Z-score and IQR methods
- Helpful for spotting climate disruptions

📌 Climate Analysis

- Regional trends showed tropical regions had consistently higher humidity and precipitation
- Temperate zones had higher variability in conditions

🌫️ Environmental Impact

- High PM2.5, CO levels correlated with low visibility and cloudy/foggy conditions
- Air quality significantly impacted by location and temperature

Challenges:

- Dealing with many granular condition categories (reduced to 6 meaningful classes)
- Time series modeling required preprocessing datetime data
- Geographic data inconsistency handled via grouping and merging techniques