**PM Accelerator Mission:** The company mission is to break down financial barriers and achieve educational fairness

**Analysis of World Weather repository**

**Introduction**

**Objective of the Analysis** The primary aim of this report is to delve into the intricate patterns of global weather conditions by leveraging a rich dataset filled with various meteorological variables. Through sophisticated statistical and machine learning methods, we seek not only to understand these patterns but also to predict future weather conditions effectively. Such insights are vital for strategic planning and informed decision-making across industries heavily influenced by weather changes, such as agriculture, transportation, and emergency management.

**Overview of the Dataset** The dataset, named "GlobalWeatherRepository.csv," comes from a comprehensive initiative to monitor weather globally. It encompasses data from various locales across the globe, showcasing a wide array of climate conditions. Key variables captured include:

* **Temperature**: This is reported in degrees Celsius and gives insights into the warmth or coolness of different regions.
* **Precipitation**: Measured in millimeters, this variable indicates the volume of rainfall areas receive, which is crucial for water resource management.
* **Wind**: This includes measurements of both speed (in km/h) and direction, which are pivotal for predicting weather fronts and storm paths.
* **Air Quality Indices**: These measurements track multiple pollutants, providing a snapshot of environmental health and aiding public health monitoring.

The data, collected over several years, provides a robust platform for identifying long-term trends and understanding seasonal weather behaviors.

**Importance of This Study** The analysis is key for several reasons:

* **Climate Research**: It contributes empirical data that can enhance our understanding of climate change, supporting ongoing research in this critical area.
* **Economic Benefits**: More accurate weather predictions can help industries reduce the economic impacts of unexpected weather changes and optimize operational efficiency.
* **Enhanced Public Safety**: Better forecasting can improve disaster readiness and response, potentially saving lives and mitigating the effects of natural disasters.

**Goals of the Report** This report aims to:

1. Conduct a detailed exploratory analysis to unearth patterns and trends within the weather data.
2. Evaluate various forecasting models to identify the most accurate in predicting different weather variables.
3. Determine how well these models capture seasonal trends and extreme weather conditions.
4. Provide recommendations on improving data collection, analytical methods, and model precision based on the insights garnered.

**Data Preprocessing**

**Handling Missing Data** Given the comprehensive nature of the global weather dataset, missing values were inevitable across various columns. Our approach to managing these gaps in data included:

* **Assessment**: We first assessed the extent and distribution of missing data across the dataset to understand its impact.
* **Simplification Decision**: For the sake of clarity and maintaining the integrity of our analysis, we chose to remove rows with missing values. This decision was based on the premise that the missing values were few enough not to bias the results significantly. This method was preferred for its straightforwardness, though we acknowledge that in future analyses, more sophisticated methods like statistical imputation might be used to retain and utilize more data.

**Managing Outliers** Outliers can dramatically skew data analysis, leading to potentially deceptive conclusions.

* **Detection Technique**: Using the Interquartile Range (IQR) method allowed us to identify extreme values effectively. This method considers outliers as any data points that lie significantly outside the middle 50% of the data.
* **Handling Approach**: We carefully reviewed each outlier to understand its origin—whether it was a measurement error or a rare meteorological event. Outliers deemed to be errors were excluded to enhance the accuracy of our analysis.

**Feature Engineering** Transforming raw data into a format suitable for analysis involved several steps:

* **Normalization**: Key variables, like temperature and wind speed, were normalized to ensure consistency in scale across the dataset, an essential factor for many predictive models.
* **Encoding Techniques**: Categorical data was transformed using one-hot encoding, making it suitable for input into machine learning models, enhancing their ability to make meaningful predictions.

**Exploratory Data Analysis (EDA)**

**Summary Statistics** We commenced our EDA by computing summary statistics for all key variables. This approach provided us with a quick snapshot of central tendencies, variability, and typical data distributions. Understanding these statistics helped us gauge the normalcy of weather conditions across different regions and identify any glaring discrepancies or unexpected values.

A blue line graph with numbers

Description automatically generatedA blue graph with white lines

Description automatically generated

**Correlation Analysis** A cornerstone of our EDA was analyzing the correlation among different weather variables:

* **Heatmap Visualization**: We used heatmaps to visually represent the correlation coefficients between variables such as temperature, precipitation, and wind speed. This not only helped in identifying relationships but also in spotting any potential issues of multicollinearity, which could affect model performance.
* **Insights Gained**: For instance, high correlations between certain air quality indices and urban locations might indicate pollution hotspots, which are crucial for environmental monitoring and policy-making.

A screenshot of a data visualization

Description automatically generated

**Visual Data Analysis** To complement our statistical analyses, we employed various data visualization techniques:

* **Time Series Plots**: We plotted key metrics like temperature and precipitation over time to identify seasonal patterns and trends. These visualizations were instrumental in confirming the cyclic nature of weather phenomena and their variability across years.
* **Scatter Plots**: By plotting temperature against precipitation, we explored their relational dynamics. These plots helped us understand how weather conditions interact in different environments, enhancing our predictive modeling strategies.

**Identifying Patterns and Anomalies**

* **Pattern Recognition**: Through EDA, we identified several recurring patterns, such as the increase in temperature during summer months and heavier precipitation during specific seasons.
* **Anomaly Detection**: Visual tools also enabled us to spot anomalies such as unexpected spikes in temperature or unseasonal rain, which could be indicative of data errors or significant environmental events.

A graph with blue lines

Description automatically generated with medium confidenceA graph of a graph showing the temperature and precipitation

Description automatically generated

**Model Building and Evaluation:**

**Gradient Boosting Model**

* **Model Setup**: We chose the Gradient Boosting Regressor for its prowess in handling non-linear relationships and interactions between features. This model is particularly adept at complex datasets like ours, where the relationships between variables can be intricate.
* **Parameter Tuning**: Utilizing techniques such as Grid Search, we meticulously tuned the model’s hyperparameters to find the optimal settings. This process ensures the model is neither overfitting nor underfitting.
* **Model Training**: The model was trained on a split dataset comprising 80% training data, ensuring it had a robust sample to learn from.

A graph showing the temperature of a year

Description automatically generated with medium confidence

A graph of a graph showing the temperature

Description automatically generated with medium confidence

**SARIMAX Model for Time Series Forecasting**

* **Rationale**: Given the sequential nature of our data, a SARIMAX model was implemented to capitalize on the seasonal patterns and autoregressive characteristics observed during the EDA phase.
* **Configuration**: The model parameters were carefully chosen based on insights from autocorrelation and partial autocorrelation analyses, which helped in identifying the significant lags and seasonal effects.
* **Fitting the Model**: This step involved adapting the model to historical data, ensuring it captures the underlying trends and cycles effectively.

A blue line with a orange line

Description automatically generated

A screenshot of a graph

Description automatically generated