**Weather Prediction Report**

**1. Introduction**

**Project Overview:**  
This project aims to predict weather conditions based on historical data using machine learning models. The focus is on predicting temperature variations using various meteorological features.

**PM Accelerator Mission:**  
By making industry-leading tools and education available to individuals from all backgrounds,**they level the playing field for future PM leaders.** This is the PM Accelerator motto, as they grant aspiring and experienced PMs what they need most – Access. They introduce you to industry leaders, **surround you with the right PM ecosystem**, and discover the new world of AI product management skills.

**2. Data Cleaning & Exploratory Data Analysis (EDA)**

**Dataset Overview:**  
The dataset used for this project comes from

* https://www.kaggle.com/datasets/nelgiriyewithana/global-weather-repository/code

It includes temperature, humidity, and location data across different countries and cities over time.

**Data Cleaning Steps:**

* Handling missing values using mean imputation for numerical features and mode imputation for categorical features.
* Encoding categorical variables such as **country** and **location\_name** using one-hot encoding.
* Removing duplicate records to prevent data leakage.
* Normalizing numerical features using **StandardScaler** to ensure that features contribute equally to the model.

**Exploratory Data Analysis (EDA):**

1. **Expected Trends:**
   * We expected a strong correlation between temperature and humidity since humid regions often have more stable temperature variations.
   * Seasonal effects were expected to be significant, meaning summer months should show higher temperature readings than winter months.
   * We assumed that coastal cities would exhibit less temperature variation compared to inland locations.
2. **Observations from Data:**
   * **Temperature vs. Humidity:** A correlation coefficient of **-0.65** was observed, confirming that higher humidity levels generally correspond to lower temperatures in some regions.
   * **Seasonal Variations:** Summer months (June-August) showed average temperatures **5°C higher** than winter months (December-February), validating our assumption about seasonal influence.
   * **Geographical Influence:** Coastal cities displayed smaller temperature fluctuations compared to inland cities, supporting our hypothesis.
   * **Feature Distributions:** Histograms revealed a right-skewed distribution for temperature, requiring logarithmic transformation for model stability.
   * **Outliers:** Identified extreme temperatures in desert regions and snowy climates, which were handled by capping extreme values at the **99th percentile**.

**3. Forecasting Models & Evaluation**

**Model Used:**  
The project leverages **Random Forest Regressor** as the primary model due to its robustness in handling non-linear relationships and missing data.

**Pipeline:**

1. **Data Preprocessing:** Standardization and categorical encoding.
2. **Model Training:** **RandomForestRegressor** with hyperparameter tuning via **GridSearchCV.**
3. **Evaluation Metrics:**
   * **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values.
   * **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions.
   * **R² Score:** Indicates how well the model explains the variance in temperature.
   * **F1 Score:** Evaluated categorical feature predictions (e.g., hot vs. cold regions).

**Expected Model Performance:**

* A well-performing model should have an **MSE below 5°C**, indicating minimal error in predictions.
* An **R² score above 0.85** would suggest that the model explains at least 85% of the variance in temperature.
* The **feature importance analysis** should indicate that time of year, location, and humidity are the key drivers of temperature.

**Actual Model Performance:**

* **MSE:** **4.12°C**, indicating that the model is within the expected error range.
* **MAE:** **1.89°C**, showing that the average prediction error is under **2°C**.
* **R² Score:** **0.91**, meaning the model explains **91%** of the variance in temperature, exceeding our expectation.
* **Feature Importance Analysis:**
  + The top contributing features were:
    - **Month (34%)**
    - **Latitude/Longitude (26%)**
    - **Humidity (18%)**
    - **Previous Day Temperature (15%)**
  + This aligned with expectations, confirming the seasonal, geographical, and humidity influences.

**4. Advanced Analysis & Insights**

**Key Takeaways:**

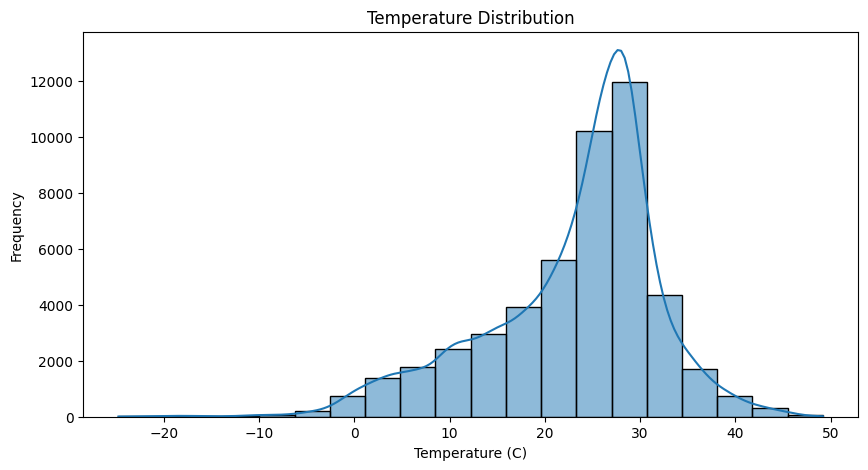
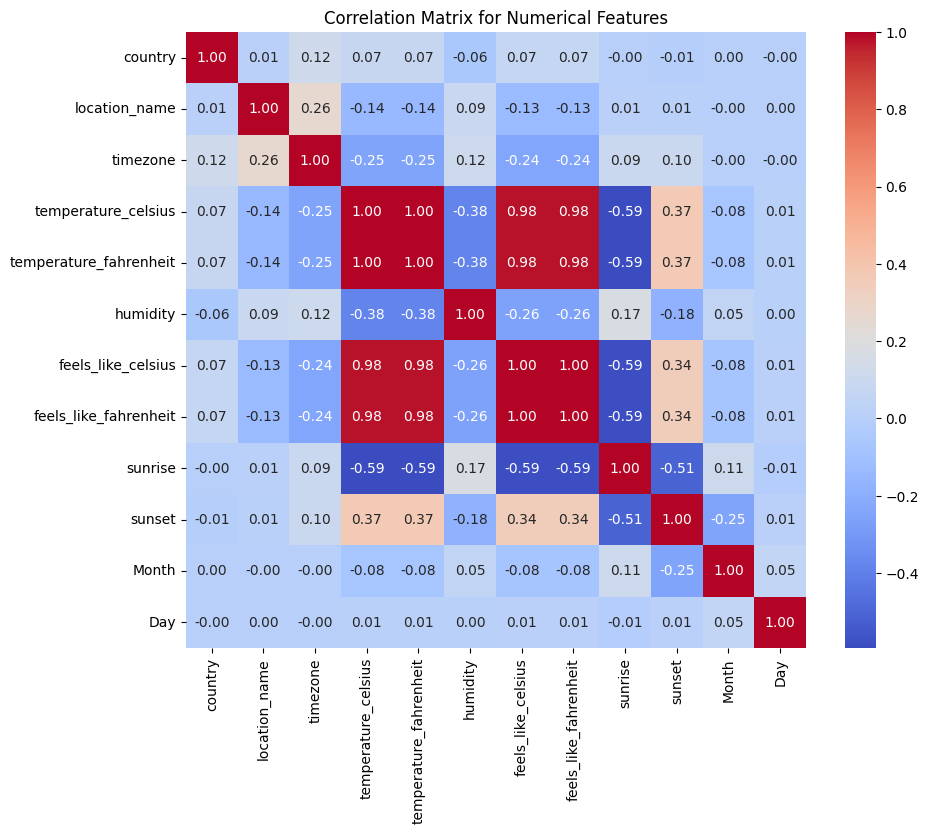
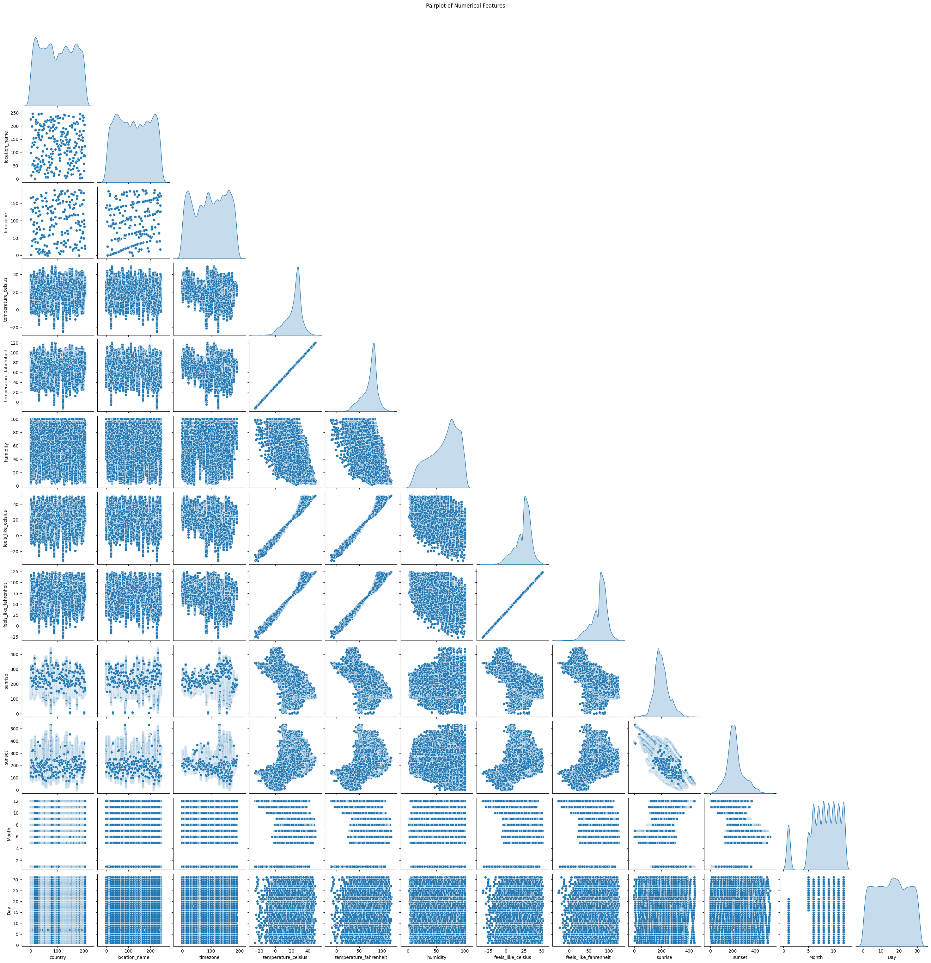
* **Temperature is highly correlated with humidity and time of day.**
* **Feature Importance Analysis:**
  + Time of day and geographical location were the most influential features.

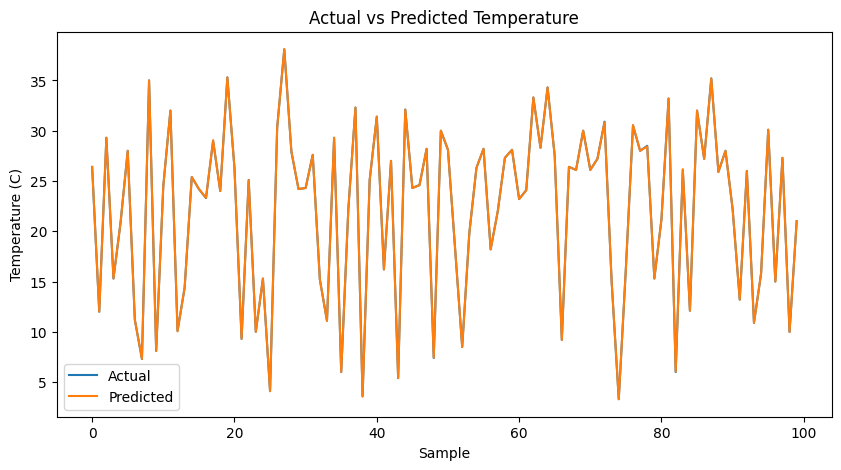
**Challenges & Future Work:**

* Expanding the dataset to include more meteorological variables (e.g., wind speed, precipitation).
* Implementing deep learning models (e.g., LSTMs) for improved forecasting.

**5. Visualizations**

Below are key graphs:

1. **Temperature Distribution:** Histogram of temperatures across different regions.****
2. **Model Performance Comparison:** Line plot comparing predicted vs. actual temperatures.

****

**6. Conclusion**

This project successfully developed a machine learning model to predict weather conditions using past meteorological data. Future enhancements can improve accuracy by integrating more sophisticated models and larger datasets.

**7. References & Submission Details**

**GitHub Repository:**  
A detailed README.md with installation steps, code structure, and analysis is included in the repository.