AI / Machine Learning Data Analysis Report

1. Data Collection Method

The dataset represents synthetic weather measurements for Somalia. Values were generated based on realistic climate ranges such as temperature, humidity, rainfall, wind speed, atmospheric pressure, and solar radiation. This synthetic approach ensures consistency and suitability for learning and experimentation.

2. Data Preparation

The dataset was structured into rows (daily records) and columns (features). Data cleaning ensured no missing values, consistent units, and realistic ranges. The dataset was then formatted into Excel and PDF-friendly structures.

3. Existing Data Problems

Missing Values: None present, but real-world data may require imputation.
Data Types: All numerical and correctly formatted.
Duplicates: No duplicate records exist.
Class Imbalance: Rain-related labels are imbalanced, with more dry days than rainy days.
Noise & Bias: Limited seasonal variation due to synthetic generation.

4. Input Variables (Features)

Temperature_C (°C), Humidity (%), WindSpeed (km/h), Pressure (hPa), SolarRadiation (W/m²). These variables act as independent inputs (X).

5. Output Variable (Label)

RainTomorrow (binary): 1 if rainfall exceeds 1 mm, otherwise 0. This is the dependent variable (y).

6. Machine Learning Applications

Regression: Predict rainfall amount using weather features.
Classification: Predict whether it will rain tomorrow.
Clustering: Group similar weather patterns without labels.

7. Conclusion

The dataset is clean, structured, and well-suited for educational AI and machine learning projects, including regression, classification, and clustering tasks.