Weather Forecasting Model for Smart Agriculture

Predictive Rain Analysis for Hyper-Local Conditions

Executive Summary

This report details the development of a machine learning model designed to predict rainfall for smart agriculture applications. Traditional weather forecasts often lack the precision needed for farming decisions at a local level. Our model addresses this gap by providing hyper-local rain predictions based on historical weather data.

The model achieves high accuracy and reliability by implementing a comprehensive data science workflow, from data preprocessing to model optimization. The final solution delivers daily rain probability forecasts for a 21-day period, enabling farmers to make informed decisions about irrigation, planting, and harvesting.

1. Data Preprocessing

1.1 Initial Data Assessment

The dataset contained daily weather observations for 300 days with the following variables:

- `avg_temperature`: Average temperature in °C
- `humidity`: Humidity in percentage
- `avg_wind_speed`: Average wind speed in km/h
- `cloud_cover`: Cloud cover percentage
- `pressure`: Atmospheric pressure
- `rain_or_not`: Binary label (Rain/No Rain)
- `date`: Date of observation

Initial inspection revealed several data quality issues:

- Missing values in key weather parameters
- Outliers in temperature, humidity, and wind speed measurements
- Inconsistent date formatting

1.2 Data Cleaning Process

The following preprocessing steps were implemented:

Date Handling:

- Converted string dates to datetime objects
- Extracted temporal features: day of year, month, and day

Outlier Treatment:

- Created boxplots to visualize outliers
- Applied percentile-based clipping (1st to 99th percentile) to handle extreme values
- This approach preserved the natural variation in weather data while removing erroneous measurements

Missing Value Imputation:

- Applied median imputation for numerical features
- This method was chosen as it's robust to outliers and preserves the data distribution

The preprocessing resulted in a clean dataset with consistent formatting and no missing values, providing a solid foundation for model development.

2. Exploratory Data Analysis

2.1 Target Variable Distribution

Analysis of the target variable showed:

- Approximately 30% of days had rainfall
- 70% of days were classified as no rain
- This slight class imbalance was addressed in the modeling phase

2.2 Feature Relationships

Correlation Analysis:

- Humidity showed the strongest positive correlation with rainfall (0.68)
- Pressure demonstrated a strong negative correlation with rainfall (-0.59)
- Temperature had a moderate correlation with rainfall (0.32)

Feature Distributions by Rain Status:

- Rainy days typically had:
  - Higher humidity levels (mean: 78% vs. 62% for non-rainy days)
  - Lower atmospheric pressure
  - Increased cloud cover
  - Moderately higher wind speeds

Seasonal Patterns:

- Monthly rainfall probability analysis revealed:
  - Highest rainfall probability in July-August (monsoon season)
  - Lowest rainfall in November-December (dry season)
  - Transitional patterns in March-April and September-October

This analysis provided valuable insights for feature engineering and model development, highlighting the key weather parameters most indicative of rainfall.

3. Feature Engineering

3.1 Temporal Features

- Created cyclical features from date components:
    - `sin_day` and `cos_day`: Sine and cosine transformations of day of year
    - These features capture the cyclical nature of seasonal weather patterns

3.2 Interaction Features

- Generated interaction terms to capture complex relationships:
    - `temp_humidity_interaction`: Product of temperature and humidity
    - `wind_pressure_interaction`: Product of wind speed and pressure
    - These interactions represent weather phenomena that aren't captured by individual variables alone

3.3 Feature Scaling

- Applied StandardScaler to normalize all features
- This ensured that all variables contributed equally to the model training process
- Particularly important for distance-based algorithms and regularized models

These engineered features enhanced the model's ability to capture complex weather patterns and seasonal variations.

4. Model Development and Evaluation

4.1 Modeling Approach

Five different algorithms were trained and evaluated:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Gradient Boosting
5. XGBoost

4.2 Evaluation Metrics

Models were evaluated using multiple metrics to ensure comprehensive assessment:

- Accuracy: Overall correctness
- Precision: Reliability of positive predictions
- Recall: Ability to detect rainy days
- F1 Score: Harmonic mean of precision and recall
- AUC-ROC: Model's ability to distinguish between classes

4.3 Results Comparison

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Regression	0.8350	0.7692	0.7143	0.7407	0.8786
Decision Tree	0.7600	0.6667	0.6429	0.6545	0.7571
Random Forest	0.8550	0.8519	0.6607	0.7442	0.9212
Gradient Boosting	0.8700	0.8438	0.7679	0.8042	0.9284
XGBoost	0.8850	0.8667	0.7857	0.8242	0.9357

XGBoost emerged as the best-performing model with the highest F1 score (0.8242) and AUC (0.9357).

4.4 Model Interpretation

Analysis of feature importance from the XGBoost model revealed:

1. Humidity (26.4%)
2. Pressure (21.3%)
3. Cloud Cover (18.7%)
4. Temperature-Humidity Interaction (12.5%)
5. Sin_Day (seasonal feature) (8.9%)

This aligned with meteorological understanding that humidity and pressure are primary indicators of precipitation.

5. Model Optimization

5.1 Hyperparameter Tuning

Grid search cross-validation was performed on the XGBoost model with the following parameters:

- n_estimators: [50, 100, 200]
- learning_rate: [0.01, 0.1, 0.2]
- max_depth: [3, 5, 7]
- subsample: [0.8, 1.0]
- colsample_bytree: [0.8, 1.0]

Optimal parameters:

- n_estimators: 100
- learning_rate: 0.1
- max_depth: 5
- subsample: 0.8
- colsample_bytree: 0.8

5.2 Performance Improvement

The tuned model showed significant improvement:

- Original XGBoost F1 Score: 0.8242
- Tuned XGBoost F1 Score: 0.8621
- Improvement: 4.6%

The optimized model balanced complexity with performance, avoiding overfitting while maintaining high predictive accuracy.

6. Rainfall Probability Forecasting

6.1 Forecast Generation

Using the optimized XGBoost model, a 21-day rainfall probability forecast was generated. The forecast includes:

- Daily rain probability (percentage)
- Binary prediction (Rain/No Rain)
- Visualization with color-coded bars indicating rain probability

6.2 Interpretation Guide

- Probability > 70%: High chance of rain, irrigation can likely be skipped
- Probability 40-70%: Moderate chance, consider reduced irrigation
- Probability < 40%: Low chance, irrigation planning should proceed
- Consecutive days with high probability: Potential for water logging, delay planting

6.3 Model Deployment

The model was saved using joblib for deployment, allowing:

- Easy loading in production environments
- Fast inference for new weather data
- Consistent scaling and preprocessing

7. Conclusions and Recommendations

7.1 Key Findings

- Humidity and pressure are the most influential factors for rainfall prediction
- Seasonal patterns significantly impact rainfall probability
- Feature interactions improve model performance
- XGBoost algorithm provides the best balance of precision and recall

7.2 Recommendations for Implementation

1. Deploy the model as part of an automated irrigation management system
2. Implement daily data collection and model updating
3. Consider combining model predictions with traditional forecasts for critical decisions
4. Monitor model performance and retrain periodically to account for climate changes

7.3 Future Improvements

- Incorporate additional data sources:
  - Satellite imagery
  - Soil moisture sensors
  - Regional weather patterns
- Develop ensemble models for further accuracy improvement
- Implement uncertainty quantification for risk assessment

This rainfall prediction model provides a powerful tool for agricultural decision-making, enabling farmers to optimize irrigation, planting, and harvesting based on reliable, hyper-local weather forecasts.