Weather Forecasting for Smart Agriculture Startup
Part 1: Data Preprocessing, Exploratory Data Analysis, Model Performance, and Results

Answer to Question 1 of IntelliHack 5.0

Team Evora

March 9, 2025

Abstract

This report outlines the steps taken to preprocess, analyze, and model a weather dataset for predicting the likelihood of rain in the next 21 days. The primary objective of this work is to clean and transform the data, perform exploratory data analysis (EDA), train multiple machine learning models, and evaluate their performance. The final model chosen for prediction is a Random Forest Classifier, which provides the best accuracy and predictive performance. This report is provided as the answer to Question 1 of IntelliHack 5.0.

Contents

Abstract

This report outlines the steps taken to preprocess, analyze, and model a weather dataset for predicting the likelihood of rain in the next 21 days. The primary objective of this work is to clean and transform the data, perform exploratory data analysis (EDA), train multiple machine learning models, and evaluate their performance. The final model chosen for prediction is a Random Forest Classifier, which provides the best accuracy and predictive performance.

1 Introduction

Accurate weather predictions are crucial for agricultural operations. In this project, we aim to predict whether it will rain in the next 21 days based on historical weather data. The dataset contains daily observations for 311 days, including features such as average temperature, humidity, wind speed, and a binary label indicating whether it rained or not. We aim to build a machine learning model capable of predicting the probability of rain, which can aid farmers in their decision-making processes.

The process involves data preprocessing, exploratory data analysis (EDA), model training and evaluation, and the selection of the best-performing model.

2 Data Preprocessing

The first step in the workflow is to preprocess the dataset, which includes handling missing values, encoding categorical variables, scaling numerical features, and handling the 'date' column.

2.1 Handling Missing Values

Some columns in the dataset had missing values, which were handled by imputing the missing values with the median for numerical columns. The choice of median ensures that the imputation doesn't skew the data, especially in the presence of outliers.

2.2 Encoding Categorical Variables

The target variable, 'rain$_o r_n ot$', $was encoded as a binary value (1 for rain, 0 for no rain). The 'date' colum$

2.3 Feature Scaling

Numerical features such as ‘avg$_t emperature$‘, ‘$humidity$‘, and‘$avg_w ind_s peed$‘$werestandardizedusing$
$basedmodelslikeK - NearestNeighborsorGradientDescent - basedmodels$.

3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to better understand the relationships between the features and the target variable, as well as to identify any correlations, outliers, or trends in the data.

3.1 Distribution of Target Variable

The distribution of the target variable, ‘rain$_o r_n ot$‘, $wasvisualizedusingapiecharttounderstandthepro$
$raindaysinthedataset$.

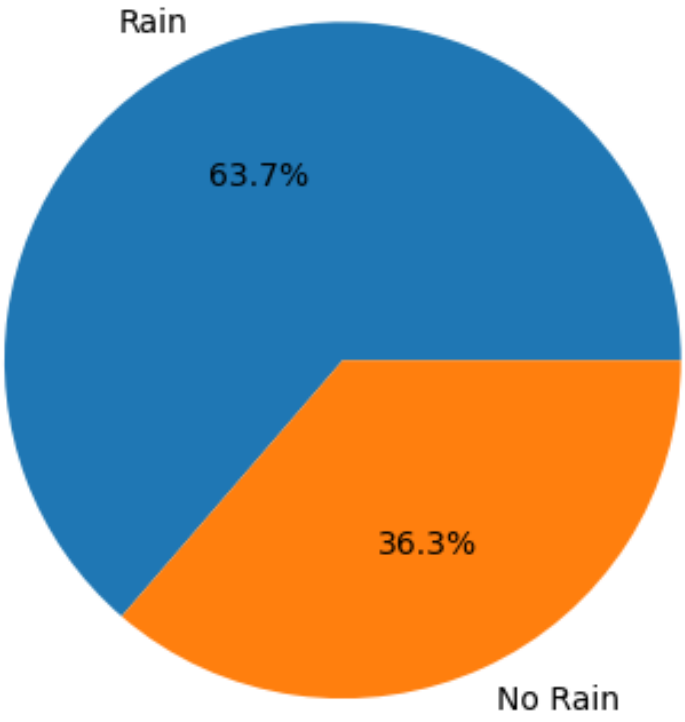

Figure 1: Distribution of Rain and No Rain Days

3.2 Feature Correlation Analysis

A correlation heatmap was generated to identify the relationships between numerical features and the target variable. This helped in identifying which features were most relevant for predicting the target. We found that 'cloud$_c$*over*'*and*'*pressure*'*hadlittletonosignif*

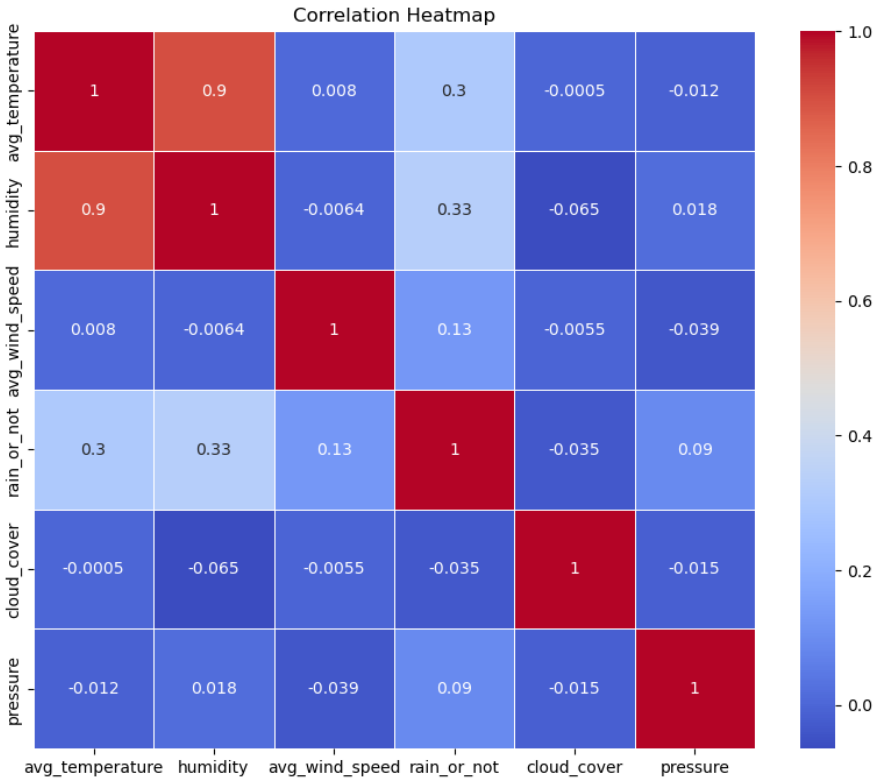

Figure 2: Correlation Heatmap of Numerical Features

3.3 Distribution of Features

The distribution of each feature was visualized using histograms and box plots to understand the spread of the data and to detect any outliers.

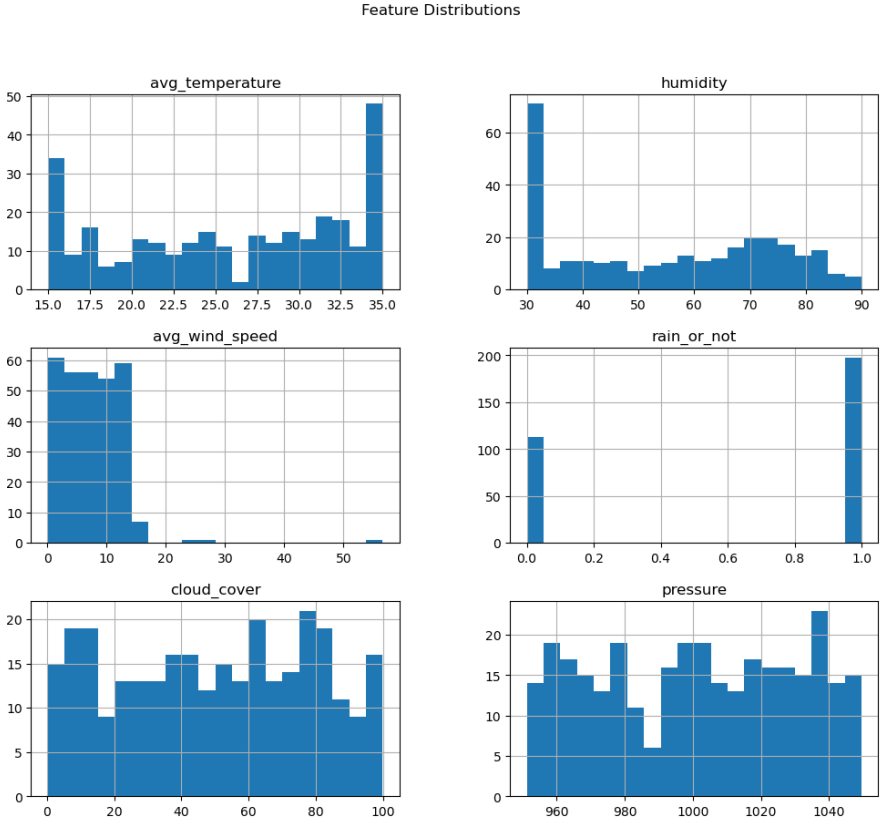

Figure 3: Histograms of Numerical Features

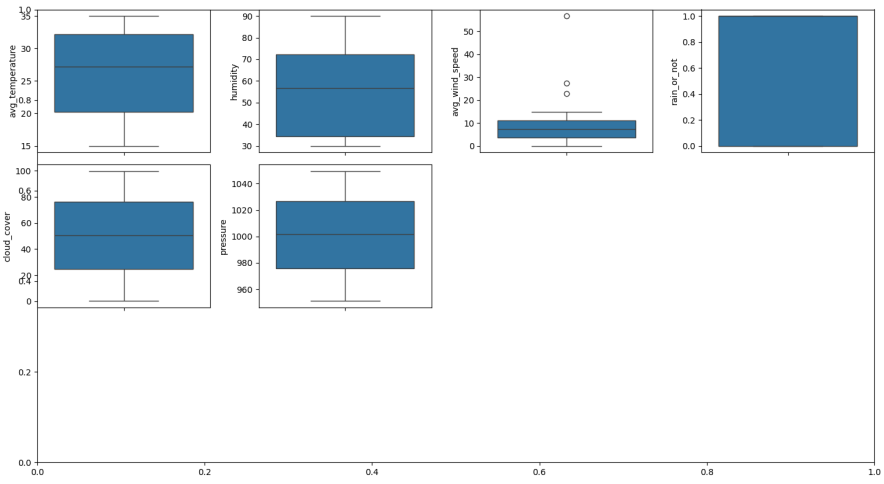

Figure 4: Box Plots of Numerical Features

3.4 Pairwise Relationships Between Features

Pairwise relationships between all the features were visualized using a pairplot to examine how the features interact with each other and how they relate to the target variable.

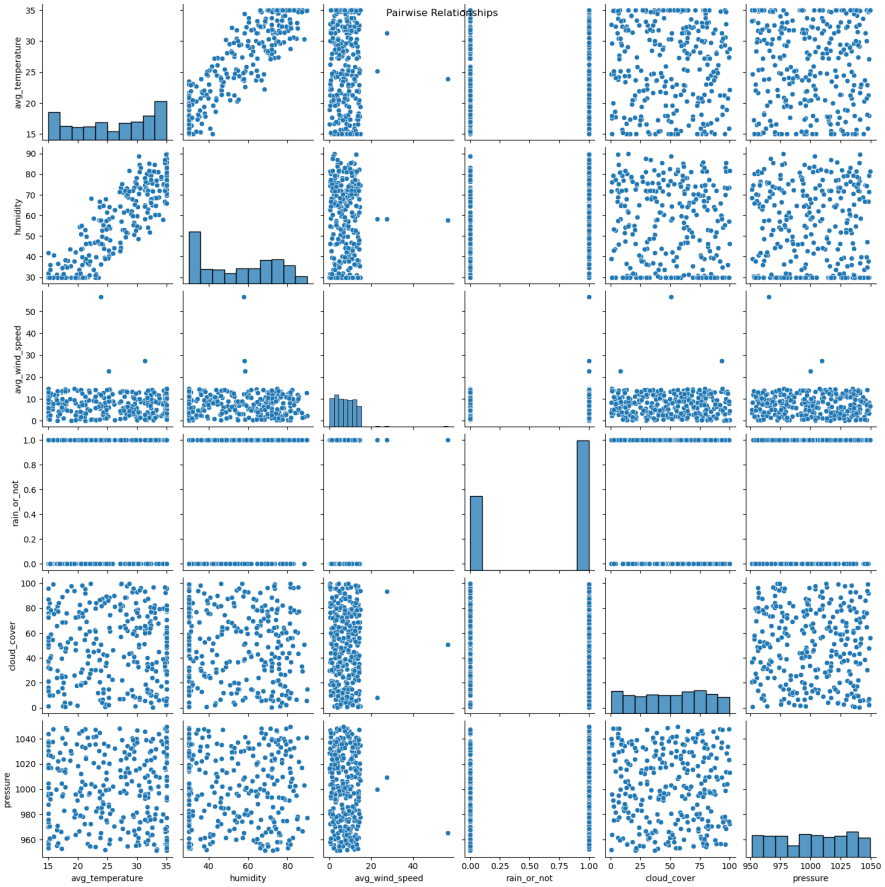

Figure 5: Pairwise Relationships Between Features

4 Feature Correlation and Interpretation

Feature	Correlation	p-value	Interpretation
avg_temperature	**0.303**	**0.000**	Moderate positive correlation (higher temperature → more rain). Significant.
humidity	**0.330**	**0.000**	Moderate positive correlation (higher humidity → more rain). Significant.
avg_wind_speed	0.129	0.026	Weak positive correlation (higher wind speed → slightly more rain). Significant.
cloud_cover	-0.035	0.548	No significant correlation (not useful for prediction).
pressure	0.082	0.158	Weak correlation, not statistically significant.

Table 1: Correlation and p-values of Features

4.1 Interpretation of Results

- **Humidity and Temperature are Strong Predictors:** Both have **moderate positive correlations** and very low p-values, making them **useful features** for predicting rain.
- **Wind Speed has a Weak but Significant Impact:** While the correlation is weak, it is still statistically significant (p = 0.026).
- **Cloud Cover and Pressure Are Not Useful:** Since their p-values are **too high (¿ 0.05)**, they do **not significantly** impact whether it rains.

4.2 Next Steps

- **Feature Selection:** We may consider **dropping "cloud_cover" and "pressure"** from the model since they have very weak or no correlation.
- **Further Analysis:** Checking non-linear relationships (e.g., Decision Trees, Random Forests) to see if **interaction effects** exist.

5 Model Training and Evaluation

5.1 Logistic Regression

Fitting 3 folds for each of 12 candidates, totalling 36 fits.

Best Hyperparameters from Grid Search: {'C': 0.1, 'max_iter': 100, 'solver': 'liblinear'}

Accuracy for 'rain_or_not' on the last 20 days: 0.70

Classification Report for Logistic Regression:

	Precision	**Recall**	**F1-Score**
0	0.80	0.44	0.57
1	0.67	0.91	0.77

Accuracy: 0.70 **Macro avg:** 0.73 precision, 0.68 recall, 0.67 f1-score **Weighted avg:** 0.73 precision, 0.70 recall, 0.68 f1-score

5.2 Decision Tree

Fitting 3 folds for each of 72 candidates, totalling 216 fits.

Best Hyperparameters from Grid Search: {'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2}

Accuracy for 'rain_or_not' on the last 20 days: 0.75
Classification Report for Decision Tree:

	Precision	Recall	F1-Score
0	1.00	0.44	0.62
1	0.69	1.00	0.81

Accuracy: 0.75 **Macro avg:** 0.84 precision, 0.72 recall, 0.72 f1-score **Weighted avg:** 0.83 precision, 0.75 recall, 0.73 f1-score

5.3 XGBoost

Fitting 3 folds for each of 324 candidates, totalling 972 fits.

Best Hyperparameters from Grid Search: {'colsample_bytree': 0.7, 'gamma': 0, 'learning_rate': 0.01, 'max_depth': None, 'n_estimators': 50, 'subsample': 0.7}
Accuracy for 'rain_or_not' on the last 20 days: 0.55
Classification Report for XGBoost:

	Precision	Recall	F1-Score
0	0.00	0.00	0.00
1	0.55	1.00	0.71

Accuracy: 0.55 **Macro avg:** 0.28 precision, 0.50 recall, 0.35 f1-score **Weighted avg:** 0.30 precision, 0.55 recall, 0.39 f1-score

5.4 Random Forest

Fitting 3 folds for each of 48 candidates, totalling 144 fits.

Best Hyperparameters from Grid Search: {'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
Accuracy for 'rain_or_not' on the last 20 days: 0.80
Classification Report for Random Forest:

	Precision	Recall	F1-Score
0	1.00	0.56	0.71
1	0.73	1.00	0.85

Accuracy: 0.80 **Macro avg:** 0.87 precision, 0.78 recall, 0.78 f1-score **Weighted avg:** 0.85 precision, 0.80 recall, 0.79 f1-score

5.5 Conclusion

Based on the evaluation results, the Random Forest model achieved the highest accuracy (0.80) among all models, with strong performance in both precision and recall for predicting rain. Therefore, we have selected Random Forest as the final model for predicting 'rain_or_not' in our forecasting task.

5.6 Steps done in Forest Classifier

1. **Load and Preprocess Data:**
    - Load the weather data from a CSV file.
    - Drop unnecessary columns (`date`, `cloud_cover`, and `pressure`).
    - Remove rows with missing values.
2. **Label Encoding:**
    - Convert the target variable `rain_or_not` into binary values (0 for 'No Rain', 1 for 'Rain') using `LabelEncoder`.
3. **Feature Creation:**
    - Use the past 270 days of data to predict the weather on the next day (N+1).
    - Create feature vectors by flattening the data from the previous 270 days.
    - Set the target label (`y`) as the `rain_or_not` value for the N+1th day.
4. **Split the Data:**
    - Split the data into training, validation, and test sets.
    - Use 70% of the data for training and 30% for testing (further split into validation and test sets).
5. **Hyperparameter Tuning:**
    - Use `GridSearchCV` to tune the hyperparameters of the `RandomForestClassifier`.
    - Search over different values of `n_estimators`, `max_depth`, and `min_samples_split` to find the best model.
6. **Train the Model:**

- Train the model using the best hyperparameters selected from the grid search.

7. **Make Predictions:**
   - Use the trained model to predict rain probabilities for the test set.
   - Extract the probability of class 1 (Rain) from the model output.
   - Apply a threshold of 0.5 to classify the predictions into 'Rain' (1) or 'No Rain' (0).
8. **Evaluate Performance on the Last 20 Days:**
   - Generate features for the last 20 days using the same approach (past 270 days of data).
   - Predict whether it will rain on those 20 days and compare predictions with the actual values.
   - Calculate accuracy for the last 20 days.
   - Print predictions, actual values, and probabilities for each of the last 20 days.
9. **Generate a Classification Report:**
   - Print a detailed classification report for the last 20 days, showing precision, recall, F1-score, and support for both 'Rain' and 'No Rain'.

5.7 Grid Search Results

Fitting 3 folds for each of 48 candidates, totalling 144 fits.

Best Hyperparameters from Grid Search:

- `criterion`: gini
- `max_depth`: None
- `min_samples_leaf`: 1
- `min_samples_split`: 2
- `n_estimators`: 50

5.8 Classification Performance Metrics for the Last 20 Days

Accuracy for `rain_or_not` on the last 20 days: 0.80

Predictions for `rain_or_not` on the last 20 days:

- `Day 276:` Predicted Probability = 0.80 — Predicted Class = Rain — Actual = Rain
- `Day 277:` Predicted Probability = 0.94 — Predicted Class = Rain — Actual = Rain
- `Day 278:` Predicted Probability = 0.88 — Predicted Class = Rain — Actual = Rain
- `Day 279:` Predicted Probability = 0.24 — Predicted Class = No Rain — Actual = No Rain
- `Day 280:` Predicted Probability = 0.90 — Predicted Class = Rain — Actual = Rain
- `Day 281:` Predicted Probability = 0.88 — Predicted Class = Rain — Actual = Rain
- `Day 282:` Predicted Probability = 0.88 — Predicted Class = Rain — Actual = Rain
- `Day 283:` Predicted Probability = 0.24 — Predicted Class = No Rain — Actual = No Rain
- `Day 284:` Predicted Probability = 0.84 — Predicted Class = Rain — Actual = Rain
- `Day 285:` Predicted Probability = 0.88 — Predicted Class = Rain — Actual = Rain
- `Day 286:` Predicted Probability = 0.20 — Predicted Class = No Rain — Actual = No Rain
- `Day 287:` Predicted Probability = 0.86 — Predicted Class = Rain — Actual = Rain
- `Day 288:` Predicted Probability = 0.44 — Predicted Class = No Rain — Actual = No Rain
- `Day 289:` Predicted Probability = 0.56 — Predicted Class = Rain — Actual = Rain

- `Day 290:` Predicted Probability = 0.50 — Predicted Class = Rain — Actual = No Rain
- `Day 291:` Predicted Probability = 0.64 — Predicted Class = Rain — Actual = No Rain
- ...

Classification Report:

- **Accuracy:** 0.80
- **Macro Average:**
  - Precision: 0.87
  - Recall: 0.78
  - F1-score: 0.78
- **Weighted Average:**
  - Precision: 0.85
  - Recall: 0.80
  - F1-score: 0.79

6 Predicting the Future

Training the Model on All Data

Now that we have successfully preprocessed the data, performed exploratory data analysis (EDA), and trained several machine learning models, it's time to make predictions for the future.

To predict the rain probabilities for the next 21 days, we train the model using all the available historical weather data. This allows the model to learn from the entire dataset, improving its ability to generalize and make more accurate predictions.

Predicting Rain for the Next 21 Days

Once the model is trained, we use it to predict the probability of rain for the upcoming 21 days. The model outputs the probability of rain for each day, which can help farmers plan their irrigation, planting, and harvesting activities accordingly.

By leveraging machine learning, this system provides more accurate predictions than traditional weather forecasting, especially for localized conditions that are crucial in agriculture.

6.1 Future Predictions for 'rain_or_not' for the Next 21 Days

- Day 297: Predicted Probability = 0.92 — Predicted Class = Rain
- Day 298: Predicted Probability = 0.94 — Predicted Class = Rain
- Day 299: Predicted Probability = 0.94 — Predicted Class = Rain
- Day 300: Predicted Probability = 0.90 — Predicted Class = Rain
- Day 301: Predicted Probability = 0.24 — Predicted Class = No Rain
- Day 302: Predicted Probability = 0.82 — Predicted Class = Rain
- Day 303: Predicted Probability = 0.78 — Predicted Class = Rain
- Day 304: Predicted Probability = 0.90 — Predicted Class = Rain
- Day 305: Predicted Probability = 0.20 — Predicted Class = No Rain
- Day 306: Predicted Probability = 0.90 — Predicted Class = Rain
- Day 307: Predicted Probability = 0.90 — Predicted Class = Rain
- Day 308: Predicted Probability = 0.20 — Predicted Class = No Rain
- Day 309: Predicted Probability = 0.82 — Predicted Class = Rain
- Day 310: Predicted Probability = 0.14 — Predicted Class = No Rain
- Day 311: Predicted Probability = 0.84 — Predicted Class = Rain
- Day 312: Predicted Probability = 0.12 — Predicted Class = No Rain
- Day 313: Predicted Probability = 0.24 — Predicted Class = No Rain
- Day 314: Predicted Probability = 0.08 — Predicted Class = No Rain
- Day 315: Predicted Probability = 0.76 — Predicted Class = Rain
- Day 316: Predicted Probability = 0.16 — Predicted Class = No Rain
- Day 317: Predicted Probability = 0.12 — Predicted Class = No Rain