Weather Forecasting Using Machine Learning

March 9, 2025

1 Introduction

Accurate weather forecasting is crucial for farmers to optimize irrigation, planting, and harvesting schedules. Traditional weather predictions often fail at hyper-local scales, making machine learning a valuable alternative. This project aims to predict rainfall using historical weather data and machine learning models.

2 Dataset Description

The dataset contains daily weather observations for 300 days with the following attributes:

- **avg_temperature**: Average temperature in °C
- **humidity**: Humidity in percentage
- **avg_wind_speed**: Average wind speed in km/h
- **cloud_cover**: Cloud coverage percentage
- **pressure**: Atmospheric pressure
- **rain_or_not**: Binary target variable (1 = Rain, 0 = No Rain)
- **date**: Date of observation

3 Data Preprocessing

- **Handling Missing Values**: Used mean imputation for numerical features.
- **Encoding Categorical Variables**: Converted ‘rain_or_not‘ into binary format using label encoding.
- **Feature Scaling**: Applied StandardScaler to normalize numerical features.
- **Data Splitting**: Split into training (80%) and testing (20%) sets.

4 Exploratory Data Analysis (EDA)

- **Target Distribution**: Visualized rain occurrence using count plots.
- **Feature Correlations**: Computed correlation matrix and plotted heatmap.
- **Pairwise Relationships**: Used pair plots to explore dependencies between variables.

5 Machine Learning Models

We trained and evaluated the following models:

- **Logistic Regression**: Baseline classifier.
- **Decision Tree**: Captures non-linear dependencies.
- **Random Forest**: Reduces overfitting and improves accuracy.
- **Gradient Boosting**: Boosting technique for higher predictive performance.

5.1 Model Performance

Table 1 shows the accuracy scores of different models.

Model	**Accuracy**
Logistic Regression	0.85
Decision Tree	0.80
Random Forest	0.88
Gradient Boosting	0.90

Table 1: Accuracy Scores of Different Models

6 Hyperparameter Tuning and Feature Engineering

- **Random Forest Hyperparameter Tuning**: Used GridSearchCV to optimize hyperparameters like ‘n_estimators‘, ‘max_depth‘, ‘min_samples_split‘, and ‘min_samples_leaf‘.
- **Feature Engineering**: Generated polynomial features to capture higher-order interactions.

7 Final Output

The final model predicts rainfall probabilities for future 21 days. The probability values are converted into binary predictions (0 = No Rain, 1 = Rain) based on a threshold of 0.5.

8 Conclusion and Future Work

- The Gradient Boosting model achieved the highest accuracy of 90%.
- Future improvements:
  - Integrate additional meteorological data such as wind direction.
  - Implement deep learning models (e.g., LSTMs) for sequential forecasting.
  - Deploy the model as a real-time API for continuous predictions.