Weather Data Analysis and Forecasting Report

Mission

By making industry-leading tools and education available to individuals from all backgrounds, we level the playing field for future PM leaders. This is the PM Accelerator motto, as we grant aspiring and experienced PMs what they need most – Access. We introduce you to industry leaders, surround you with the right PM ecosystem, and discover the new world of AI product management skills.

1. Introduction

The purpose of this report is to analyze weather data, focusing on temperature trends, air quality correlations, and geographical variability. Several techniques were used to clean the data, explore trends, and forecast future weather patterns. This analysis also incorporates advanced methods such as anomaly detection, climate pattern analysis, and feature importance assessment.

---

2. Data Overview

The dataset contains the following columns:

- **Country**: Country of the recorded weather data.
- **Location Name**: Specific location within the country.
- **Latitude, Longitude**: Coordinates for each location.
- **Temperature**: Both Celsius and Fahrenheit temperature readings.
- **Air Quality**: Various air quality measures, such as CO, Ozone, and PM2.5.
- **Humidity**: Humidity level in percentage.
- **Precipitation**: Precipitation levels in millimeters.
- **Wind Speed**: Wind speed in miles per hour.
- **Sunrise/Sunset**: Times for daily sunrise and sunset.
- **Other features**: Time of last update, visibility, and UV index.

---

3. Data Cleaning

Data cleaning was performed to ensure that the dataset was ready for analysis. The steps include:

- **Handling Missing Values**: Columns with more than 50% missing data were dropped. For columns with numerical values, missing values were filled using the median of the respective columns.

- **Outlier Detection and Handling**: Outliers were detected using the Interquartile Range (IQR) method. Extreme values in temperature and humidity were adjusted to within the acceptable range.
- **Normalization**: Numerical features such as temperature and humidity were normalized using MinMaxScaler to ensure a consistent scale across all features.

---

4. Exploratory Data Analysis (EDA)

4.1 Temperature Distribution

A histogram was created to visualize the distribution of temperature values across all locations. The distribution showed a bell-shaped curve, indicating a normal distribution for temperature values.

4.2 Correlation Analysis

A correlation matrix was calculated to explore the relationships between different weather features. The heatmap revealed strong correlations between temperature and humidity, indicating that these two variables often vary together.

---

5. Forecasting Models

5.1 ARIMA Model

An ARIMA model was used to forecast future temperatures based on historical data. The model was trained on the temperature data, and predictions for the next 10 time steps were made.

Forecasted temperatures for the next 10 days:

ARIMA Forecast: 39064 0.381586

39065 0.354767

39066 0.411244

39067 0.399069

39068 0.411835

39069 0.404441

39070 0.393521

39071 0.394086

39072 0.402432

39073 0.4012755.2

Random Forest & XGBoost Models

Random Forest and XGBoost models were trained to predict temperature based on features such as humidity, precipitation, wind speed, and UV index. These models were evaluated using R-squared and Mean Absolute Error (MAE) metrics.

Model Evaluation:

- **Random Forest**:
    - R-squared: 0.6438097782239218
    - MAE: 0.07914872992168548
- **XGBoost**:
    - R-squared: 0.674067935379974
    - MAE: 0.07884101322309223

5.3 Feature Importance

The importance of different features in predicting temperature was evaluated using XGBoost's feature importance plot. The most important features included humidity, precipitation, and wind speed.

---

6. Advanced Analyses

6.1 Anomaly Detection

An Isolation Forest algorithm was used to detect anomalies in the data. The anomalies, particularly in temperature and humidity, were highlighted in the plot, revealing unusual weather events or outlier data points.

6.2 Climate Pattern Analysis

Temperature trends were analyzed across countries, and the average temperature was calculated for each country. The top countries with the highest and lowest average temperatures were visualized.

6.3 Geographical Patterns

A geographical heatmap was created using latitude and longitude coordinates, showing how temperature varies across different regions of the world. The heatmap highlights regions with higher and lower temperatures.

---

7. Results and Insights

- The analysis revealed a clear correlation between temperature and humidity across different regions. Locations with higher humidity tend to have slightly lower temperatures.
- The ARIMA model provided reasonable forecasts for future temperatures, though it may benefit from incorporating additional features like air quality.
- The Random Forest and XGBoost models showed that humidity and wind speed were significant predictors of temperature variations.
- Anomalies detected in the data suggest unusual weather patterns or errors in data collection.
- The geographical heatmap revealed temperature clusters, with tropical regions experiencing higher temperatures compared to temperate zones.

---

8. Conclusion

This analysis provides valuable insights into weather patterns, temperature forecasting, and air quality analysis. The combination of statistical methods and machine learning models enables a robust understanding of the factors influencing weather conditions across different regions. The insights can be used to better prepare for weather-related events and improve environmental awareness.

---

9. Future Work

- Incorporate more granular weather data, such as daily or hourly readings, to improve forecast accuracy.
- Use additional machine learning models, such as neural networks, for improved predictive power.
- Expand the geographical analysis to include more countries and regions for broader insights.