Weather Long-term Time Series Forecasting Report

1. Dataset Description

The analysis was performed on a comprehensive time-series weather dataset using big data processing techniques.

1.1 Source & Scope

- **Time Period:** The data covers weather observations from **2020-01-01** to **2021-01-01**.
- **Records:** The dataset contains **52,696** individual observations.
- **Processing Environment:** The project utilized **PySpark** (SparkContext v4.0.1) for large-scale data manipulation and modeling.

1.2 Key Columns

The dataset includes 21 features, with primary weather variables:

- **date**: Timestamp of the observation.
- **p**: Atmospheric Pressure (Mean: 989.99 hPa).
- **T**: Air Temperature in Celsius (Mean: 10.82°C).
- **Tdew**: Dew Point.
- **rh**: Relative Humidity (Mean: 72.49%).
- **wv / max. wv**: Wind Velocity and Maximum Wind Velocity.
- **SWDR / PAR**: Solar and Photosynthetically Active Radiation measurements.

1.3 Data Quality

- Initial checks indicated **0 missing values** after preliminary data loading.
- However, descriptive statistics show minimum values of **-9999.00** for variables like wv and max. PAR, suggesting that these are placeholders for missing or bad data that were not properly handled during the initial cleaning phase.

2. Operations Performed

The project focused on preparation, modeling, and output generation using a Python/Spark environment.

2.1 Data Cleaning & Exploration

- Data types were confirmed, including the conversion of the date column to a datetime format.
- Descriptive statistics were generated across all 21 columns to understand the range, distribution, and central tendency of the weather parameters.

2.2 Predictive Modeling

- A **Linear Regression (LR)** model was implemented and trained to predict a target weather variable.
- Features were prepared using a **VectorAssembler** for the linear regression input.
- The model generated predictions, which were saved alongside the original temperature data.

2.3 Descriptive Visualization

- A **Temperature Distribution Pie Chart** was generated by binning the temperature (T) data into ranges (e.g., 0-10, 11-20, 21-30, 31-40). This visualization helps understand the frequency of different temperature ranges over the year.

2.4 Data Export

- The final predictions, including the original date and T values, were exported to a CSV file named **final_weather_predictions.csv**.

3. Key Insights

3.1 Temperature Extremes and Range

- **Minimum Temperature:** The lowest recorded temperature was **-6.44°C**.
- **Maximum Temperature:** The highest recorded temperature was **34.80°C**.
- **Overall Average:** The dataset's mean temperature was approximately **10.82°C**.

3.2 Atmospheric Conditions

- **Humidity:** The mean Relative Humidity (rh) was high at **72.49%**, suggesting generally moist conditions.
- **Pressure Stability:** Atmospheric pressure (p) remained relatively stable, averaging near **990 hPa**.
- **Rainfall:** The mean rainfall was extremely low (**0.0118**), suggesting infrequent or low precipitation events throughout the recording period.

3.3 Modeling and Prediction

- The project successfully implemented a **Linear Regression** model suitable for baseline forecasting of a continuous weather variable.
- The prediction results were persisted, making the model's output available for further post-analysis and integration.

4. Recommendations

4.1 Data Quality Remediation

- **Data Imputation:** Immediately address the placeholder values of **-9999** found in columns like wv (wind velocity) and max. PAR. These values should be properly filtered or replaced using appropriate imputation techniques (e.g., mean imputation, interpolation) to prevent model skew.

4.2 Advanced Forecasting & Model Evaluation

- **Time Series Models:** Since weather data is a classic time series, explore specialized models like **ARIMA**, **SARIMA**, or **Prophet** to capture temporal dependencies and seasonality, which could significantly improve prediction accuracy over simple Linear Regression.
- **Performance Metrics:** The next phase of the project must include a dedicated section for model evaluation, providing key metrics such as **R-squared**, **Mean Absolute Error (MAE)**, and **Root Mean Square Error (RMSE)** to quantify predictive capability.

4.3 Feature Engineering

- **Lagged Features:** Create lagged versions of the target variable and other highly correlated features (e.g., Tdew, rh) to provide the model with a historical context, which is crucial for weather forecasting.
- **Temporal Features:** Extract features like **Hour of Day** and **Day of Year** from the date column to help the model learn daily and annual cyclical patterns.