Weather Prediction Model Documentation

This document provides a comprehensive overview of the process followed to build and evaluate models for predicting various weather conditions based on historical weather data from multiple locations. The goal was to develop models capable of predicting weather for both known locations within the dataset and new, unseen locations by leveraging geographical coordinates as features.

1. Data Loading and Initial Inspection

The process began by loading five separate datasets containing daily measurements for:

- Rainfall (`Rainfall.csv`)
- Relative Humidity (`Relative Humidity.csv`)
- Maximum Temperature (Tmax) (`Tmax.csv`)
- Minimum Temperature (Tmin) (`Tmin.csv`)
- Wind Speed (`Wind Speed.csv`)

Each dataset contained columns for 'Station ID', 'Geogr1' (likely Latitude), 'Geogr2' (likely Longitude), 'Name', 'Year', 'Month', 'Time', and daily readings from '01' to '31'.

Initial inspection revealed the presence of missing values across all datasets, particularly in the daily reading columns. Data types were also reviewed to ensure suitability for numerical processing.

2. Data Cleaning

The data cleaning phase addressed missing values and potential inconsistencies:

- Rows with missing identification information (Station ID, Geogr1, Geogr2, Name, Year, Month, Time) in the Wind Speed DataFrame were dropped.
- Missing values in daily reading columns with less than 10% missing data were imputed with the mean of the respective column.
- Daily reading columns '29', '30', and '31', which had a high percentage of missing values across the datasets, were dropped to avoid introducing significant bias through imputation.
- Data types of the daily reading columns were converted to float, coercing errors to handle any non-numeric entries.
- Outliers in the daily reading columns were identified using box plots and handled by capping the values at the 99th percentile to retain the majority of the data while mitigating the influence of extreme values.

3. Data Preprocessing and Feature Engineering

The data was transformed and new features were created to prepare it for modeling:

- The datasets were melted from a wide format (daily columns) to a long format, resulting in a single 'Day' column and a value column for the specific weather measurement.
- A 'Date' column was created by combining 'Year', 'Month', and 'Day', with invalid date combinations handled by coercing them to `NaT`.
- The original 'Time' column was dropped.
- **Crucially for new location prediction:** 'Geogr1' and 'Geogr2' (geographical coordinates) were retained as features, and the one-hot encoding of 'Station ID' and 'Name' was removed in favor of using these geographical features to enable generalization to new locations.
- Temporal features, 'Year', 'Month', and 'DayOfWeek', were extracted from the 'Date' column.
- Lag features (`_lag1`) were created for each weather condition, representing the value from the previous day, grouped by 'Station ID' to capture temporal dependencies specific to each location.
- Remaining missing values after processing were handled by imputing with the mean.
- Original identification columns ('Station ID', 'Name', 'Date', 'Year', 'Month', 'Day') were dropped as they were no longer needed in their original format for modeling.

4. Model Selection

The Gradient Boosting Regressor was selected as the regression model for predicting each of the weather conditions. This model is suitable for handling complex relationships between features and continuous target variables.

5. Model Training and Evaluation (with Geographical Coordinates)

Separate Gradient Boosting Regressor models were trained for each of the five weather conditions: Rainfall, Relative Humidity, Tmax, Tmin, and Wind Speed. The processed data, including geographical coordinates and engineered features, was split into training (80%), validation (10%), and testing (10%) sets. Each model was trained on the training set.

The models were evaluated on the test set using the following metrics: Mean Squared Error (MSE), R-squared (R2), and Mean Absolute Error (MAE).

Weather Condition	MSE	R2	MAE
Rainfall	11.47	0.86	0.86
Relative Humidity	6.52	0.97	0.99
Tmax	0.16	0.97	0.14

Weather Condition	MSE	R2	MAE
Tmin	0.20	0.93	0.15
Wind Speed	197.81	0.48	11.68

Evaluation Summary:

- The models for Relative Humidity, Tmax, and Tmin demonstrated strong performance with high R-squared values (above 0.93) and low MSE and MAE, indicating accurate predictions for these variables.
- The Rainfall model also performed well with an R-squared of 0.86 and a relatively low MAE.
- The Wind Speed model showed the lowest performance among the five, with an R-squared of 0.48 and a higher MAE. This suggests that predicting Wind Speed is more challenging with the current features and model, potentially due to the inherent variability of wind speed or limitations in the available data.

6. Prediction (for New Locations)

The trained models, which now utilize geographical coordinates as features, can be used to make predictions for weather conditions in new locations not present in the original dataset. This is achieved by providing the geographical coordinates (Geogr1 and Geogr2) of the new location, along with the temporal features (Year, Month, DayOfWeek) and lag features (based on the previous day's weather at that location).

A demonstration was performed to predict weather conditions for Kumasi (a location potentially not in the training data in this context) for a specific future date (August 10th, 2025) by providing its coordinates and dummy lag features.

7. Model Saving

The trained models for each weather condition were saved as joblib files. Two options were demonstrated:

- Saving each model individually to the directory `/content/trained_weather_models_geo`.
- Saving all trained models together in a single joblib file (`combined_weather_models_geo.joblib`) in the same directory.

These saved models can be loaded into a web application backend or other deployment environment to serve predictions.

Conclusion

The process successfully involved loading, cleaning, and preprocessing weather data, engineering relevant features including geographical coordinates for new location prediction, training individual Gradient Boosting Regressor models for each weather condition, and evaluating their performance. The models for Temperature and Humidity performed very well, while the Wind Speed model's performance was more modest. The trained models are now ready to be used for making predictions for both known and potentially new locations based on their geographical coordinates. Further work could focus on improving the Wind Speed model and exploring additional features or modeling techniques.

1.