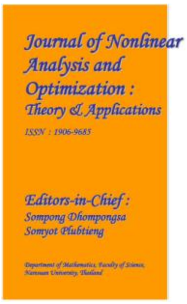
Journal of Nonlinear Analysis and Optimization
Vol. 16, Issue. 1, No.1 : 2025
ISSN : **1906-9685**



Journal of Nonlinear Analysis and Optimization : Theory & Applications
ISSN : 1906-9685
Editors-in-Chief :
Sompong Dhompongsa
Somyot Plubtieng
Department of Mathematics, Faculty of Science, Naresuan University, Thailand

PREDICTION OF AGRICULTURAL CROP YIELD IN INDIA USING MACHINE LEARNING ALGORITHM

Dr. M. Lalithambigai Assistant Professor, Department of Artificial Intelligence and Machine Learning Sri Krishna Adithya College of Arts and Science, Coimbatore
Rithanyaasri. M. V, B.Sc Artificial Intelligence and Machine Learning, Sri Krishna Adithya College of Arts and Science, Coimbatore
Niroshini. V, B.Sc Artificial Intelligence and Machine Learning, , Sri Krishna Adithya College of Arts and Science, Coimbatore

ABSTRACT

Agriculture is a cornerstone of the Indian economy, playing a vital role in ensuring food security. With the rapid increase in human population, crop yield prediction has become a critical challenge in agriculture. Crop yield is influenced by numerous factors, including weather conditions, rainfall, humidity, temperature, and pesticide usage. Additionally, having accurate information about historical crop yields is crucial for making reliable predictions and mitigating agricultural risks. Traditionally, crop yield predictions were based on farmers' experience with specific fields and crops. However, advancements in technology have enabled the use of machine learning models, such as the Random Forest algorithm, to analyze data patterns and predict crop yields more accurately. This study focuses on predicting agricultural yields in India by leveraging machine learning techniques, achieving an impressive accuracy rate of 98%.

KEYWORDS:
Prediction, Security, Machine Learning, Random Forest Algorithm, Prediction.

1.INTRODUCTION

India, as a predominantly agrarian nation, depends significantly on its agricultural sector for both sustenance and economic progress. However, the unpredictable nature of crop yields, influenced by factors like climate change, soil quality, pest infestations, and other environmental conditions, presents substantial challenges for both farmers and policymakers. The ability to accurately predict crop yields is essential for effective planning, resource distribution, risk management, and ensuring food security.In recent years, advancements in technologies such as data science, remote sensing, and machine learning have opened up new possibilities to enhance the accuracy of crop yield forecasts. These technologies allow for the integration of vast datasets, including historical yield information, weather data, soil quality evaluations, satellite imagery, and agricultural practices. By analyzing this wealth of data, predictive models can be constructed to offer more precise and dependable crop yield predictions. The dataset in question includes agricultural data for a variety of crops cultivated in several Indian states from 1997 to 2020. It encompasses vital features necessary for crop yield prediction, such as crop types, planting years, cropping seasons, states, cultivation areas, production volumes, annual rainfall, fertilizer usage, pesticide application, and calculated yields. This detailed dataset is an invaluable resource for agricultural experts, researchers, and data scientists involved in crop yield forecasting and agricultural analysis. It provides insights into the relationships between various agricultural factors (e.g., rainfall, fertilizer, pesticide use) and crop productivity across different regions and crop varieties. Researchers can leverage this data to build robust machine learning models for predicting crop yields and to identify trends in agricultural production. The analysis, using data from 1997 to 2020, including factors like annual rainfall, fertilizer, and pesticide usage, has been utilized to predict the yields of several crops such as maize, potato, coconut, cotton, and sweet potato. This prediction process has been carried out using machine learning techniques like the random forest regressor to enhance the accuracy of agricultural crop yield predictions in India.

2.DATA DESCRIPTION

Data description is the process of summarizing, organizing, and presenting information about a dataset to facilitate its understanding and analysis. Additionally, data description provides information such as variable names and data source information such as variable names and data source information, may be included to provide context and aid in data interpretation. Effective data description is a fundamental step in data analysis and plays a crucial role in informing decision-making and drawing insights from the data.

The dataset comprises information regarding agricultural crop production in various states of India, spanning different years and seasons. Each entry includes details such as the type of crop, the year of cultivation, the season of cultivation, the state where it was cultivated, the area under cultivation, the production yield, the annual rainfall, and the usage of fertilizers and pesticides.

The dataset reveals significant insights into the dynamics of agricultural productivity across different regions and over time. It captures the interplay between environmental factors such as rainfall, agricultural practices such as fertilizer and pesticide usage, and crop yields. Additionally, it provides a comprehensive overview of the distribution of crops, their cultivation periods, and the agricultural landscape across Indian states.

Through thorough analysis and modelling techniques, this dataset aims to predict and understand the factors influencing crop yields, facilitating informed decision-making processes for agricultural stakeholders, policymakers, and researchers.

3.DATA PREPROCESSING & FEATURE SELECTION

DATA PREPROCESSING

- **Handling Missing Values**: The code checks for missing values using df.isnull().sum(). If there are any missing values, appropriate strategies such as imputation or removal might be applied.
- **Exploratory Data Analysis (EDA):** The code explores unique values in each column using set(df[i].tolist()) to understand the data better. Additionally, it checks for duplicate records using df.duplicated().sum() and describes the dataset using df.describe().
- **Visualization**: Various visualizations are created to understand the data distribution, trends, and relationships between different variables. Scatter plots, line plots, bar plots, and scatter plots are used to visualize different aspects of the data.
- **One-Hot Encoding**: Categorical variables are converted into numerical using one-hot encoding (pd.get_dummies()).
- **Feature Scaling/Transformation**: Power transformation (Yeo-Johnson) is applied to handle skewed data distributions.
- **Data Splitting**: The dataset is split into training and testing sets using train_test_split().
- **Multicollinearity Handling**: Variance Inflation Factor (VIF) is calculated to detect multicollinearity among independent variables. Highly correlated features are dropped to improve model performance.

FEATURE SELECTION

- **Correlation Analysis**: This method calculates the relationship between each feature and the target variable (crop yield) to identify which features are most relevant. Features with higher correlation coefficients are more likely to be significant for making accurate predictions.
- **Feature Importance**: In tree-based models like Random Forest, feature importance can be used to rank features based on their contribution to reducing impurity or variance. Features that score higher in importance are deemed more influential for prediction.
- **Recursive Feature Elimination (RFE)**: RFE is an iterative process that starts with all available features and eliminates the least important ones step by step, stopping when the desired number of

features is reached. This method depends on the coefficients from models like linear regression or the feature importance values from models like Random Forest.

- **Lasso Regression (L1 Regularization)**: Lasso regression introduces a penalty term to the regression model that encourages some feature coefficients to shrink to zero. The features that remain with non-zero coefficients are selected, effectively performing feature selection.
- **Principal Component Analysis (PCA)**: PCA is a technique used for reducing the dimensionality of the data by transforming the original features into a new set of orthogonal variables, known as principal components. This helps to simplify the dataset while retaining essential information.
- **Forward/Backward Selection**: These are stepwise selection techniques. In forward selection, the process begins with an empty set of features and iteratively adds the most significant ones. In backward selection, the method starts with all features and removes the least important ones based on a specific criterion, such as AIC or BIC.
- **Mutual Information**: Mutual information assesses the dependency between features and the target variable. Features that have higher mutual information scores with the target are considered more informative and useful for the model.

4.METHODOLOGY

Import Libraries: Load essential libraries like Pandas, NumPy, Matplotlib and Seaborn, for data processing, visualization, and modeling.

Load & Preprocess Data: Clean the dataset by handling missing values, normalizing, scaling, and encoding categorical variables.

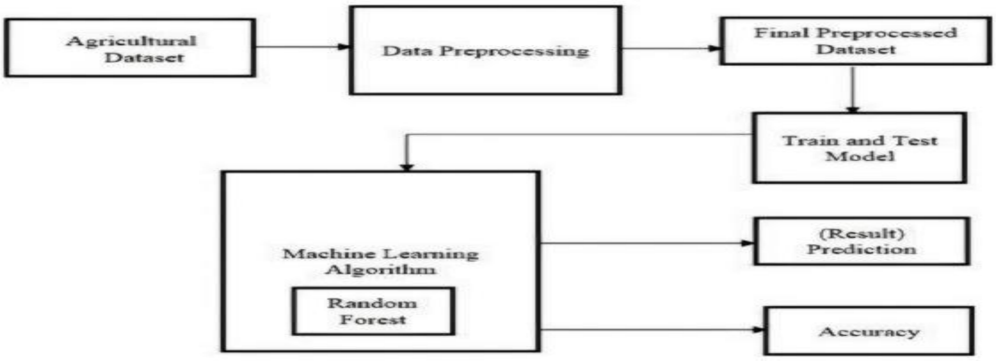
Visualization: Use histograms, box plots, and heatmaps to explore data distribution, variable relationships, and trends.

State-wise Production Analysis: Analyze agricultural output across Indian states to identify top-producing regions and influencing factors.

Crop-Specific Analysis (Wheat): Focus on wheat yield patterns, growth conditions, and the impact of factors like weather, soil, and location.

Data Transformation: Encode categorical variables and scale numerical features for compatibility with machine learning models.

Modeling: Train a machine learning model (e.g., Random Forest) to predict crop yield, splitting data into training/testing sets and evaluating performance using metrics like accuracy and precision.



5.CLASSIFICATION ALGORITHM

RANDOM FOREST REGRESSOR

Ensemble Learning:

• Random Forest Regressor is an ensemble learning method, meaning it combines the predictions of multiple individual models (decision trees) to improve overall performance and robustness.

Random Forest Algorithm:

• Random Forest builds multiple decision trees by randomly selecting subsets of features and data points.

• Each decision tree is trained on a bootstrapped sample of the training data (sampling with replacement).

- During the training process, at each node of each tree, a random subset of features is considered for splitting.
- The predictions from all the individual trees are aggregated to produce the final prediction.

Prediction:

- For regression tasks, the final prediction from the Random Forest Regressor is typically the average (mean or median) prediction of all the individual trees.
- Alternatively, weighted averaging or other aggregation techniques can be used.

Benefits:

- Random Forest Regressor is robust to overfitting and noise in the data, thanks to the ensemble approach and randomization.
- It can handle both numerical and categorical features without requiring feature scaling or one-hot encoding.
- It's relatively easy to use and less prone to parameter tuning compared to other complex models.

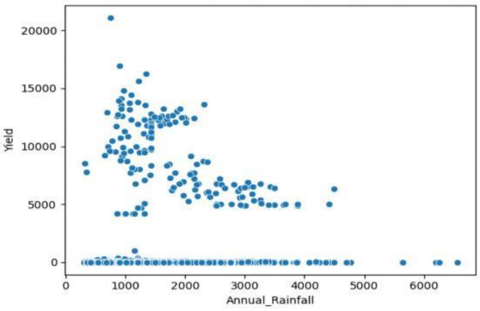
Applications:

- Random Forest Regressor is widely used in various regression tasks such as predicting housing prices, stock prices, sales forecasts, and more.

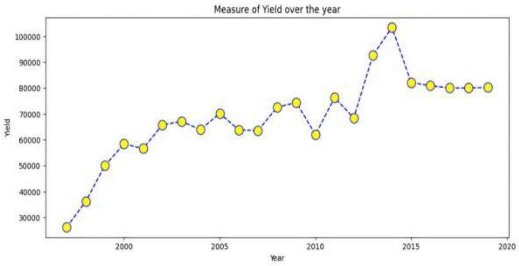
6.DATA VISUALIZATION

Data visualization is the graphical representation of information and data through visual element such as chart, graph and maps data visualization tools provide an accessible way to understand complex dataset.

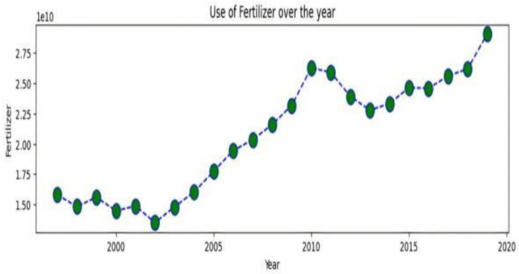
1.Visualization of Annual Rainfall



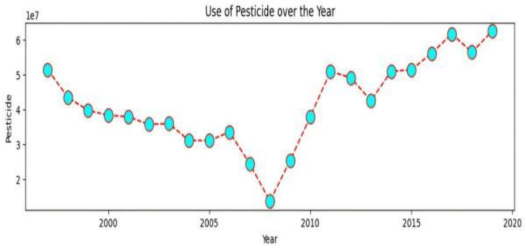
2.Visualization of yield over the year



3.Visualization of the use of fertilizer over the year



4.Visualisation of the pesticide over the year



7.FUTURE ENHANCEMENT

Hyperparameter Tuning could further improve the performance of Random Forest Regressor by tuning their hyperparameters using techniques like grid search or randomized search. Feature Engineering Experiment with creating new features or transforming existing ones to capture more meaningful information for the models. Ensemble Methods Consider ensemble methods like stacking or blending, where you combine predictions from multiple models to achieve better performance. Additional Metrics While accuracy is essential, consider including other evaluation metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) to get a more comprehensive understanding of model performance. Regularization implements regularization techniques like L1 or L2 regularization to prevent overfitting, especially for models like Linear Regression. Cross-Validation Implement k-fold cross-validation to get more reliable estimates of model performance and to detect overfitting.

8.CONCLUSION

Random forest regressor emerged as the best-performing model for predicting crop yield in the given dataset. It exhibited higher accuracy compared to other model both in training and testing phase, indicating its effectiveness in capturing the underlying patterns in the data making accurate prediction. The script provides a comprehensive analysis of agricultural crop yield prediction for Indian states using machine learning techniques. Through data visualization and modeling, it offers insights into factors influencing crop yield, such as annual rainfall, fertilizer usage, and area under cultivation. Various machine learning models are employed to predict crop yield, with evaluations indicating the effectiveness of each model. Recommendations for improving crop productivity and agricultural practices can be derived from the insights provided by this analysis. Random forest regression gives the accuracy for Agricultural crop yield prediction is 98%**.**

REFERENCES

[1]. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture Andrew Crane-Droesch Published on 26 October 2018

[2]. Crop Yield Prediction using Machine Learning and Deep Learning Techniques by PanelKavita Jhajharia [a], Pratistha Mathur [a], Sanchit Jain [a], Sukriti Nijhawan, Manipal University Jaipur Published on 31 January 2023.

[3]. Agricultural Analysis and Crop Yield Prediction of Habiganj using Multispectral Bands of Satellite Imagery with Machine Learning by Fariha Shahrin; Labiba Zahin
December 2020

[4]. Rice crop yield prediction in India using support vector machines by Niketa Gandhi, Leis J. Armstrong Published on 21 November 2016

[5]. Crop Yield Prediction using Machine Learning Algorithm by D.Jayanarayana Reddy,M. Rudra Kumar Published on26 May 2021

[6].R. Medar, V. S. Rajpurohit and S. Shweta, "Crop Yield Prediction using Machine Learning Techniques," *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, Bombay, India, 2019, pp. 1-5, doi: 10.1109/I2CT45611.2019.9033611.

[7]. Sonal Agarwal and Sandhya Tarar 2021 *J. Phys.: Conf. Ser.* 1714 012012

[8].Elbasi, E.; Zaki, C.; Topcu, A.E.; Abdelbaki, W.; Zreikat, A.I.; Cina, E.; Shdefat, A.; Saker, L. Crop Prediction Model Using Machine Learning Algorithms. *Appl. Sci.* 2023, *13*, 9288. https://doi.org/10.3390/app13169288

[9].Abbas, F.; Afzaal, H.; Farooque, A.A.; Tang, S. Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms. *Agronomy* 2020, *10*, 1046. https://doi.org/10.3390/agronomy10071046

[10].Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha and V. G. R. R. Devi, "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector," 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 736-741, doi: 10.1109/ICCES48766.2020.9137868