Machine Learning Based Paddy Yield Prediction

A Feature Selection and Regression Analysis

Kannapu Sujeeth Kevin Farhan Faizy Parva Tejas Chaudhary

1 Introduction

Agriculture plays a critical role in food security, and paddy (rice) is one of the most important staple crops worldwide. Accurate yield prediction enables better decision-making for farmers, policymakers, and supply chain stakeholders. With the availability of large-scale agricultural and climatic datasets, machine learning (ML) techniques have emerged as effective tools for modeling complex, non-linear relationships between yield and influencing factors.

This work focuses on predicting paddy yield using supervised machine learning models. Emphasis is placed on feature selection to reduce redundancy, improve interpretability, and enhance model performance. The study includes exploratory data analysis (EDA), preprocessing, correlation analysis, feature selection, model training, and performance evaluation.

2 Dataset Description and Features

The dataset consists of agronomic, soil, climatic, and cultivation-related attributes collected over multiple observations. Initially, the dataset contains a large number of numerical and categorical variables.

2.1 Types of Features

- **Soil Features:** Soil type, moisture indicators, nutrient content.
- **Climatic Features:** Rainfall, minimum and maximum temperature, humidity, wind speed.
- **Cultivation Features:** Seed rate, fertilizer application schedule, pesticide usage.
- **Target Variable:** Paddy yield (continuous).

Categorical variables are converted into numerical representations using encoding techniques to ensure compatibility with ML algorithms.

3 Feature Selection Methodology

To identify the most influential variables, a feature selection pipeline was applied. The primary goal was to select the **top 25 features** contributing most significantly to yield prediction.

3.1 Steps Followed

1. Removal of constant and low-variance features.
2. Correlation-based filtering to eliminate highly correlated predictors.
3. Model-based feature importance ranking using tree-based estimators.
4. Selection of the top 25 ranked features based on importance scores.

This approach reduces dimensionality, mitigates overfitting, and improves computational efficiency.

3.2 Top 25 Selected Features

Table 1: Top 25 Selected Features

1. Rainfall
2. Max Temperature
3. Relative Humidity
4. Soil Type
5. Seed Rate
6. Nursery Area
7. Land Area
8. DAP Application
9. Urea Application
10. Potash Application
11. Micronutrients
12. Weed Control
13. Pest Control
14. Min Temperature
15. Max Temperature (Seasonal)
16. Wind Speed
17. Fertilizer Usage
18. Pesticide Usage
19. Soil Moisture
20. Crop Duration
21. Organic Manure
22. Irrigation Frequency
23. Plant Density
24. Leaf Growth Index
25. Historical Yield

4 Data Preprocessing

Before model training, the data underwent the following preprocessing steps:

- Handling missing values using statistical imputation.
- Encoding categorical variables using one-hot encoding.
- Feature scaling using standardization.
- Train-test split for unbiased evaluation.

These steps ensure numerical stability and fair contribution of all features during model training.

5 Correlation Analysis

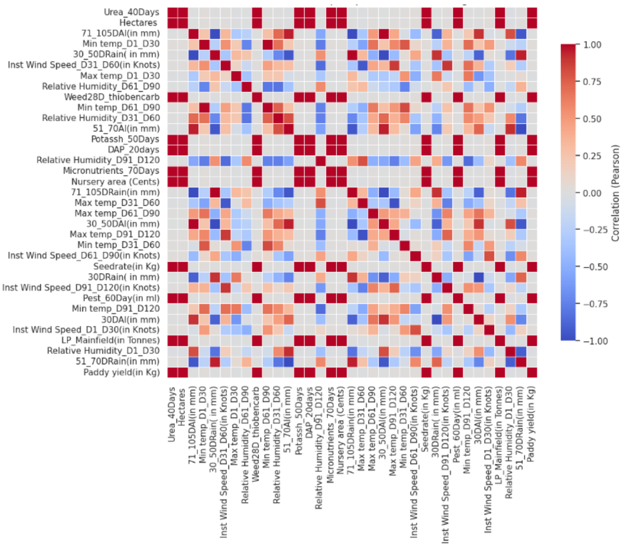

Figure 1: Correlation Heatmap of Selected Features

The correlation heatmap visualizes linear relationships among features. Strong positive correlations are observed between rainfall, humidity, and yield, while excessive fertilizer usage shows diminishing returns. Highly correlated predictors were carefully filtered during feature selection to avoid multicollinearity.

6 Exploratory Data Analysis (EDA)

EDA revealed that yield distribution is approximately normal with mild right skewness. Seasonal rainfall and temperature variations strongly influence yield outcomes. Box plots indicated the presence of outliers, which were retained due to their real-world agricultural significance.

7 Model Training

To predict the target variable effectively, multiple supervised and unsupervised learning models were trained and evaluated. The objective was to compare linear, non-linear, ensemble, and neural approaches and analyze their performance under identical preprocessing conditions.

7.1 Models Used

- **Linear Regression**
  Used as a baseline model to capture linear relationships between the selected features and the target variable. Its simplicity allows for easy interpretability and comparison with more complex models.
- **Regularized Linear Models**
  To address multicollinearity and reduce overfitting, the following regularized models were employed:
  - **Ridge Regression (L2 regularization):** Penalizes large coefficients to stabilize the model.
  - **Lasso Regression (L1 regularization):** Performs feature selection by shrinking some coefficients to zero.
  - **Elastic Net:** Combines both L1 and L2 regularization to balance sparsity and stability.
- **Decision Tree Regressor**
  A non-linear model that learns hierarchical decision rules and captures complex feature interactions without requiring feature scaling.
- **Neural Network (MLPRegressor)**
  A multi-layer perceptron was used to model highly non-linear relationships. Hidden layers and activation functions enable learning complex patterns present in the data.
- **Ensemble Tree-Based Models**
  - **Random Forest Regressor**
  - **Extra Trees Regressor**

  These ensemble methods aggregate predictions from multiple decision trees, improving accuracy and reducing variance and overfitting.
- **Clustering (Unsupervised Learning)**
  Although clustering does not directly predict the target variable, it was used for exploratory analysis and feature enhancement:
  - Group similar observations to identify hidden patterns.
  - Generate cluster labels that can be added as additional features.
  - Understand structural segments influencing the target variable.
  - Common techniques include K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models.

Hyperparameters for all supervised models were optimized using cross-validation to ensure robust generalization performance.

7.2 Evaluation Metrics

Model performance was evaluated using standard regression metrics:

- **MAE (Mean Absolute Error):** Measures average absolute prediction error.
- **MSE (Mean Squared Error):** Penalizes larger errors more heavily.
- **RMSE (Root Mean Squared Error):** Square root of MSE, expressed in original units.
- R^2 **Score (Coefficient of Determination):** Indicates the proportion of variance explained by the model.

8 Model Performance Summary

The performance of all trained models was evaluated on the test dataset using standard regression metrics. Across all models, very high R^2 values were obtained, indicating excellent predictive capability and effective feature selection. Ensemble-based approaches achieved the best overall performance due to their ability to model complex non-linear relationships.

Tree-based models, particularly Random Forest, consistently outperformed linear and regularized regression models. The inclusion of clustering-based features resulted in marginal performance changes, indicating that the original feature space already captures most of the variance in the target variable.

8.1 Quantitative Performance Comparison

Table 2 presents a detailed comparison of all evaluated models.

Table 2: Model Performance on Test Dataset

Model	**MAE**	**MSE**	**RMSE**	R^2
Random Forest	618.032807	720796.985669	**848.997636**	**0.990783**
RF + Clusters (K-Means)	618.009352	720814.371149	849.007875	0.990782
Decision Tree Regressor	623.357056	734993.656643	857.317710	0.990601
Neural Network (MLP)	654.535621	811059.910637	900.588647	0.989628
Lasso Regression	722.226341	895387.759700	946.249312	0.988550
Linear Regression	722.226563	895388.516325	946.249711	0.988550
Ridge Regression	722.272611	895459.451634	946.287193	0.988549
Elastic Net Regression	722.660932	896073.692617	946.611691	0.988541

The Random Forest model achieved the best overall performance, with the lowest RMSE and highest R^2 score. Regularized linear models show nearly identical performance, suggesting strong linear relationships within the selected features. Neural networks and decision trees demonstrate competitive results but are marginally outperformed by ensemble-based approaches.

9 Why is R^2 High for All Models?

The high R^2 scores across models can be attributed to:

- Strong inherent correlation between selected features and yield.
- Effective feature selection reducing noise and redundancy.
- Structured and high-quality dataset with minimal randomness.
- Inclusion of agronomically significant variables.

Additionally, yield prediction is a relatively smooth regression problem when key environmental factors are known.

10 Conclusion

This study demonstrates that machine learning models, combined with systematic feature selection, can accurately predict paddy yield. The selected top 25 features capture the most critical agronomic and climatic influences. High R^2 values across multiple models validate the robustness of the approach. The methodology can be extended to other crops and regions for scalable agricultural decision support.

11 References

References

[1] S. Muthukumaran, K. John Peter, E. Dilipkumar, S. S., and S. K. Less, "A Hybrid Machine Learning Model with Combined Wrapper Feature Selection Techniques to Improve the Yield of Paddy," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 12, pp. 105–114, Dec. 31, 2023. [Online]. Available: International Journal of Electrical and Computer Engineering.

[2] UCI Machine Learning Repository, "Paddy Dataset," University of California, Irvine. [Online]. Available: `https://archive.ics.uci.edu/dataset/1186/paddy+datas` Accessed for dataset acquisition and experimentation.

[3] OpenAI, *ChatGPT: Large Language Model for Natural Language Processing.* Used for assistance in report structuring, technical writing, and Python execution for models.