COMPARISION OF MACHINE LEARNING ALGORITHM FOR CROP YIELD PREDICTION

Dr. Rakesh Kumar B.,
Associate Professor,
St. Aloysius Deemed To Be University,
Beeri, Kotekar, Mangalore

Mrs. Shashikala Shetty
Associate Professor
SDM College of Business Management, Mangalore

ABSTRACT

Crop yield prediction has become increasingly vital in addressing challenges in agriculture and ensuring global food security, especially with the growing population. Traditionally, farmers relied on their experience and observations to decide which crops to plant, often based on previous harvests, neighboring crops, or regional preferences. However, this approach overlooks critical factors, such as essential soil nutrients like nitrogen, phosphorus, and potassium, which are crucial for healthy crop growth. This lack of awareness can lead to poor yields and inefficient farming practices.

Machine learning offers a modern solution by enabling farmers to make data-driven decisions about which crops to plant during the growing season. This paper explores the use of various machine learning algorithms in crop prediction, aiming to enhance agricultural productivity and support sustainable farming practices.

Keywords: Machine Learning Algorithm, Agriculture, Crop Yield Prediction, Random Forest, Linear Regression, Decision Tree, K-Nearest Neighbors Regressor (KNN).

I. INTRODUCTION

Agriculture is the foundation of human survival, with crop yield playing a vital role in ensuring food security, especially as the world's population continues to rise. In the past, farmers typically relied on their experience and local knowledge to predict crop yields, often basing their decisions on trends from previous seasons, neighboring areas, or commonly grown crops. However, this traditional approach overlooks critical factors like soil nutrients such as nitrogen, phosphorus, and potassium which are crucial for healthy crop growth. As a result, these methods can lead to lower crop yields and inefficient farming practices.

Recently, machine learning (ML) has emerged as a valuable tool in overcoming these challenges. By harnessing ML algorithms, farmers can make more informed decisions about which crops to plant, taking into account not only regional trends but also important factors such as soil quality, climate conditions, and historical data. This study delves into the use of various machine learning algorithms, including Linear Regression (Lr), K-Nearest Neighbors (Knn), Decision Tree (Dt), and Random Forest (Rf), to predict crop yield. These algorithms are evaluated for their accuracy and effectiveness in improving agricultural productivity, helping to foster more sustainable farming practices.

II. LITERATURE REVIEW

The paper "A Comparative Study of Agricultural Crop Yield Prediction Using Machine Learning Techniques" by **M. Uma Maheswari and Ramani Ramasamy (2023)** explores various ML algorithms, including Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbours

(KNN), Extreme Gradient Boosting (XGBoost), and Random Forest (RF), for crop yield prediction.

Lontsi Saadio Cedric et al. (2022) explored machine learning models like Decision Trees, Random Forests, and SVM to predict crop yields in West Africa, addressing challenges like climate change and resource limitations. Their study found Random Forest to be highly effective, highlighting the potential of ML in improving agricultural planning and food security.

The paper "Crop Selection and its Yield Prediction" by **Aksheya Suresh et al**. uses machine learning to help farmers select suitable crops and predict yields based on location and season. It employs Decision Tree for crop classification and Linear Regression for yield prediction, achieving an accuracy of 88.7%.

Ruby Aworka et al. (2022): This study developed a crop prediction framework using Support Vector Machine (SVM), Gradient Boosting Machine, and Random Forest (RF) algorithms, focusing on factors such as pesticide use, rainfall, temperature, and climate parameters. The RF algorithm achieved a high accuracy of $R^2 = 92.2\%$, outperforming the Gradient Boosting models.

Guojie Ruan et al. (2022): This research utilized daily meteorological data to extract weather features critical for crop yield prediction. It highlighted the effectiveness of ensemble classifiers like XGBoost and RF, which outperformed other models by leveraging seasonal changes and land nutrient values.

Mohamed Bouni et al. (2021): This work compared Random Tree, Naive Bayes (NB), K-Nearest Neighbors (KNN), and Deep Reinforcement Learning (DRL) for crop prediction based on soil and climate parameters. The study found that DRL and Random Tree provided similar accuracy, with DRL's performance improving with larger datasets.

Mummaleti Keerthana et al. (2021): This study suggested that combining unsupervised and supervised learning techniques could enhance crop yield prediction accuracy. It emphasized the importance of parameters such as country, crop name, year, yield value, and environmental factors, with AdaBoost Regressor achieving an accuracy of 95.7%.

Kiran Moraye et al. (2021): This research utilized district-level agricultural production data and weather datasets to predict crop yields. The RF algorithm was noted for its ability to generate accurate predictions through a voting mechanism among decision trees, achieving an accuracy of 87%.

Potnuru Sai Nishant et al. (2020): This study proposed a web application for forecasting crop productivity across India, employing enhanced regression methods and stacking regression concepts to improve yield predictions.

Maya Gopal P. S. et al. (2020): This research identified critical factors for accurate crop yield prediction and proposed various data mining methods, including KNN, SVM, Artificial Neural Networks (ANN), RF, and regression techniques.

III PROPOSED APPROACH

A. Dataset

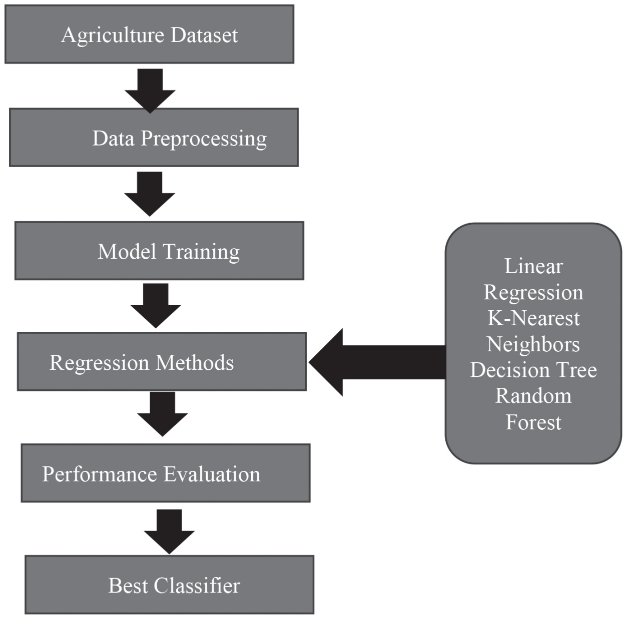
The dataset used in this experiment came from Kaggle's publicly available dataset, a community of data scientists who provide datasets for various analyses and model building. Area, Item, Year, hg/ha_yield,average_rain_fall_mm_per_year,pesticides_tonnes,avg_temp. The data were then pre-processed and contained 28243 records. This was accomplished using Microsoft Excel. CSV was the file format (Comma Separated values). (p. 1)

B. Data Preprocessing

Data pre-processing is the process of converting raw data into a smooth set of data. The data is collected from various sources and in multiple formats; the technique is effective but hard to implement. The data in the dataset has been preprocessed to remove redundant, missing, and inconsistent values. The final process of data preprocessing is the separation of training and testing data [10]. Because learning the concept requires as many records as possible, the records are typically divided unequally. The trained model is the initial dataset used to train ML techniques to learn and produce accurate predictions (in this specific instance, 80 percent of the total set of data is used for training).

C. System Architecture

This system is to predict crop yield by applying different ML approaches.



IV. MATERIALS AND METHODOLOGY

Dataset Acquisition

The dataset used in this study was obtained from Kaggle, a popular platform that provides datasets for various data science and machine learning tasks. The dataset contains important information for crop yield prediction, including:

- **Area**: The location where the data was collected.
- **Item**: The type of crop being analyzed.
- **Year**: The year the data was recorded.
- **hg/ha_yield**: Crop yield, measured in hectograms per hectare.
- **Average_rain_fall_mm_per_year**: The average annual rainfall in millimeters.
- **Pesticides_tonnes**: The amount of pesticides used, measured in tonnes.
- **Avg_temp**: The average yearly temperature.

Data Preparation

To ensure that the predictions are accurate, the dataset underwent several preprocessing steps:

- **Handling Missing Data**: Any missing values were addressed using imputation techniques to ensure completeness.
- **Scaling Features**: Continuous variables were normalized to make sure all features are on a similar scale.
- **Encoding Categorical Variables**: Non-numeric data were transformed into numerical values using techniques like one-hot encoding.
- **Feature Selection**: Important features were selected using correlation analysis to help improve the accuracy of the models and reduce unnecessary complexity.

Training and Evaluation

The machine learning models were trained on a subset of the data and tested on another subset to evaluate their performance. The models were assessed based on two main metrics:

- **Mean Squared Error (MSE)**: This measures the average squared differences between predicted and actual values.
- **R² Score (Coefficient of Determination)**: This indicates how well the models explain the variation in crop yields.

Machine Learning Models

The research tested four different machine learning models, each bringing a unique approach to crop yield prediction:

- **Decision Tree Regressor (DT)**: This model splits the data into hierarchical structures based on specific features to make predictions.
- **Random Forest Regressor (RF)**: This is an ensemble model that combines multiple decision trees to improve prediction accuracy.
- **Linear Regression (LR)**: A simple model that establishes a linear relationship between input features and the target variable.
- **K-Nearest Neighbors Regressor (KNN)**: This

model makes predictions based on the average of the closest data points in the feature space.

Comparative Assessment

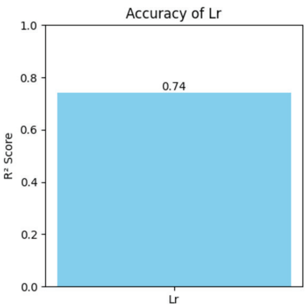
Once the models were trained and tested, their performance was compared using the key metrics to determine which algorithm was the most effective for predicting crop yields. This comparison helped highlight the strengths and weaknesses of each model.

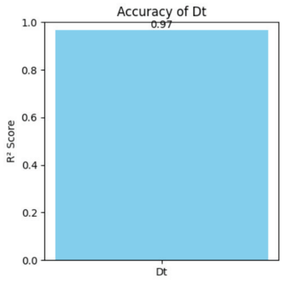
Tools and Libraries

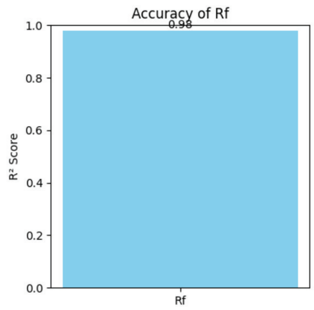
Python was used for the entire implementation, with key libraries including Scikit-learn for machine learning, Pandas for data manipulation, and NumPy for numerical operations.

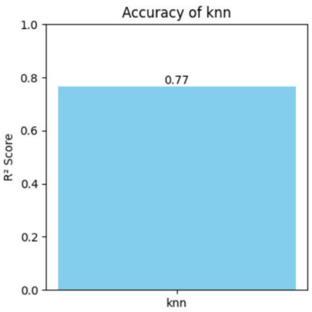
Result Analysis

The results were analyzed to identify the best-performing model. The analysis also explored the factors that influenced each model's performance, such as the trade-offs between prediction accuracy and computational efficiency.







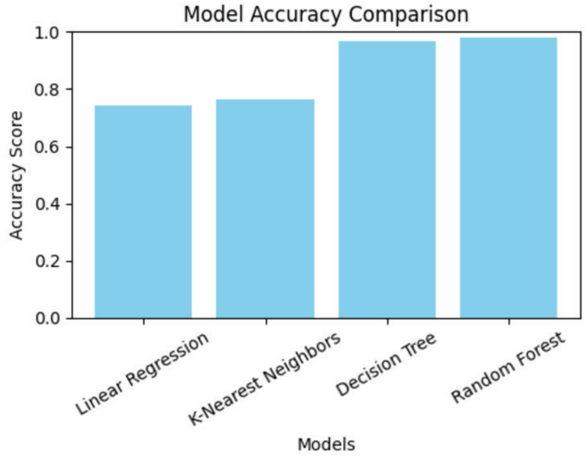


V RESULT AND DISCUSSION

Performance Evaluation Criteria

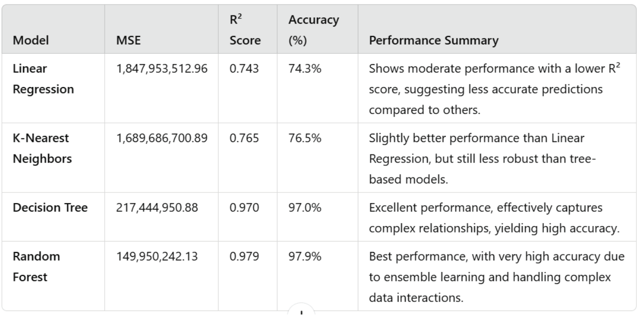
The main aim of crop yield prediction is to analyze the regression algorithm's performance based on various input parameters such as climate, soil conditions, and farming practices. These algorithms, including Linear Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF), are used to predict continuous outcomes like crop yield. Instead of using a confusion matrix (common in classification problems), regression models evaluate the model's accuracy using performance metrics like Mean Squared Error (MSE) and R² score. MSE measures the average squared difference between predicted and actual values, while the R² score indicates how well the model explains the variability in the data. These evaluation metrics help assess how well the regression algorithms can predict crop yield based on the provided data.

Accuracy (R² Score): Represents how well the model's predictions align with actual outcomes. It ranges from 0 to 1, with higher values indicating better prediction accuracy.



Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values. Lower MSE values indicate better performance.

Performance Summary: Based on the model's prediction capability (as per MSE and R² Score), the performance summary would highlight how effectively the model can predict crop yields. Here's how it could look:



Model	MSE	R^2 Score	Accuracy (%)	Performance Summary
Linear Regression	1,847,953,512.96	0.743	74.3%	Shows moderate performance with a lower R^2 score, suggesting less accurate predictions compared to others.
K-Nearest Neighbors	1,689,686,700.89	0.765	76.5%	Slightly better performance than Linear Regression, but still less robust than tree-based models.
Decision Tree	217,444,950.88	0.970	97.0%	Excellent performance, effectively captures complex relationships, yielding high accuracy.
Random Forest	149,950,242.13	0.979	97.9%	Best performance, with very high accuracy due to ensemble learning and handling complex data interactions.

VI. CONCLUSION

In this research, machine learning algorithms were applied to predict crop yields based on various agricultural parameters such as average rainfall, temperature, pesticide usage, and the type of crop grown. The study employed four regression models—Linear Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), and Random Forest (RF)—to assess the performance of each in accurately predicting crop yields.

Among the models tested, Random Forest (RF) demonstrated the highest performance with an accuracy score of 97.9%, followed by Decision Tree (DT) with 96.9%. K-Nearest Neighbors (KNN) and Linear Regression (LR) achieved lower accuracy scores of 76.5% and 74.3%, respectively. The results show that ensemble methods like Random Forest, which aggregate predictions from multiple trees, are better suited for this task due to their ability to capture complex, non-linear relationships in the data.

This research highlights the potential of machine learning in agricultural decision-making, offering a data-driven approach to predicting crop yields, which can ultimately assist farmers in optimizing crop production and improving food security. Furthermore, the performance evaluation criteria demonstrate the importance of choosing the right regression model based on the accuracy of predictions, with Random Forest proving to be the most effective algorithm for crop yield prediction in this study.

Future work could explore other machine learning techniques, such as Gradient Boosting or Neural Networks, and integrate more factors such as soil quality, irrigation patterns, and market conditions to improve prediction accuracy further.

VII. REFERENCES

[1]. M. Uma Maheswari and Ramani Ramasamy, *A Comparative Study of Agricultural Crop Yield Prediction Using Machine Learning Techniques*, 2023.

DOI: 10.1109/ICACCS57279.2023.10112854

[2]. Lontsi Saadio Cedric et al., *Crops Yield Prediction Based on Machine Learning Models: Case of West African Countries*, Elsevier, April 2022.

doi:https://doi.org/10.1016/j.atech.2022.100049[3]. Mohamed Bouni et.al, (2022, May). Towards an efficient Recommender System in Smart agriculture: A deep reinforcement learning approach. Volume 203 Elsevier, https://doi.org/10.1016/j.procs.2022.07.124 Procedia Computer Science 203 (2022) 825–830

[4]. Mummaleti Keerthana; K J M Meghana et.al (2021, March), An Ensemble Algorithm for Crop Yield Prediction, IEEE, DOI: 10.1109/ICICV50876.2021.9388479

[5]. Kiran Moraye et.al ,(2021, JUNE).Crop Yield Prediction Using Random Forest Algorithm for Major Cities in Maharashtra State, 'International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-9, Issue-2, March 2021, https://doi.org/10.21276/ijircst.2021.9.2.7

[6]. Potnuru Sai Nishant,Pinapa Sai Venkat et al,(2021)' Crop yield prediction based on Indian Agriculture using machine learning', IEEE, doi:10.1109/INCET49848.2020.9154036

[7]. V. Geetha; A. Punitha; et.al(2020) 'An Effective Crop Prediction Using Random Forest Algorithm, IEEE, doi:10.1109/ICSCAN49426.2020.9262311.

[8]. Anna Chlingaryan, Salah Sukkarieh, et al (2018), Machine learning approaches for Crop Yield prediction and nitrogen status estimation in precision agriculture: A review, Elsevier, https://doi.org/10.1016/j.compag.2018.05.012

[9]. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: by panel Anna Chlingaryan ,Salah Sukkarieh , Brett Whelan
https://doi.org/10.1016/j.compag.2018.05.012

❖❖❖❖❖❖❖