Weather Data Analysis Project

1. Project Description:

In this project, you will perform a comprehensive analysis of a weather dataset. Your task will involve data preprocessing, exploratory data analysis (EDA), feature engineering, visualizations, and data transformations using various Python tools. The goal of this project is to gain a deeper understanding of the data, uncover patterns, and derive insights from the weather conditions.

2. Data:

weather.csv(Provided in Brightspace)

3. Project Function (Python Resources to Use):

You are expected to utilize the following Python libraries to complete the project:

- **Pandas**
- **Numpy**
- **Matplotlib** and **Seaborn**(For data visualization)
- **Sklearn**
- **Regular Expression Library**

4. Project Goals (100 Marks)

    1. Data Preprocessing (20 Marks)

        - Handle missing values (10 Marks)

        - Convert categorical variables (RainToday, RainTomorrow, WindDir9am) to numerical variables / Ordinal Encoding (10 Marks)

    2. Exploratory Data Analysis (EDA) (20 Marks)

        - Summary statistics and dataset description (10 Marks)

        - Visualize distributions of key variables (e.g., temperature, humidity) (10 Marks)

    3. Feature Engineering (20 Marks)

        - Normalize temperature and humidity columns (5 Marks)

        - Create `TempRange` (MaxTemp - MinTemp) and `AvgHumidity` (average of Humidity9am and Humidity3pm) (15 Marks)

    4. Advanced Visualizations (15 Marks)

        - Scatterplot: TempRange vs. Rainfall (7 Marks)

        - Boxplot: Sunshine vs. RainTomorrow (8 Marks)

5. Correlation Analysis (10 Marks)

    - Create a correlation heatmap, including new features (10 Marks)

6. Regular Expressions (15 Marks)

    - Extract wind directions starting with 'N' (5 Marks)

    - Clean WindDir9am column using regex (10 Marks)

Expected Outputs:

1. Cleaned dataset with no missing values.
2. Visualizations:
    a. Count plot of RainTomorrow
    b. Scatterplot: TempRange vs. Rainfall
    c. Boxplot: Sunshine vs. RainTomorrow
    d. Correlation heatmap with new features
3. New features (`TempRange`, `AvgHumidity`) added to dataset.
4. Regex-based manipulations of wind direction data.

Bonus Section: Naive Bayes Classification (Optional for Extra Credit - 20 Marks)

In this bonus task, you will predict whether it will rain tomorrow (`RainTomorrow`) using weather data. You'll preprocess the dataset, build a classification model, and evaluate its performance using relevant metrics. Focus on data preparation, model accuracy, and visualizing the results.

Key Outputs and Marking (20 Marks Total):

1. Data Preparation (5 Marks): Clean and preprocess the dataset, handling missing values and selecting appropriate features.

2. Model Implementation(5 Marks): Build a Naive Bayes classifier to predict `RainTomorrow`.

3. Evaluation(5 Marks): Measure the performance using accuracy, confusion matrix, and other relevant metrics.

4. Visualization (3 Marks): Visualize key results, including a confusion matrix heatmap.

5. Analysis & Insights (2 Marks): Provide insights based on the model's performance.