PMAccelerator 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.

Global Weather Repository Analysis Report:

1. Introduction

This report provides an in-depth analysis of the Global Weather Repository dataset, which contains daily weather information for cities worldwide. The primary goal of this analysis is to explore global weather patterns, identify key trends, and develop a forecasting model for predicting future temperature values.

2. Data Cleaning and Preprocessing

Upon initial inspection of the dataset, several data issues were identified and addressed as follows:

- **Handling Missing Values**: Missing data were found in multiple columns. For numerical variables, missing values were imputed using the median value of each respective column, ensuring that the central tendency was preserved. For categorical variables, the mode (most frequent value) was used to fill in the gaps.

- **Outlier Detection and Handling**: Outliers in key weather variables were identified using the Interquartile Range (IQR) method. Any values falling outside 1.5 times the IQR were capped at the lower and upper bounds to prevent extreme values from skewing the analysis.

- **Date Conversion**: The 'last_updated' column was converted into a proper datetime format to facilitate time series analysis, ensuring accurate temporal tracking of the weather data.

3. Exploratory Data Analysis

The exploratory data analysis (EDA) aimed to uncover meaningful insights into weather patterns across different regions. Key findings are summarized below:

3.1 Temperature Distribution

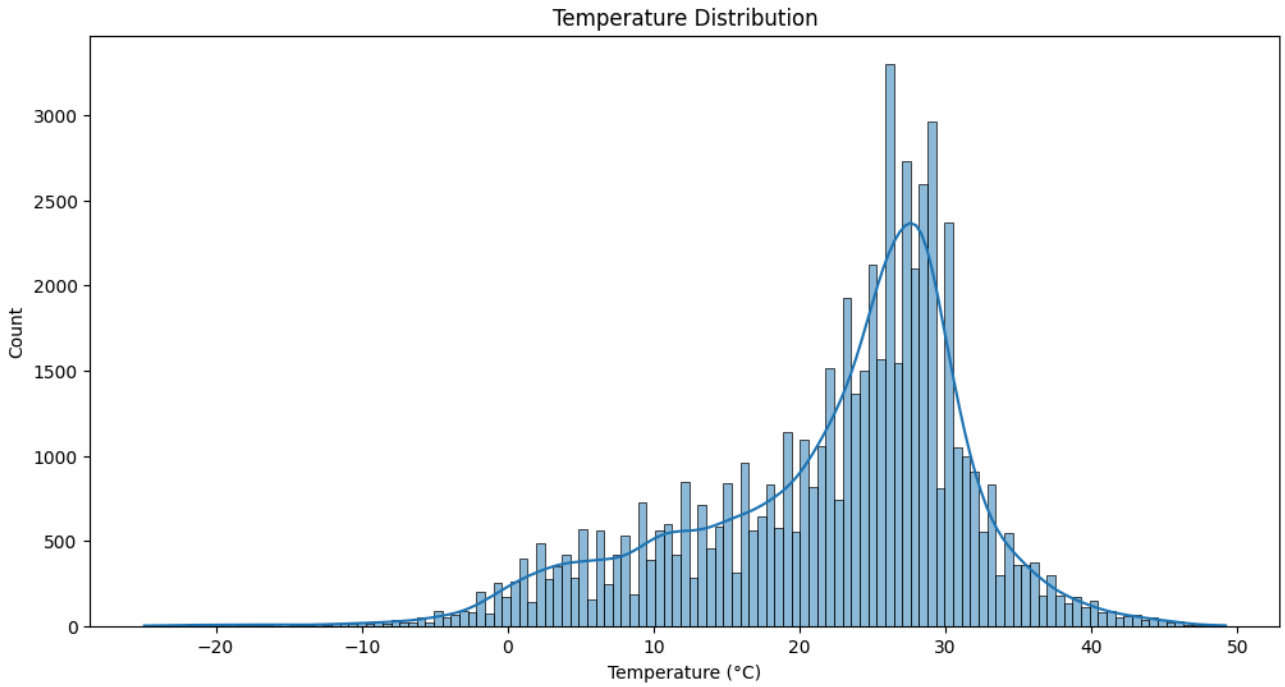

Fig: Showing the count of temperature recorded

The temperature distribution analysis provides a visualization of the range and frequency of temperatures recorded globally. This analysis helps identify common temperature ranges and highlights any bimodal patterns that could indicate seasonal variations across different regions.

3.2 Temperature by Country

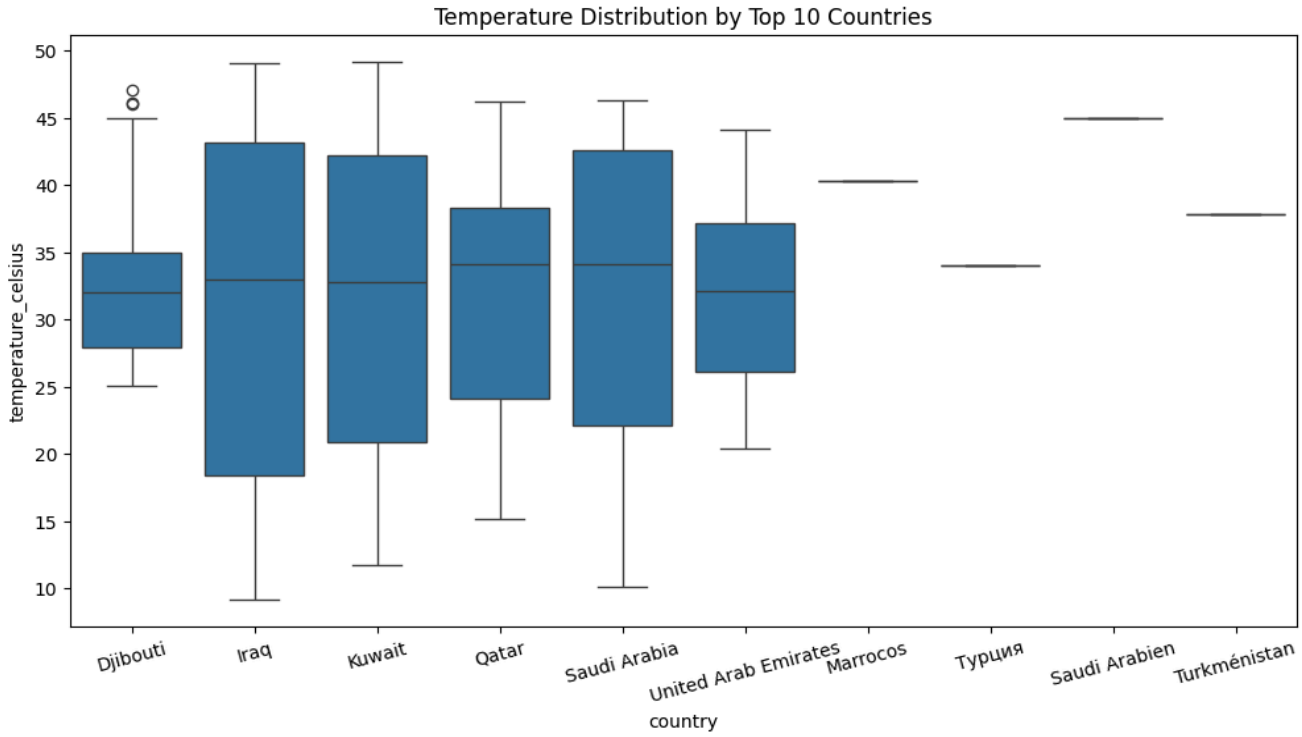

Fig: Showing box-plot of top-10 countries with higher median temperature

A box plot visualization of temperatures by country reveals significant variations in temperature distributions:

- Countries with the highest and lowest median temperatures were identified.
- Countries exhibiting the widest temperature ranges, indicating seasonal fluctuations, were highlighted.
- Outliers, where temperatures were unusually high or low, were also detected, pointing to specific regional anomalies.

3.3 Correlation Analysis

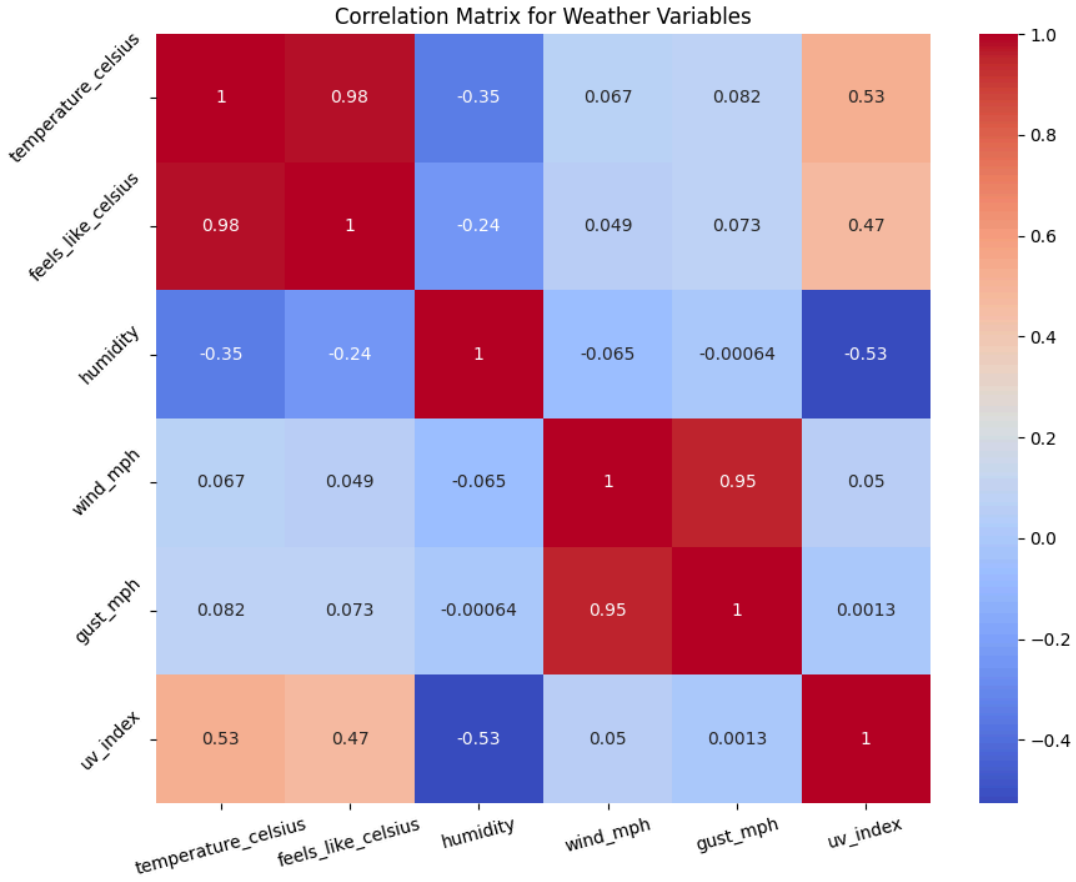

Fig: Correlation matrix having several parameters

The correlation matrix was used to examine the relationships between key weather variables. Notable correlations include:

- **Temperature and Feels-Like Temperature**: There is a strong positive correlation between the temperature and the feels-like temperature, as expected.
- **Temperature and Humidity**: A moderate negative correlation was observed, indicating that as temperature increases, humidity tends to decrease.

- **Wind Speed and Temperature**: A weak correlation was found between wind speed and temperature, suggesting that wind speed does not have a significant direct impact on temperature.
- **Temperature and UV-index**: A moderate positive correlation was found between temperature and UV index, indicating slight increase in temperature with the increment of UV index.

These insights into the relationships between weather variables are crucial for developing a robust forecasting model.

3.4 Precipitation Analysis

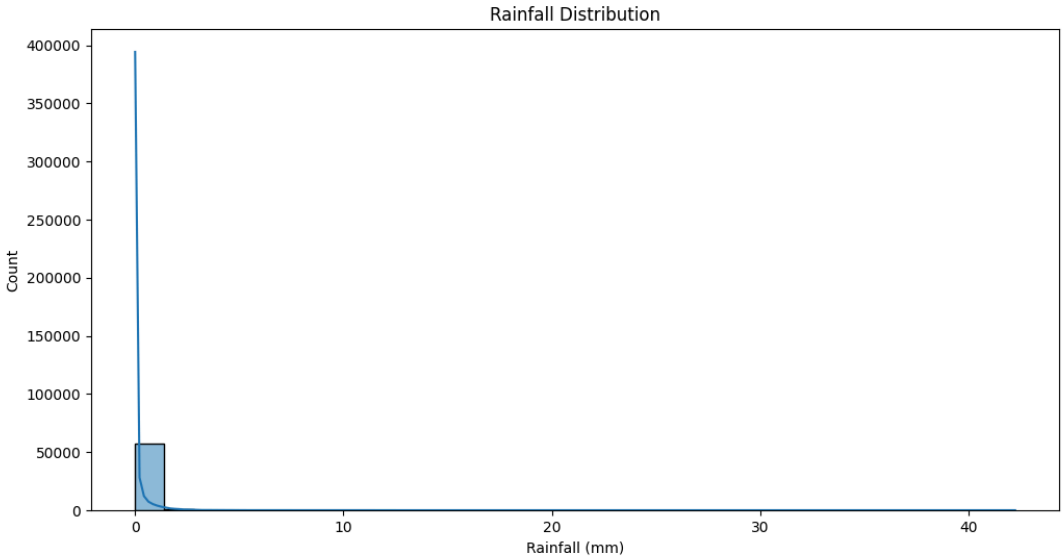

Fig: Showing the count of rainfall(mm) occurred

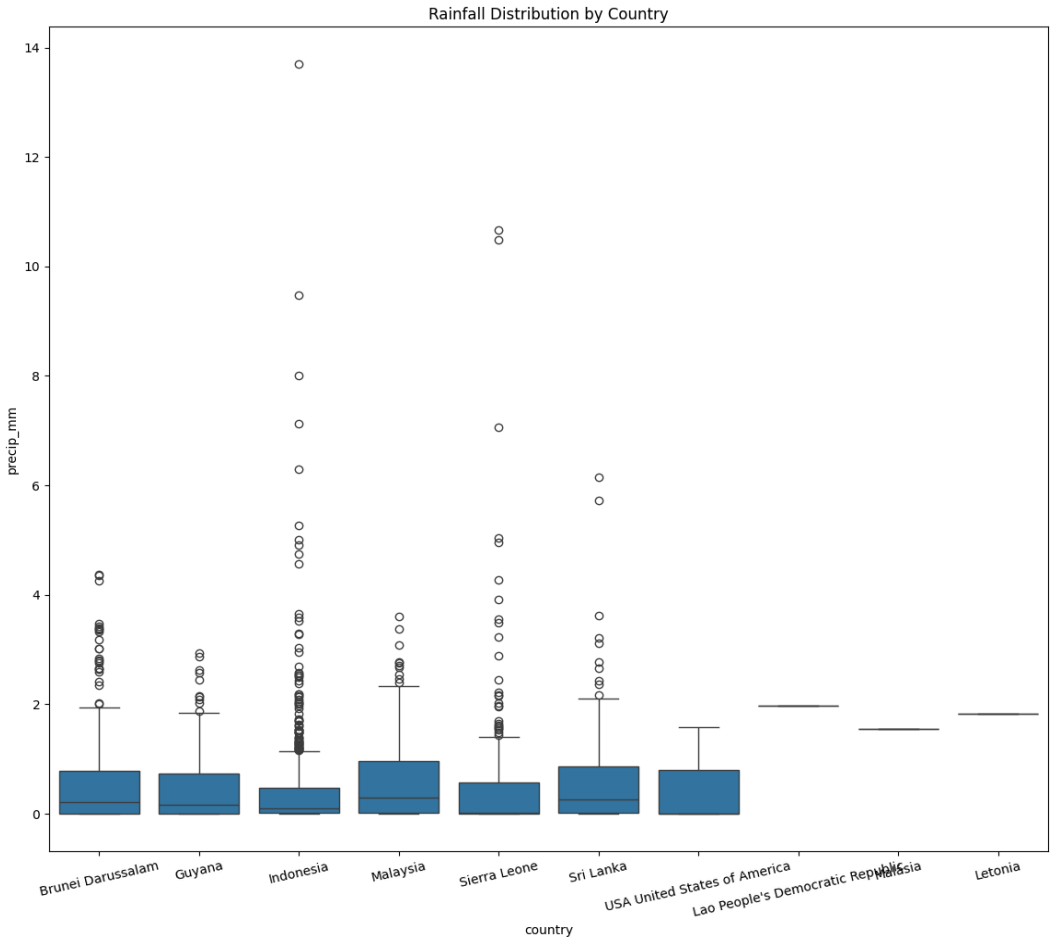

Fig: Showing box-plot of top-10 countries with higher mean precipitation

Precipitation data analysis showed a highly skewed distribution, where most records reported minimal or no precipitation, while heavier rainfall events were much less frequent. Additionally, significant variations in rainfall patterns were observed across different countries, with some regions consistently experiencing higher levels of precipitation than others.

3.5 Temperature Time Series

The time series visualization of average daily temperatures provides insights into global temperature trends:

- **Seasonal Patterns**: Clear seasonal variations in temperature were observed globally, with identifiable peaks and troughs corresponding to typical seasonal cycles.
- **Long-Term Trends**: Potential long-term trends, such as gradual warming or cooling, were identified.
- **Unusual Fluctuations**: Periods of unexpected temperature fluctuations were also highlighted, which could be indicative of extreme weather events or anomalies.

4. Time Series Forecasting

4.1 Model Development

For forecasting temperature trends, we implemented multiple regression models, including Linear Regression, Random Forest Regressor, Gradient Boosting, and XGBoost. These models were selected based on their ability to handle time series data effectively and capture complex temperature patterns. The primary reasons for selecting these models include:

- Ability to capture non-linear relationships in weather data
- Robust performance in the presence of outliers
- Capability to handle multiple input features efficiently

The models used historical temperature data with a 1-day lag window to predict future temperatures. This means each prediction was based on the temperature pattern from the previous day. A lag of 1 was found to yield the best results given the dataset size (303 observations) and the 80-20 train-test split.

4.2 Model Evaluation

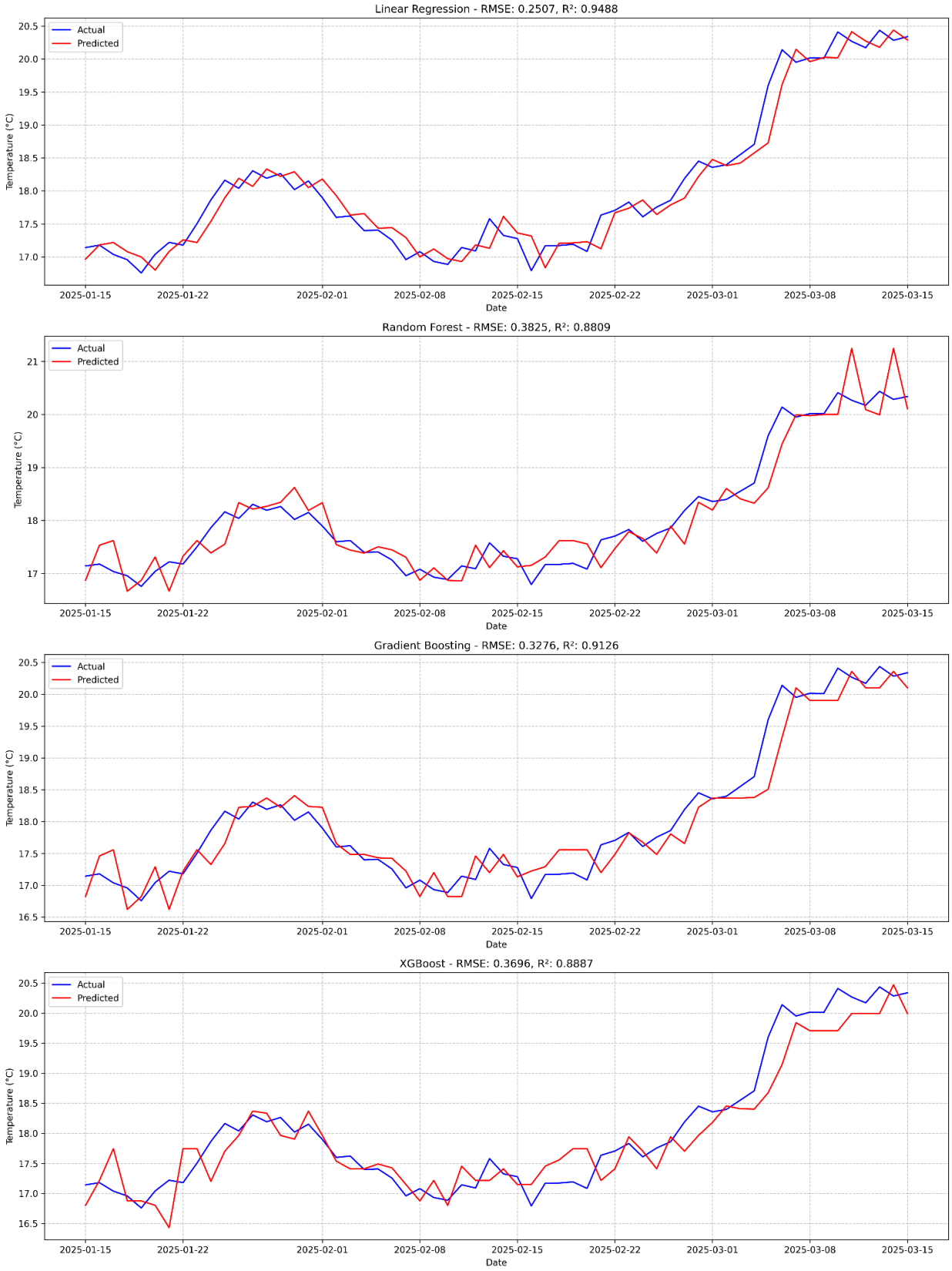

Fig: Comparison of actual and predicted temperature values from different ML models

The forecasting models were evaluated on a test set comprising 20% of the most recent data.

The performance metrics for each model are:

Models	Root Mean Squared Error(RMSE)	R-Squared(R^2)
Linear Regression	**0.251**	**0.949**
Random Forest Regressor	0.382	0.881
Gradient Boosting Regressor	0.328	0.913
XGBoost Regressor	0.370	0.889

Among the models tested, the Linear Regression and GradientBoost models exhibited superior performance, capturing temperature trends with higher accuracy. The visual comparison between predicted and actual values indicates that the models successfully track general trends but may struggle with extreme fluctuations.

5. Conclusions and Recommendations

5.1 Key Findings

- Temperature patterns exhibit clear temporal variations.
- Short-term forecasting (1-day lag) provides strong predictive performance given the dataset.
- The Linear Regression and GradientBoost models demonstrated the best overall accuracy.
- RMSE and R^2 values confirm that these models effectively capture temperature trends with minimal error.

5.2 Recommendations

- **Enhanced Feature Engineering**: Incorporate additional factors such as seasonal variations, holidays, and meteorological influences like wind speed and humidity.
- **Model Ensemble**: Consider blending multiple forecasting approaches (e.g., combining Random Forest with ARIMA) to improve accuracy.
- **Adaptive Time Windows**: Experiment with dynamic lag values to optimize forecasting across different time scales.

5.3 Limitations

- The analysis does not account for long-term climate change effects.
- Unpredictable extreme weather events remain challenging to forecast accurately.
- The dataset covers global temperature trends, potentially overlooking microclimate variations.

---

6. Future Work

- Investigate the impact of climate change on temperature trends using extended historical data.
- Develop deep learning-based models, such as LSTMs or Transformers, for more sophisticated time series predictions.
- Implement real-time weather forecasting dashboards for practical applications.