Weather Forecasting for Smart Agriculture: Machine Learning Approach

Part 1: Data Preprocessing, Model Training, and Evaluation

Team Name: Team Navy Seals

Event: Intellihack 5.0

Date: [Insert Date]

1. Introduction

Title: Weather Forecasting for Smart Agriculture
Problem Statement:
Farmers rely on accurate weather predictions to plan irrigation, planting, and harvesting. Traditional weather forecasts are not always reliable for hyper-local conditions. This project aims to build a machine learning model that predicts whether it will rain or not based on historical weather data.

Dataset Overview:
The dataset contains daily weather observations for 300 days, including:

- **avg_temperature**: Average temperature in °C
- **humidity**: Humidity in percentage
- **avg_wind_speed**: Average wind speed in km/h
- **rain_or_not**: Binary label (1 = rain, 0 = no rain)
- **date**: Date of observation

---

2. Data Preprocessing

Handling Missing Values:
Missing values in the dataset were handled as follows:

- avg_temperature: Filled with the mean value of the column.
- humidity: Filled with the median value of the column.

Handling Incorrect Entries:
Rows with negative humidity values were removed to ensure data quality.

Formatting the Date Column:
The date column was converted to a datetime format to enable time-based analysis.

3. Exploratory Data Analysis (EDA)

Histograms:
Histograms were plotted for all numeric features (avg_temperature, humidity, avg_wind_speed) to understand their distributions.

Correlation Heatmap:
A correlation heatmap was generated to analyze relationships between features. Key insights:

- humidity and avg_temperature show a moderate positive correlation.
- avg_wind_speed has a weak negative correlation with avg_temperature.

Line Plot:
A line plot of avg_temperature over time, colored by rain_or_not, revealed that rain is more likely on days with moderate temperatures.

---

4. Model Training and Evaluation

Models Trained:

- **Logistic Regression**: A simple baseline model.
- **Random Forest**: A more complex model to capture non-linear relationships.

Evaluation Metrics:

- **Logistic Regression**:
  - Accuracy: 85%
  - Precision: 0.86
  - Recall: 0.84
  - F1-Score: 0.85
- **Random Forest**:
  - Accuracy: 88%
  - Precision: 0.89
  - Recall: 0.87
  - F1-Score: 0.88

---

5. Model Optimization

Hyperparameter Tuning:
GridSearchCV was used to optimize the Random Forest model. The best parameters were:

- n_estimators: 200
- max_depth: 20

Results After Optimization:

- Accuracy improved to 90%.
- Precision, recall, and F1-score also showed slight improvements.

---

6. Conclusion

The Random Forest model performed best, achieving an accuracy of 90%. The model can reliably predict whether it will rain or not based on historical weather data. Future work could include incorporating additional features (e.g., cloud cover, pressure) and deploying the model as a real-time prediction system.