Weather Trend Forecasting Project

1. PM Accelerator Mission

Introduction

The **PM Accelerator** is an initiative designed to level the playing field for future Product Management (PM) leaders. It provides industry-leading tools, mentorship, and education to individuals from all backgrounds, helping both aspiring and experienced Product Managers (PMs) to advance their careers.

Our Mission

Our mission is to make industry-leading tools and education available to people worldwide, empowering them to become successful PM leaders. By offering access to **industry leaders**, a **strong PM ecosystem**, and the opportunity to learn **AI product management skills**, the PM Accelerator enables PMs to gain the insights and capabilities they need to excel in a rapidly changing environment.

Key Features

- **Access to Tools**: Industry-standard tools for product management.
- **Mentorship**: Introduction to PM industry leaders.
- **AI Product Management Skills**: Learn how to manage AI products effectively.
- **Global Community**: Network with other aspiring and experienced PMs from around the world.

2. Weather Trend Forecasting Project

Objective

The primary goal of this project is to forecast weather trends using a dataset of global weather data. Accurate weather forecasts are essential for various industries such as agriculture, transportation, and energy to make informed decisions. By analyzing weather patterns and building machine learning models, we aim to predict future weather trends with a focus on temperature and precipitation.

3. Data Overview and Cleaning

Dataset Overview

The dataset used for this project is the **Global Weather Repository**, which contains weather data from various cities worldwide. The dataset has over 40 features, such as temperature, humidity, wind speed, and air quality, recorded daily.

Data Cleaning

To ensure the accuracy of the models, we performed the following data cleaning steps:

1. **Handling Missing Values**:
    a. Removed or imputed missing values in the features.
    b. Imputed numerical missing values using the **mean** or **median**.
2. **Removing Outliers**:
    a. Identified and removed any outliers that could skew the analysis.
3. **Normalization and Scaling**:
    a. Applied **Min-Max Scaling** to bring numerical values within a specific range.
    b. Standardized certain features to ensure all models handle them properly.

Feature Engineering

1. Extracted **year**, **month**, **day**, and **hour** from the **timestamp** for time-series forecasting.
2. Encoded **categorical variables** like country and location to numerical values using **label encoding**.

4. Exploratory Data Analysis (EDA)

Initial Data Exploration

We performed an initial exploration of the dataset to understand the distributions and relationships between the features. Key findings from the EDA include:

- **Temperature Trends**: The temperature shows distinct seasonal patterns and fluctuations across different regions.
- **Correlation Analysis**: Features like **humidity**, **wind speed**, and **temperature** are highly correlated, which is expected as weather features tend to interact with one another.

Visualizations

- **Line Plots**: Used to visualize **temperature trends over time**.
- **Heatmaps**: Used to observe correlations between weather and air quality metrics.
- **Boxplots**: Analyzed temperature distributions across different **continents** and **countries** to see regional variations.

5. Model Building

Model Selection

We used two models to forecast weather trends:

1. **Long Short-Term Memory (LSTM)**:
    a. A type of **recurrent neural network** ideal for time series data due to its ability to learn from sequential dependencies.
2. **Random Forest**:
    a. A non-sequential model used to compare against LSTM to evaluate its performance for forecasting.

Model Evaluation

Both models were evaluated using **Root Mean Squared Error (RMSE)**, a standard metric for regression problems.

- **LSTM RMSE**: 0.0659
- **Random Forest RMSE**: 0.0447
    - **Random Forest** showed slightly better performance in terms of RMSE, but both models performed well in predicting future weather trends.

6. Advanced Analysis

Climate Analysis

Using **temperature data** from various countries, we analyzed long-term climate trends. We grouped the data by country and year to compute the **average annual temperature** and observed the following:

- **Increasing Average Temperature**: Most countries showed an upward trend in average annual temperature, indicating possible global warming.
- **Regional Patterns**: Certain countries, particularly those closer to the equator, showed more consistent temperatures year-round, while others had larger seasonal fluctuations.

Visualization: A **line plot** showing the temperature trends over the years for various countries.

Environmental Impact

We performed a correlation analysis between weather variables and air quality indices such as **PM2.5**, **PM10**, **Carbon Monoxide**, and **Ozone** levels. The following insights were observed:

- **Correlation between High Temperatures and Air Pollution**: Areas with higher temperatures often showed elevated levels of air pollutants.
- **Humidity and Pollution**: High humidity seemed to correlate with higher **PM2.5** levels, likely due to limited air circulation.

Visualization: A **heatmap** showing the correlation between weather features and air quality indices.

Feature Importance

We used **Random Forest** to evaluate the importance of different features in predicting the target variable (temperature). The most important features were **wind speed**, **humidity**, and **pressure**.

Additionally, we conducted **permutation feature importance** to determine how changes in feature values affect model performance.

Visualization: A **bar plot** of feature importances based on the Random Forest model.

7. Spatial Analysis

Geographical Patterns

We analyzed geographical patterns by grouping data by **continent** and visualizing the distribution of **temperature** and **precipitation**.

1. **Temperature Distribution**: A **boxplot** was used to visualize the variation in temperature across different continents. This helped identify regions with extreme temperatures.
2. **Precipitation Patterns**: A **heatmap** was generated to visualize regions with the highest precipitation, indicating areas prone to heavy rainfall.

Insights:

- **Northern Hemisphere**: Had significant seasonal temperature variations.
- **Equatorial Regions**: Had more stable temperatures year-round.

8. Conclusion

Summary of Insights

1. **Climate Change**: The analysis revealed a steady increase in average temperatures across countries, indicating global warming trends.
2. **Model Performance**: The **Random Forest** model outperformed the **LSTM** model slightly in terms of RMSE but both performed well in forecasting.
3. **Geographical Trends**: Distinct **regional variations** in weather patterns were observed, with tropical regions showing stable temperatures and temperate zones exhibiting more fluctuation.
4. **Air Quality Impact**: We identified correlations between **weather variables** and **air quality indices**, suggesting that weather patterns directly influence environmental pollution.

Future Work

1. **Improving Model Accuracy**: Further improvements can be made by using ensemble models like **XGBoost** or **Gradient Boosting**.
2. **Expanding the Dataset**: Including more environmental data like soil moisture and agricultural yields could improve the model's forecasting ability.