Unveiling Tomorrow's Weather : Predictive Analytics on Climate Patterns

By

Michael Robin K

Surandeer V

Naveen M

INTRODUCTION

Weather prediction is a critical aspect of planning and decision-making various fields, including agriculture, disaster management, travel, and daily life. Accurate weather forecasts significantly enhance efficiency, safety, and preparedness.

As data science interns, we embarked on a project titled "Unveiling Tomorrow's Weather: Predictive Analytics on Climate Patterns". The objective of this project is to leverage machine learning techniques and data visualization tools to improve weather prediction accuracy and user interaction.

PROJECT OBJECTIVES

To develop a Global Weather Forecast application:

- Weatherwave provide users with real-time weather data from various global models.
- Offer interactive features for location selection and detailed weather visualizations.
- Ensure the application is accessible and user-friendly for both everyday users and professionals.

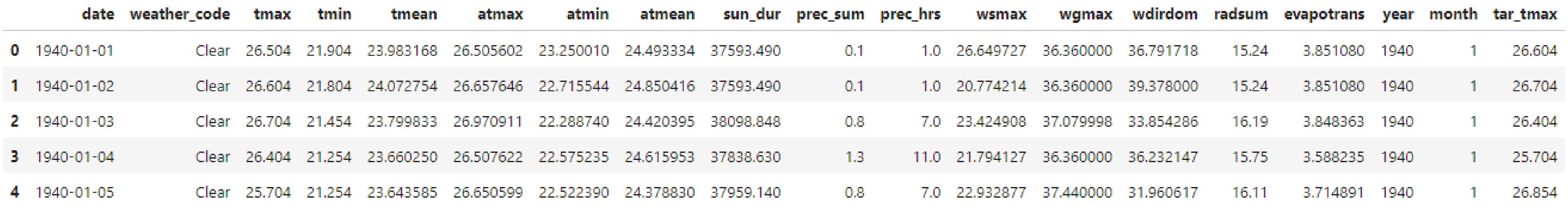
To create a Feature-Based Weather Prediction application:

- Utilize advanced machine learning models to predict maximum temperature based on various input weather features.
- Enable users to input their own weather data and receive predictions from multiple models.
- Compare predictions from different models to assess their reliability and performance.

Dataset Overview

The weather dataset provides a comprehensive collection of weather measurements including temperature, precipitation, wind speed, evapotranspiration and more.

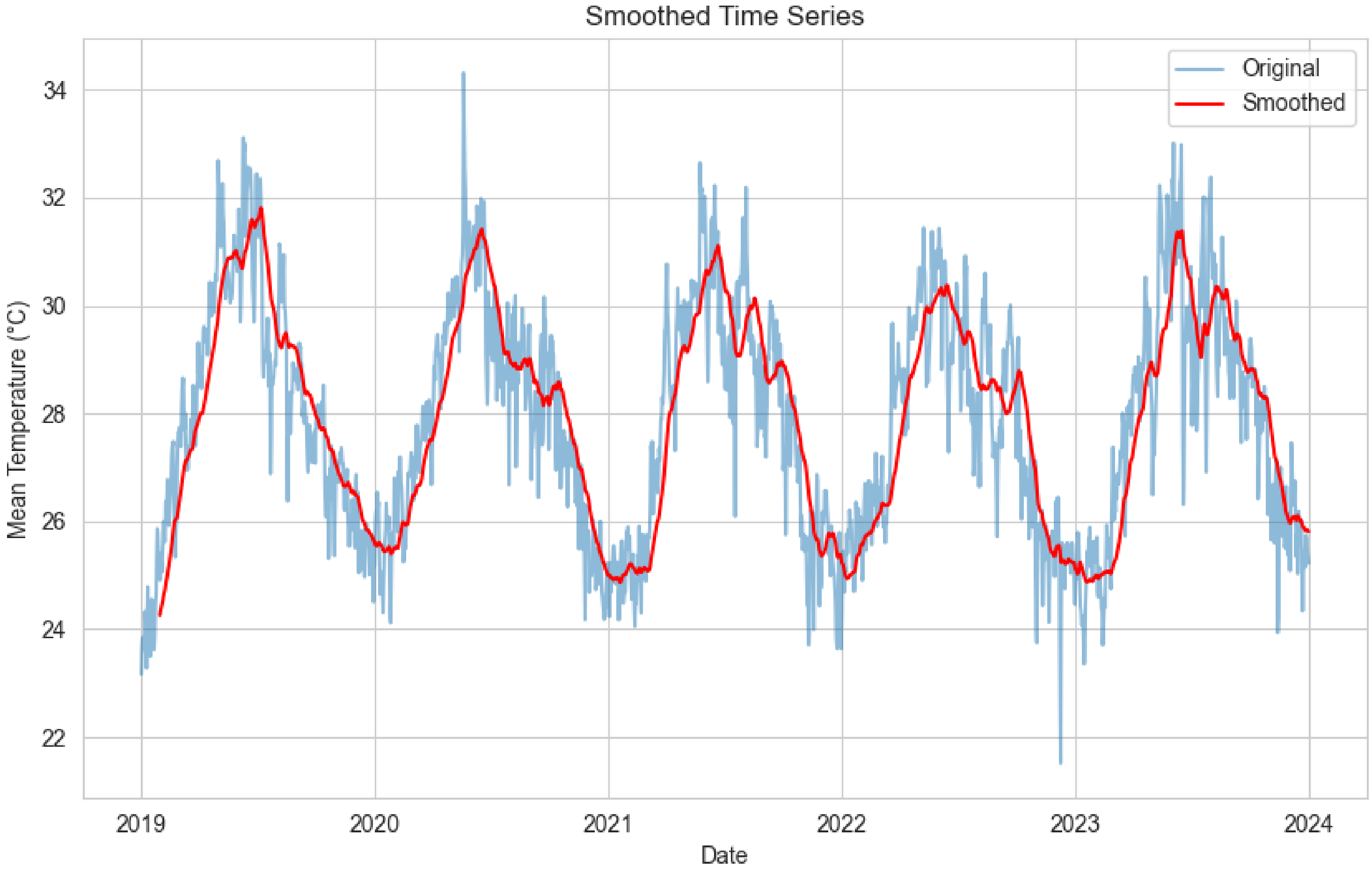
Understanding the structure and characteristics of this dataset is crucial for accurate weather prediction. Below are the first few rows of the dataset to provide an initial glimpse.



	date	weather_code	tmax	tmin	tmean	atmax	atmin	atmean	sun_dur	prec_sum	prec_hrs	wsmax	wgmax	wdirdom	radsum	evapotrans	year	month	tar_tmax
0	1940-01-01	Clear	26.504	21.904	23.983168	26.505602	23.250010	24.493334	37593.490	0.1	1.0	26.649727	36.360000	36.791718	15.24	3.851080	1940	1	26.604
1	1940-01-02	Clear	26.604	21.804	24.072754	26.657646	22.715544	24.850416	37593.490	0.1	1.0	20.774214	36.360000	39.378000	15.24	3.851080	1940	1	26.704
2	1940-01-03	Clear	26.704	21.454	23.799833	26.970911	22.288740	24.420395	38098.848	0.8	7.0	23.424908	37.079998	33.854286	16.19	3.848363	1940	1	26.404
3	1940-01-04	Clear	26.404	21.254	23.660250	26.507622	22.575235	24.615953	37838.630	1.3	11.0	21.794127	36.360000	36.232147	15.75	3.588235	1940	1	25.704
4	1940-01-05	Clear	25.704	21.254	23.643585	26.650599	22.522390	24.378830	37959.140	0.8	7.0	22.932877	37.440000	31.960617	16.11	3.714891	1940	1	26.854

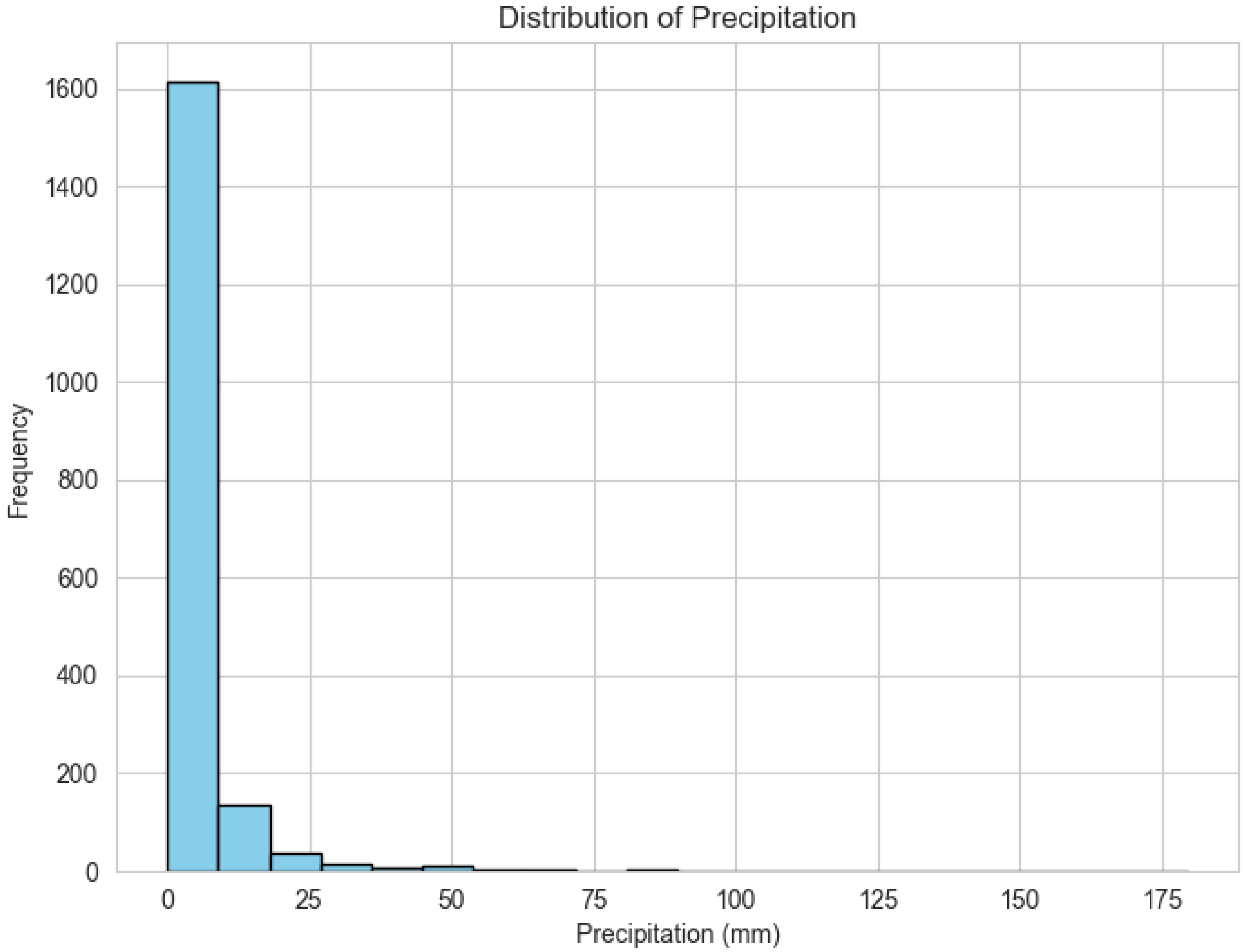
Temperature Trends

Analyzing temperature trends over time is essential to understand seasonal variations and long-term changes. The graph below illustrates the maximum temperature trends, showing how temperature fluctuates across different periods. Such trends are key to developing robust weather prediction models.



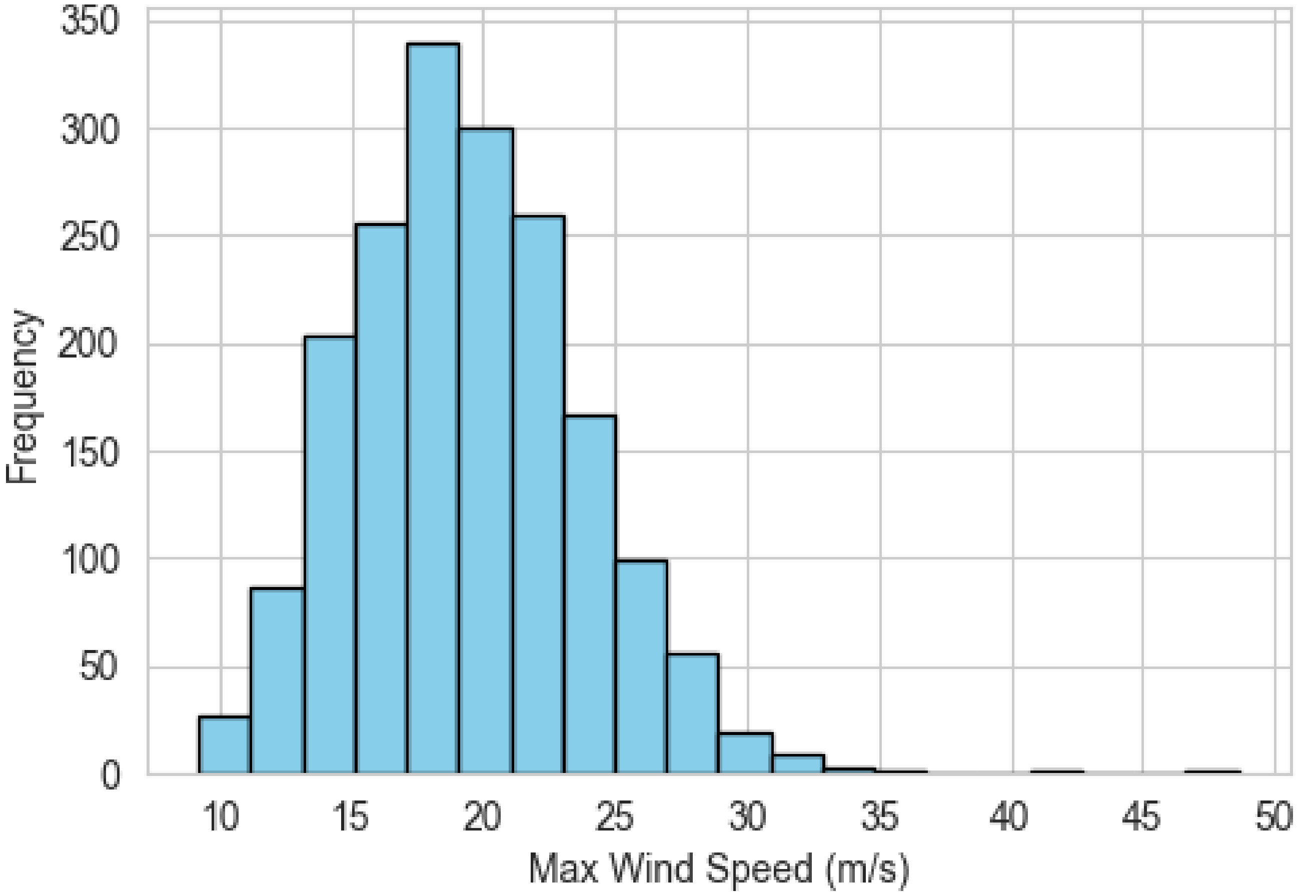
Precipitation Analysis

Precipitation data is vital for predicting rainfall patterns and their potential impact. The following plot shows the trends in precipitation over time, highlighting periods of heavy rainfall and drought. Analyzing these trends helps in forecasting future precipitation events and understanding their correlation with other weather parameters.



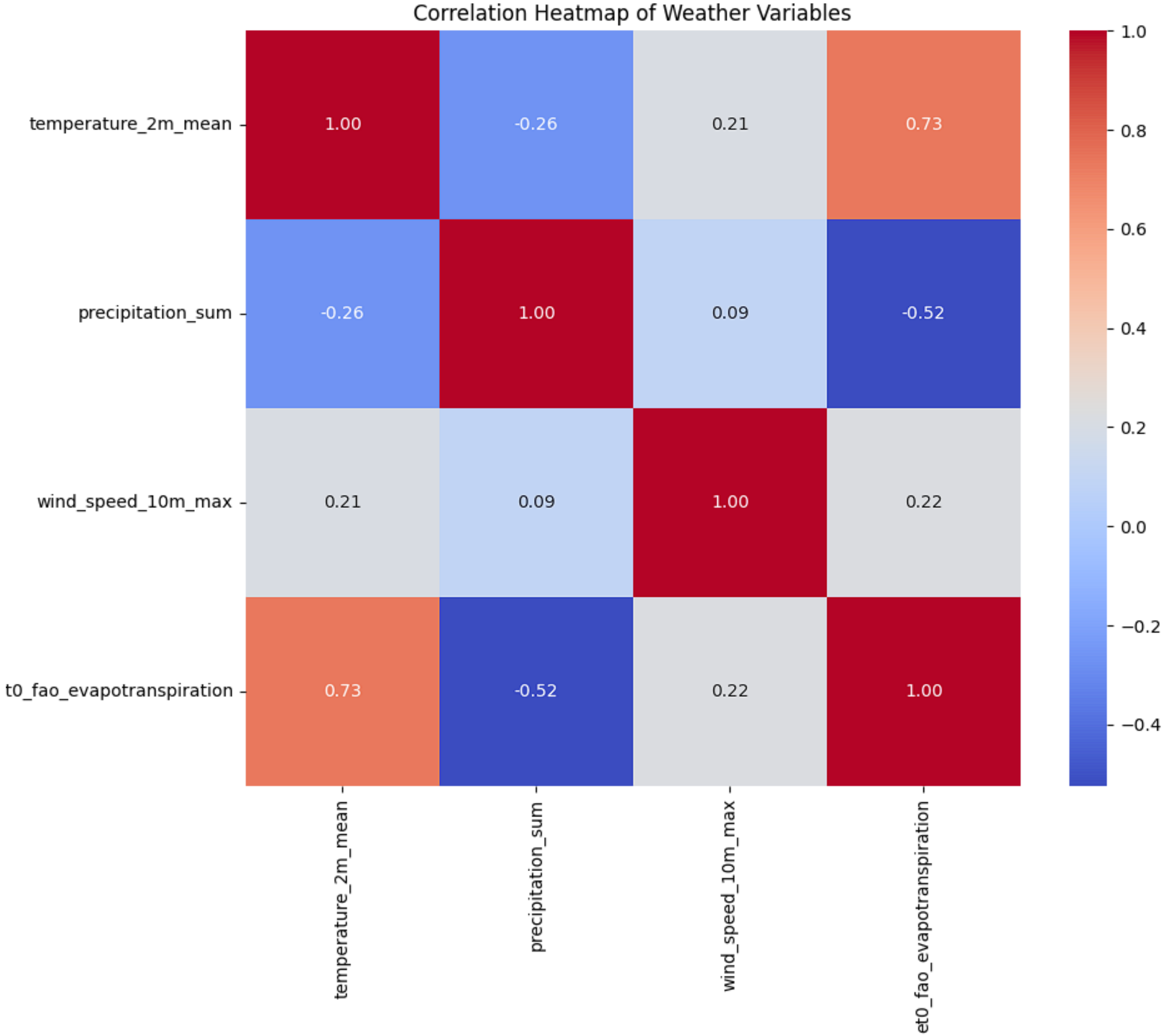
Wind Speed Analysis

Wind speed is a critical factor in weather prediction, influencing various atmospheric processes. The analysis of wind speed trends provides insights into wind patterns, helping to predict storm events and other wind-related phenomena. The graph below illustrates these trends, showing how wind speed varies over time.



Correlation heatmap

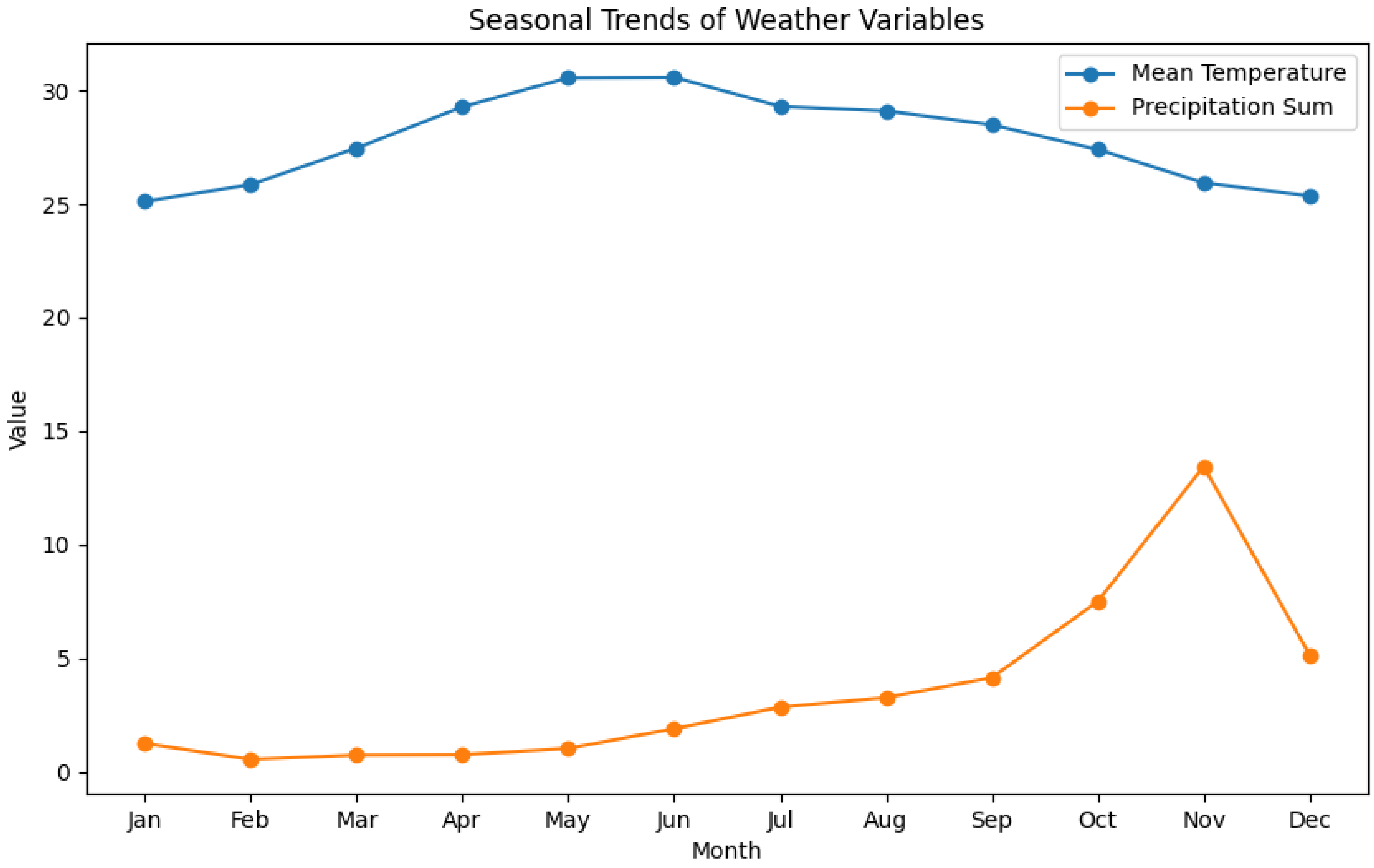
Understanding the relationships between different weather variables is crucial for building accurate predictive models. Correlation analysis helps identify how variables such as temperature, precipitation, and wind speed interact with each other. The heatmap below shows the correlation coefficients, indicating the strength and direction of these relationships.



Correlation Heatmap of Weather Variables

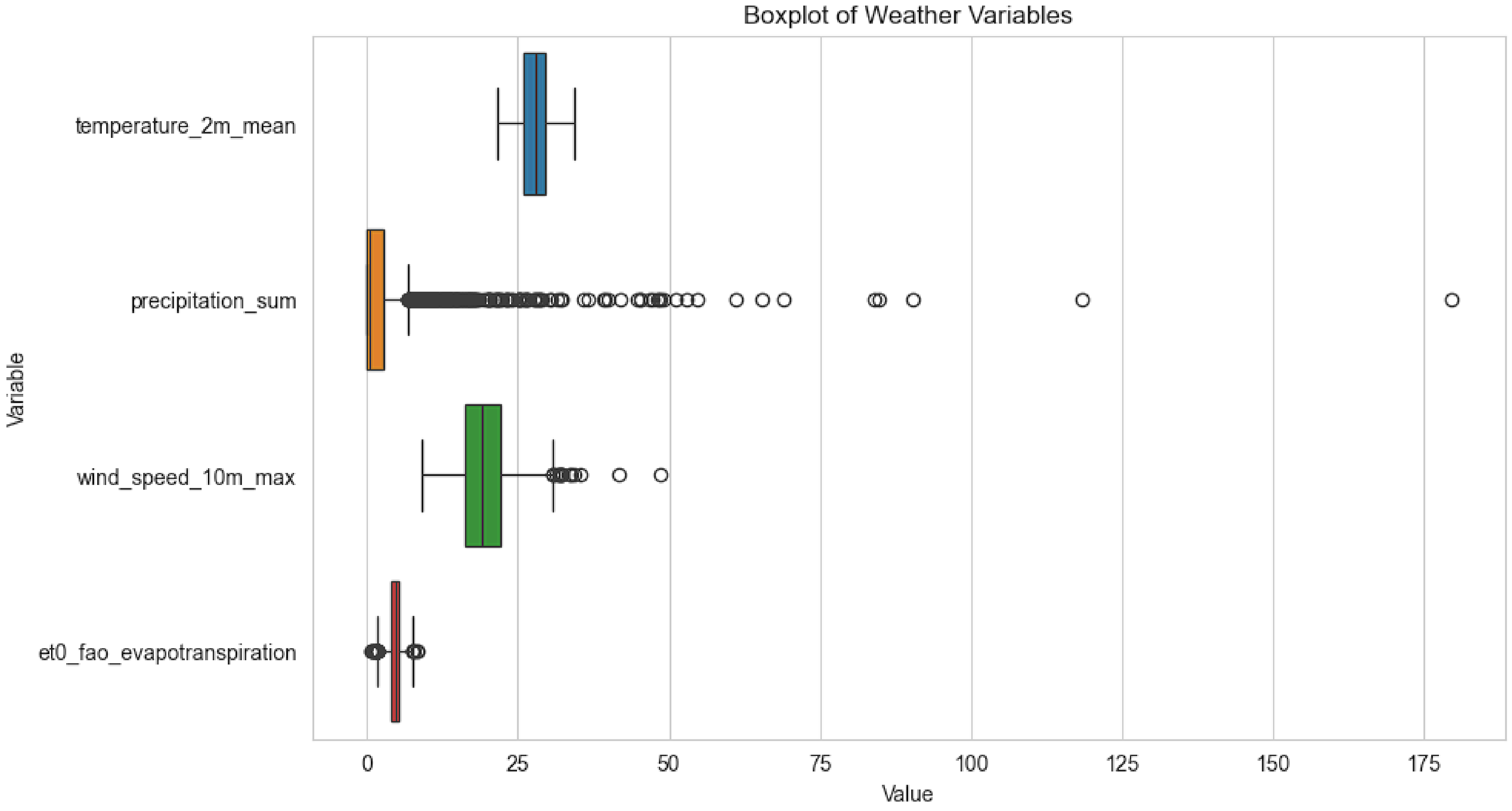
Seasonal Patterns

Weather patterns often exhibit seasonal variations. By examining data across different seasons, we can identify specific patterns and anomalies associated with each season. This seasonal analysis helps improve the accuracy of weather predictions by accounting for expected changes in different weather parameters throughout the year.



Boxplot Analysis

Boxplots provide a visual summary of the distribution of weather variables, highlighting the median, quartiles, and potential outliers. This type of analysis helps to understand the variability and central tendency of different weather parameters. The boxplot below illustrates the spread of key weather variables such as temperature, precipitation, and wind speed, offering a clear picture of their distributions and any anomalies present.



Models Description

XGBoost Model:

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that excels in prediction tasks involving structured/tabular data. It is designed for speed and performance, leveraging parallel processing to handle large datasets efficiently. One of its key features is the ability to handle missing data, which is common in real-world weather datasets. XGBoost includes built-in regularization to prevent overfitting, enhancing the model's generalization capabilities. Using a boosting framework, XGBoost sequentially improves weak learners to create a strong predictive model. This makes it ideal for predicting weather variables such as temperature, precipitation, and wind speed based on historical data.

LSTM Model:

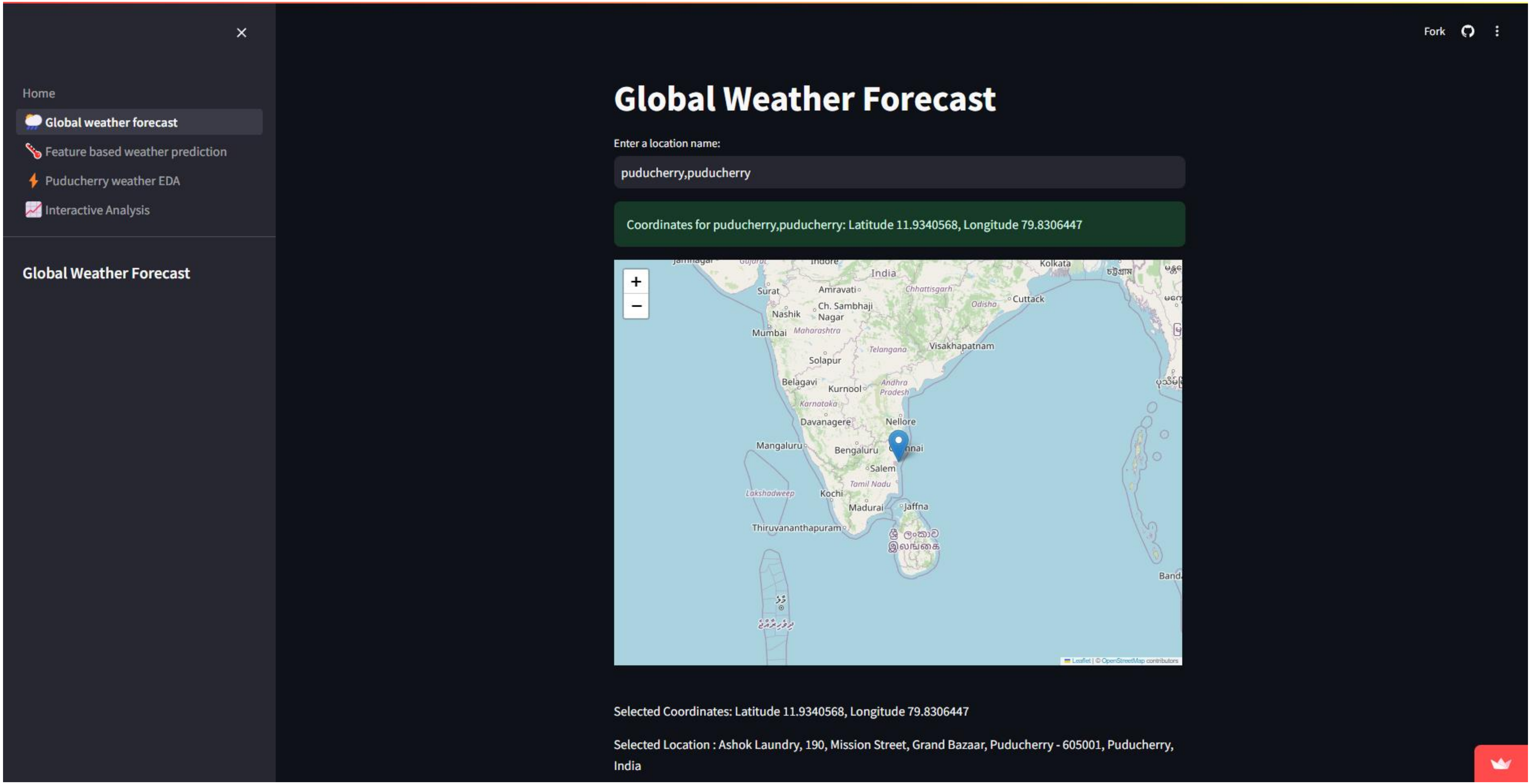
Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) that can learn long-term dependencies, making them suitable for sequential data. LSTM networks are particularly effective for time series forecasting due to their memory cells, which can store information for long periods. They utilize gate mechanisms—input, output, and forget gates—to control the flow of information, capturing temporal dependencies effectively. LSTMs are excellent for modeling temporal sequences and patterns in weather data over time, making them effective for predicting weather trends based on sequential data such as temperature readings over days or months.

Ridge Regression Model:

Ridge Regression is a type of linear regression that includes a regularization term to prevent overfitting, making it robust for predictive modeling. The regularization introduces a penalty on the size of the coefficients, helping maintain model simplicity and avoid overfitting. This technique balances the tradeoff between bias and variance, improving model performance on unseen data. Ridge Regression is particularly useful in scenarios with highly correlated predictors, as is common in weather datasets. It is suitable for predicting continuous weather variables where linear relationships are assumed, such as estimating temperature based on multiple weather factors.

Visualizing Model Performance:

To evaluate the performance of each model, metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are used for both training and testing datasets. Graphical representations, such as bar charts or line graphs, can be employed to compare the model errors. This visual comparison helps to highlight which model performs best based on the chosen metrics and discuss the trade-offs between model complexity and performance.



Global Weather Forecast

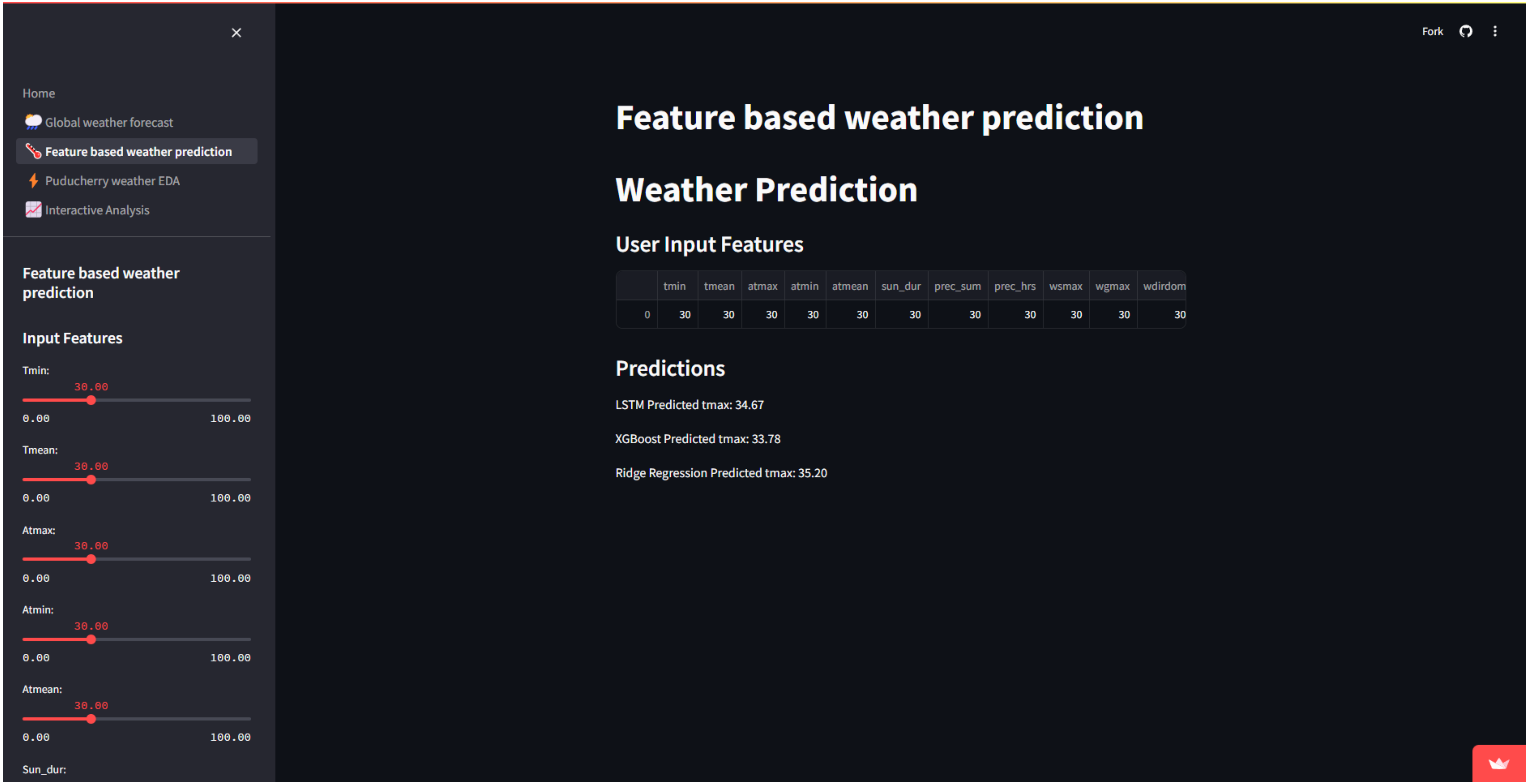
Enter a location name:

puducherry,puducherry

Coordinates for puducherry,puducherry: Latitude 11.9340568, Longitude 79.8306447

Selected Coordinates: Latitude 11.9340568, Longitude 79.8306447

Selected Location : Ashok Laundry, 190, Mission Street, Grand Bazaar, Puducherry - 605001, Puducherry, India



Feature based weather prediction

Input Features

Tmin: 30.00 (0.00 – 100.00)

Tmean: 30.00 (0.00 – 100.00)

Atmax: 30.00 (0.00 – 100.00)

Atmin: 30.00 (0.00 – 100.00)

Atmean: 30.00 (0.00 – 100.00)

Sun_dur:

Feature based weather prediction

Weather Prediction

User Input Features

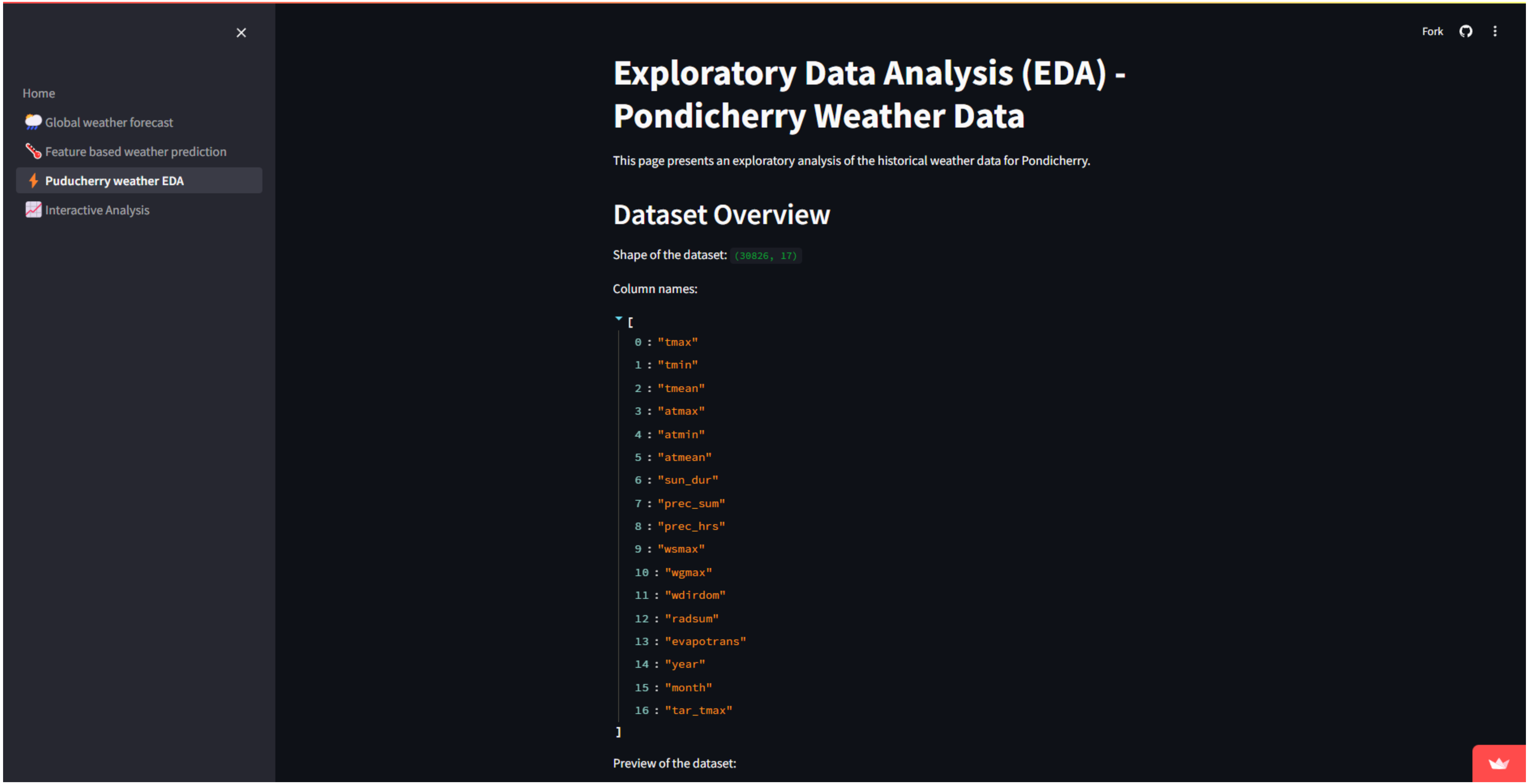
	tmin	tmean	atmax	atmin	atmean	sun_dur	prec_sum	prec_hrs	wsmax	wgmax	wdirdom
0	30	30	30	30	30	30	30	30	30	30	30

Predictions

LSTM Predicted tmax: 34.67

XGBoost Predicted tmax: 33.78

Ridge Regression Predicted tmax: 35.20



Exploratory Data Analysis (EDA) - Pondicherry Weather Data

This page presents an exploratory analysis of the historical weather data for Pondicherry.

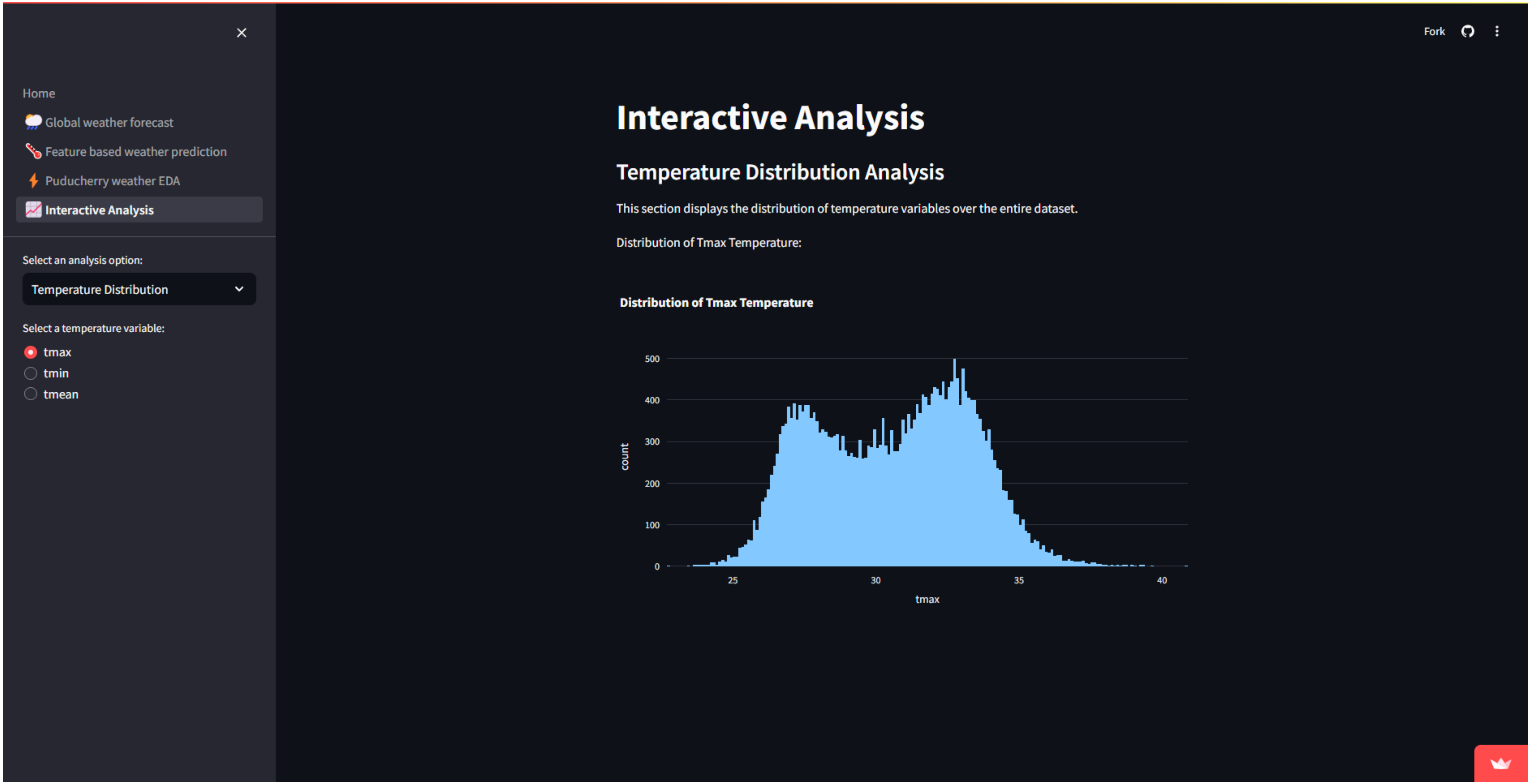
Dataset Overview

Shape of the dataset: `(30826, 17)`

Column names:

```
[
  0 : "tmax"
  1 : "tmin"
  2 : "tmean"
  3 : "atmax"
  4 : "atmin"
  5 : "atmean"
  6 : "sun_dur"
  7 : "prec_sum"
  8 : "prec_hrs"
  9 : "wsmax"
  10 : "wgmax"
  11 : "wdirdom"
  12 : "radsum"
  13 : "evapotrans"
  14 : "year"
  15 : "month"
  16 : "tar_tmax"
]
```

Preview of the dataset:



Interactive Analysis

Temperature Distribution Analysis

This section displays the distribution of temperature variables over the entire dataset.

Distribution of Tmax Temperature:

Distribution of Tmax Temperature

Select an analysis option: Temperature Distribution

Select a temperature variable:

● tmax

○ tmin

○ tmean

CONCLUSION

- The exploratory data analysis (EDA) of the weather dataset has provided valuable insights into the key weather variables and their interrelationships. By visualizing distributions, trends, and correlations, we gained a deeper understanding of the data, which is crucial for building accurate predictive models. The EDA highlighted the variability and patterns in weather variables such as temperature, precipitation, and wind speed, guiding the feature selection process for the models.

- In the model-building phase, we employed three distinct approaches to forecast weather variables: XGBoost, LSTM, and Ridge Regression. Each model brought its unique strengths to the table. XGBoost, with its powerful gradient boosting framework, demonstrated high efficiency and accuracy in handling tabular data and dealing with missing values. The LSTM model, leveraging its capability to capture long-term dependencies, proved effective in predicting weather trends over time, thanks to its recurrent neural network architecture. Ridge Regression provided a straightforward and interpretable model, well-suited for handling multicollinearity and producing reliable predictions for continuous weather variables.

- The comparative analysis of these models showed that each has its optimal application scenarios. XGBoost excelled in making precise predictions for specific weather variables, while LSTM was superior in modeling temporal sequences and forecasting trends. Ridge Regression offered a balanced approach, reducing overfitting and maintaining interpretability.