Rain Prediction Model - Final Report

1. Introduction

Cruella de Vil, a renowned fashion designer, faces a challenge in protecting her **vegan fur coats** from rain damage. The goal of this project is to **predict rainy days with at least 80% accuracy** to help her plan her wardrobe before **Paris Fashion Week (October 2024).**

2. Problem Statement

The weather conditions significantly impact the **quality and appearance of Cruella's fur coats**. Rain exposure makes the fur rough and difficult to detangle, resulting in a cheap look. Our objective is to **develop a weather prediction model** that forecasts rain with high accuracy, enabling proactive protection of her designs.

3. Data Understanding & Wrangling

Dataset Overview

- The dataset contains **weather records** from **multiple cities over two years (2024-2025).**
- The data includes features such as **temperature, humidity, wind speed, precipitation, cloud cover, and pressure.**

Key Data Cleaning Steps:

- **Date Conversion**: Transformed the `Date` column into `datetime` format.
- **Duplicate Handling**: Identified and investigated duplicate `Date-Location` entries. Found **five recordings per day** per city, likely due to multiple daily observations.
- **Missing Values**: Checked for null values—no significant issues found.
- **Outliers**: Initial analysis did not highlight major anomalies.

4. Exploratory Data Analysis (EDA)

EDA helped uncover patterns in the data:

- **Rainfall Distribution**: Examined the frequency of rainy vs. non-rainy days.
- **Feature Correlations**:
    - **Humidity and precipitation** showed strong positive correlation.

    - **Pressure and precipitation** had an inverse relationship.
- # Hypothesis Testing:

- Determined whether the rain data recorded on the same dates represented multiple readings taken throughout the day or separate days, using variations in recorded temperatures for verification.

5. Feature Engineering

To enhance model performance, new features were created:

- **Humidity-Temperature Interaction**: Captured the combined effect of humidity and temperature.
- **Precipitation Binning**: Grouped precipitation into `Low` and `High` categories.
- **Pressure & Temperature Categories**: Categorical labels assigned to simplify numerical features.
- **Wind Speed Grouping**: Classified wind speeds into descriptive categories (`Calm`, `Breezy`, etc.).
- **Rain Indicator**: A binary feature indicating the presence of rain.

6. Modeling & Evaluation

Models Tested:

- **Decision Tree Classifier**
- **Random Forest Classifier**
- **Logistic Regression**

Training & Evaluation Steps:

- **One-Hot Encoding** applied to categorical variables.
- **Train-Test Split:** 80% training, 20% testing.
- **Hyperparameter Tuning:**
    - **RandomizedSearchCV** and **Bayesian Optimization** used for parameter tuning.
- **Performance Metrics:**
    - **Accuracy, Precision, Recall, and F1-score** were evaluated.
    - **Confusion Matrix** provided insights into false positives and false negatives.

7. Findings & Recommendations

Key Findings:

- **Decision Tree was selected as the best model** due to its **higher accuracy (85%), reliance on only two features, and excellent interpretability.**

- **Humidity and precipitation** showed a **strong positive correlation (above 0.7) with rainfall.**

- **Atmospheric Pressure had an inverse correlation (-0.6) with rainfall**, indicating lower pressure is associated with rain.

- **Hyperparameter tuning** improved Decision Tree performance, achieving **Recall: 82%, Precision: 87%, and F1-score: 84%.**

Recommendations:

1. **Deploy Decision Tree as the Final Model**: Its simplicity, high accuracy, and interpretability make it the best choice.

2. **Improve Data Collection Methods**: Collect additional hourly weather observations for better granularity.

3. **Monitor and Retrain Regularly**: Retrain every 6 months with updated data to maintain accuracy.

4. **Enhance Data Sources**: Integrate external weather APIs such as NOAA or OpenWeather.

5. **Future Enhancements**: Explore ensemble techniques to improve performance further.

8. Conclusion

The project successfully developed a rain prediction model to aid **Cruella de Vil** in preserving her fashion assets. The model provides a **data-driven approach** to mitigate weather-related damage. Future improvements could include **deep learning models** or **ensemble techniques** for even greater accuracy.