Research Article

Annals Of Agricultural Science And Technology

Exploring Machine Learning Techniques For Accurate Crop Yield Prediction.

S.Thavareesan*[1], J.Sriranganesan[2] And T. Nishatharan[1].

[1]*Department of Computing, Faculty of Science, Eastern University, Sri Lanka.*
[2]*Department of Mathematics, Faculty of Science, Eastern University, Sri Lanka.*

Corresponding author
Dr. S. Thavareesan ,
Department of Computing, Faculty of Science, Eastern University, Sri Lanka.
Email : sajeethas@esn.ac.lk

Received Date : December 20, 2024
Accepted Date : December 21, 2024
Published Date : February 05, 2025

Copyright *© 2024 Dr. S. Thavareesan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

ABSTRACT

The agricultural sector plays a critical role in South Asia's economy, providing livelihoods for millions and ensuring food security. However, challenges such as unpredictable weather patterns, limited arable land, and increasing population pressure significantly affect crop yields. Accurate crop yield prediction is essential for addressing these issues, as it enables informed decision- making on resource allocation, crop planning, and risk management. This study evaluates the performance of various machine learning regression models for predicting crop yields in five South Asian countries: Sri Lanka, Bangladesh, India, Nepal, and Pakistan. The dataset used includes crop yield data for the ten most widely consumed crops, alongside weather-related factors like rainfall, temperature, and pesticide usage, spanning from 1961 to 2016. The models assessed include XGBoost Regressor, Decision Tree Regressor, Gradient Boosting Regressor, Random Forest Regressor, K-Nearest Neighbors (KNN), and linear models such as Linear Regression, Ridge, Lasso, Elastic Net, and Support Vector Regression (SVR). Performance was measured using Mean Squared Error (MSE) and R^2 scores. The results demonstrate that XGBoost Regressor achieved the lowest MSE and highest R^2 score, making it the most accurate model for crop yield prediction. Decision Tree Regressor and Gradient Boosting Regressor also performed well, while SVR and simpler linear models (Linear Regression and Ridge Regression) showed poorer results. These findings emphasize the effectiveness of advanced machine learning techniques, especially XGBoost, in enhancing crop yield predictions and supporting more efficient agricultural decision- making in South Asia.

Keywords: Crop yield prediction, machine learning, South Asia, R^2 score, Mean Squared Error (MSE).

INTRODUCTION

The agricultural sector in South Asia plays a pivotal role in the region's economy, providing livelihoods for millions and contributing significantly to food security. However, South Asian countries face several challenges, including unpredictable weather patterns, limited arable land, and increasing population pressure, all of which impact crop yields. Accurate crop yield prediction is critical for addressing these challenges, as it helps farmers, policymakers, and agricultural experts make informed decisions about resource allocation, crop planning, and risk management.

Traditionally, crop yield prediction relied on empirical knowledge, statistical models, and basic agricultural practices. However, with the availability of large-scale data from satellite imagery, meteorological observations, and soil sensors, machine learning (ML) techniques have become a powerful tool in enhancing the accuracy of crop yield forecasts. These techniques can process complex datasets to detect hidden patterns, make predictions, and provide valuable insights that traditional methods may overlook.

This paper presents an analysis of various machine learning techniques applied to crop yield prediction in South Asian countries, where agriculture is heavily dependent on weather, soil quality, and other environmental factors. We aim to explore how machine learning models—such as Gradient Boosting Regressor, Random Forest Regressor, Support Vector Regression (SVR), Decision Tree Regressor, K-Nearest Neighbors (KNN), Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, XGBoost Regressor - can be used to predict crop yields with high accuracy in the diverse agricultural contexts of South Asia. The paper also evaluates the strengths, weaknesses, and practical applications of these techniques, considering the specific challenges faced by farmers in this region.

In the following sections, we review the application of machine

learning models in crop yield prediction, focusing on studies conducted in countries like India, Bangladesh, Pakistan, and Sri Lanka. We discuss the impact of environmental variables, such as rainfall, temperature, and soil conditions, as well as the role of remote sensing and time-series data in improving prediction accuracy. By analyzing these machine learning techniques and their implementations in South Asia, this paper aims to provide insights into the future of crop yield prediction and the potential benefits of adopting machine learning in agricultural practices across the region.

RELATED WORK

In this paper[1] several machine learning algorithms like KNN, Lasso regression, Ridge regression, linear regression and Decision tree, were used to forecast the crop yield. Among all the models developed , the KNN algorithm emerged as the most effective and highest accuracy for crop yield forecasting. Also they highlighted the enormous potential of machine learning in providing accurate and reliable predictions for forecasting crop yield using historical data and climate variables. These results would be helpful for crop management and decision making in agriculture.

According to a study[2] emphasizes how machine learning has the potential to revolutionize crop management techniques. By leveraging historical data and advanced predictive algorithm, machine learning can provide precise insights into crop yields, optimal resource allocation and efficient farming practices. Crop prediction is done by classification model and yield prediction uses regression models to learn from the data. Multiple machine learning models are analyzed based on performance metrics. backend. Among the used models Random Forest Regression gives best results for yield prediction. For crop prediction, Naïve Bayes classifier gives most accurate results with highest accuracy.

This research paper [3] provides valuable insights into the factors influencing crop yield and demonstrates the effectiveness of machine learning models in predicting and understanding agricultural outcomes. The original dataset was utilized to train various regression machine learning models, and their performance was compared using metrics such as the R-squared score and Root Mean Squared Error. The Extra Trees Regressor model achieved the highest R-squared score indicating its good prediction accuracy.

Research paper [4], proposed an Machine learning based model, Smart Crop Selection (SCS), which is based on data of metrological and soil factors. These factors include nitrogen, phosphorus, potassium, CO2, pH, temperature, humidity of soil, and rainfall. Existing IoT-based systems are not efficient as compared to their proposed model due to limited consideration of these factors. In the proposed model, real-time sensory data is sent to Firebase cloud for analysis. Its results are also visualized on the Android app. SCS ensembles the following five Machine Learning algorithms like Decision tree, SVM, KNN, Random Forest, and Gaussian Naïve Bayes to increase performance and accuracy. For rainfall prediction, a dataset containing historical data of the last fifteen years is acquired from Bahawalpur Agricultural Department.

Authors [5] said that the Data Mining techniques are the better selections for predicting yield of crop. Different Data Mining techniques are used in agriculture for estimating the upcoming year's crop production. Crop Yield Prediction includes predicting yield of the crop from previous historical data like rainfall, temperature and groundwater level. KNN model is using to classifies the groundwater level dataset to predict the future test data record dataset. It could be useful in analyzing the ground water in the past and which predict the future level.

A study [6] used several models like random forest, decision tree classifier, support vector machine, KNN, and logic regression to find the best predictive model. So that they have been suggested most suitable crops to grow based on the available climatic conditions and soil conditions. With the highest accuracy score, the random forest produced the greatest results out of all of them.

Research paper [7] find the best model for crop prediction, which can help farmers decide the type of crop to grow based on the climatic conditions and nutrients present in the soil. This paper compares popular algorithms such as K-Nearest Neighbor (KNN), Decision Tree, and Random Forest Classifier using two different criterions Gini and Entropy. Results expose that Random Forest gives the highest accuracy among the three algorithms.

This paper[8] proposes a feasible and user-friendly yield prediction system for the farmers. The proposed system provides connectivity to farmers via a mobile application. GPS helps to identify the user location. Machine learning algorithms allow choosing the most profitable crop list or predicting the crop yield for a user-selected crop. Selected Machine Learning algorithms such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF), Multivariate Linear Regression (MLR), and K-Nearest Neighbour (KNN) are used to predict the crop yield. The various algorithms are compared with their accuracy. The results obtained indicate that Random Forest Regression is the best among the set of standard algorithms used on the given datasets with high accuracy.

In this research [9] developed a model using three machine learning algorithm such as KNN, support vector machine and naïve Bayes for the purpose of predicting crop yields. Crop yield data set is used for experimental work. Accuracy, sensitivity and specificity are used to compare the performance. The

experimental data set contains data pertaining to the crop as well as other information The training of machines to learn and create models for future predictions is widely applied in all fields in current world. Agriculture is a cornerstone of the global economy, and with the ongoing growth of the human population, understanding global crop yields is essential for tackling food security challenges and mitigating the effects of climate change. Crop yield prediction is a challenging and important agricultural problem all over the world. The Agricultural yield primarily depends on weather conditions (rain, temperature, etc), pesticides and accurate information about history of crop yield in the past. Crop yield prediction is important when making decisions related to agricultural risk management and future predictions.

The ultimate goal of this research is to apply various machine learning algorithms to predict crop yields in South Asian countries based on the provided data, including weather variables (temperature and rainfall), pesticides used, and historical yield data.

The paper presents a comparison of various machine learning algorithms, such as Gradient Boosting Regressor, Random Forest Regressor, Support Vector Regression (SVR), Decision Tree Regressor, K-Nearest Neighbors (KNN), Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, XGBoost Regressor, in crop yield prediction in South Asian countries.

METHODOLOGY

The paper compares various machine learning algorithms for predicting crop yields in South Asian countries. Our study focused on five countries in the region: Sri Lanka, Bangladesh, India, Nepal, and Pakistan. We collected data from publicly available sources, including the Food and Agriculture Organization (FAO) and the World Bank.

We gathered crop yield data for the ten most widely consumed crops in these countries, including Manioc, Maize, Plantains, Potatoes, Rice, Paddy, Sorghum, Soybeans, Sweet Potatoes, Wheat, and Yams. The dataset provides information on country, crop name, year, and yield per year, covering the period from 1961 to 2016.Recognizing the impact of weather on agriculture, we also collected data on annual rainfall and average temperature for each country from the World Bank. The rainfall data covers the period from 1985 to 2016, while the temperature data is available from 1849 to 2013. Additionally, pesticide usage data for each crop and country was sourced from the FAO.

All this data was merged to create a dataset with eight attributes: crop name, country name, year, yield per year, average rainfall per year, pesticide usage, and average temperature. The final dataset contains 6090 instances of crop yield data, ranging from 1990 to 2013.

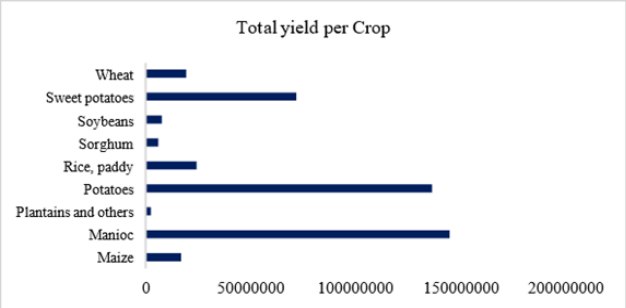
India has the highest crop yield production in the dataset, while Nepal has the lowest. **Table 1**, shows the total yields per country.

Table 1: County vs Total Yield.

Country	Total Yield
India	327420324
Pakistan	73897434
Bangladesh	15440318
Sri Lanka	11217741
Nepal	4113713

Table 2 displays the total yield for each crop across all the South Asian countries we analyzed. Maize was the most produced crop in South Asian countries, followed by potatoes. Plantains and other crops had the lowest production.

Table 2: Total Yield per Crop.



India stands out as the leading producer, contributing heavily to the yields of Manioc, Potatoes, Rice, and Sweet potatoes. Maize is primarily produced in India and Pakistan, but it lags behind crops like Manioc and Potatoes in overall yield. Sri Lanka contributes relatively lower yields across all crops, except for Manioc and Plantains, where it plays a notable role. Plantains and others is the least produced crop in the region, mainly grown in Sri Lanka. **Table 3** shows the clear picture of the distribution of crop yields across South Asia, with India as the dominant producer in most categories.

Table 3. Crop vs Main Country.

Crop	**Main Producers**
Manioc	India (Mostly)
Potatoes	India (Mostly)
Sweet Potatoes	India, Pakistan
Rice ,Paddy	India, Pakistan
Maize	India, Pakistan
Soybeans	India, Pakistan
Sorghum	India, Pakistan
Wheat	India, Pakistan
Plantains and Others	Sri Lanka

Annals Of Agricultural Science And Technology

Data Preprocessing

Data preprocessing is a method used to transform raw data into a clean, structured dataset. Essentially, when data is collected from various sources, it is often in an unrefined state, which makes it unsuitable for analysis. The dataset contains two categorical columns, which are variables that hold label values instead of numeric ones. Categorical data typically has a limited set of possible values, such as the items and countries in this case. Since many machine learning algorithms cannot process label data directly and require numeric input for both variables and outputs, the categorical data must be transformed into a numerical format.

One common technique for this transformation is One-Hot Encoding. This method converts categorical variables into a format suitable for machine learning models, helping improve prediction accuracy. One-Hot Encoding creates a binary column for each category, representing the presence or absence of a category in the dataset, and returns a matrix of these binary values.

The dataset above includes features with varying magnitudes, units, and ranges. Features with larger magnitudes will dominate the distance calculations, potentially leading to biased results in machine learning models. To address this issue, it's essential to scale the features so that they all have comparable magnitudes. Scaling ensures that no single feature disproportionately influences the model. This can be achieved by applying various scaling techniques, which standardize or normalize the features, bringing them to the same scale.

Training and testing Data

The dataset is typically divided into two subsets: the training dataset and the test dataset. The split is often uneven because training a model generally requires as much data as possible. Common splits are 70/30 or 80/20 for training and testing.

The training dataset is used initially to train the machine learning algorithm, allowing it to learn patterns and generate accurate predictions. In this case, 80% of the dataset is allocated for training. The test dataset, on the other hand, is used to evaluate how well the algorithm performs after being trained. It is crucial not to reuse the training dataset for testing because the algorithm would already "know" the expected output, which would invalidate the testing process. Typically, 20% of the dataset is reserved for testing.

Machine Learning algorithms

For the crop yield prediction, we applied a range of machine learning algorithms, each chosen to capture different aspects of the data and improve prediction accuracy. The models and their respective hyperparameters are as follows:

- GradientBoostingRegressor: We used this model with 200 estimators (trees), a maximum depth of 3 for each tree, and a fixed random seed (random_state=0) to ensure reproducibility. This model builds trees sequentially, where each tree corrects the errors of the previous one.
- RandomForestRegressor: This model also used 200 estimators and a maximum tree depth of 3, with a random seed for reproducibility. Random forests build multiple decision trees independently and then aggregate their results, which helps reduce overfitting.
- Support Vector Regressor (SVR): The SVR model was used with default settings. This model is based on the principle of finding a hyperplane that best fits the data in a high-dimensional space, making it particularly useful for non-linear relationships.
- DecisionTreeRegressor: A simple decision tree model was used without any additional tuning. It splits the dataset into subsets based on feature values, making decisions at each node to predict the target variable.
- KNeighborsRegressor: This model was set with 5 neighbors (n_neighbors=5). K-Nearest Neighbors (KNN) is a non-parametric algorithm that makes predictions based on the average of the closest data points in feature space.
- Linear Regression: A basic linear regression model was used to capture linear relationships between the features and the target variable.
- Ridge Regression: This model applied an L2 regularization technique with an alpha value of 1.0, which helps prevent overfitting by shrinking the coefficients of less important features.
- Lasso Regression: With an alpha value of 0.1, this model applied L1 regularization, which helps in feature selection by shrinking some coefficients to zero.
- ElasticNet Regression: This model used both L1 and L2 regularization, with an alpha of 0.1 and an L1 ratio of 0.5, balancing between Lasso and Ridge regularization techniques.
- XGBoost Regressor: This powerful boosting algorithm used 100 estimators and a learning rate of 0.1. XGBoost combines multiple weak learners (trees) sequentially, where each tree corrects the mistakes of the previous one, often delivering superior performance.

These diverse models were applied to the crop yield prediction task, with each model offering a different approach to handling the data. By using a mix of simple and advanced models, the aim was to identify the best-performing model for the specific dataset and prediction task.

Evaluation matrices

For evaluating the performance of the machine learning models used for crop yield prediction, we employed two key

metrics: Mean Squared Error (MSE) and R-squared (R^2) error.

Mean Squared Error (MSE)

MSE is a commonly used metric for regression tasks. It calculates the average of the squared differences between the predicted and actual values. The formula for MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2$$

Where:

- yi is the actual value of the target variable,
- ŷi is the predicted value,
- n is the total number of data points.

A lower MSE indicates better performance, as it suggests that the predicted values are closer to the actual values.

R-squared (R^2) Error

R^2 measures how well the model's predictions match the actual data. It represents the proportion of variance in the target variable that is explained by the model. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}i)^2}{\sum_{i=1}^{n}(yi - \bar{y})^2}$$

Where:

- yi is the actual value,
- ŷi is the predicted value,
- ȳ is the mean of the actual values.

R^2 values range from 0 to 1:

- An R^2 of 1 means that the model perfectly predicts the target variable.
- An R^2 of 0 means that the model does no better than predicting the mean of the target variable.

In summary, MSE is used to measure the accuracy of the model's predictions, with lower values indicating better performance. R^2 provides insight into the proportion of variance explained by the model, with higher values indicating better fit to the data. Together, these metrics help assess how well each machine learning model predicts crop yield.

RESULTS AND DISCUSSION

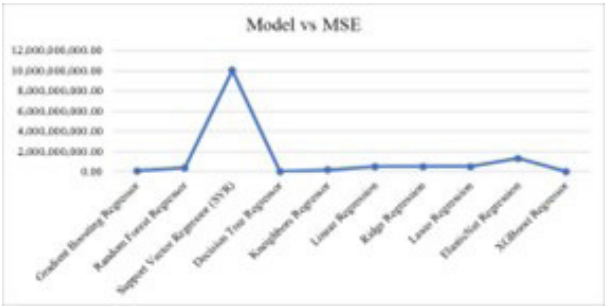
The performance of several machine learning regression models for predicting crop yield was evaluated using Mean Squared Error (MSE) as the evaluation metric. **Table 4** shown below summarizing the MSE values for each model:

Table 4. Model vs MSE.

Model	Mean Squared Error(MSE)
Gradient Boosting Regressor	81,062,296.60
Random Forest Regressor	401,848,760.27
Support Vector Regression (SVR)	10,029,525,063.91
Decision Tree Regressor	37,058,000.83
K-Nearest Neighbors (KNN)	180,411,614.63
Linear Regression	523,963,690.69
Ridge Regression	524,275,886.24
Lasso Regression	523,968,228.05
Elastic Net	1,324,932,659.62
XGBoostRegression	15,615,822.00

The data reveals (See Figure 1) significant variation in the performance of different regression models, as indicated by their Mean Squared Error (MSE). XGBoost Regressor stands out as the best-performing model, with the lowest MSE of 15,615,822.00, indicating that it produces the most accurate predictions among the models tested. Following closely behind, the Decision Tree Regressor demonstrates strong performance with an MSE of 37,058,000.83, outperforming Gradient Boosting Regressor, which has a higher MSE of 81,062,296.60. While Gradient Boosting is still effective, its performance lags slightly behind the Decision Tree in this case. K-Nearest Neighbors (KNN) and Random Forest Regressor show progressively higher MSE values, with KNN at 180,411,614.63 and Random Forest at 401,848,760.27, suggesting reduced prediction accuracy compared to tree-based and boosting models. The simpler linear models—Linear Regression, Ridge Regression, and Lasso Regression—all exhibit similar MSE values around 523 million, indicating they are less effective in capturing the data patterns. Elastic Net performs even worse with an MSE of 1,324,932,659.62, while Support Vector Regression (SVR) has the highest MSE by far, at 10,029,525,063.91, signifying its poor predictive performance in this scenario. Overall, the results suggest that boosting methods like XGBoost provide superior performance, while SVR and linear models should be avoided for this particular dataset due to their relatively poor performance.

Figure 1: Model vs MSE.

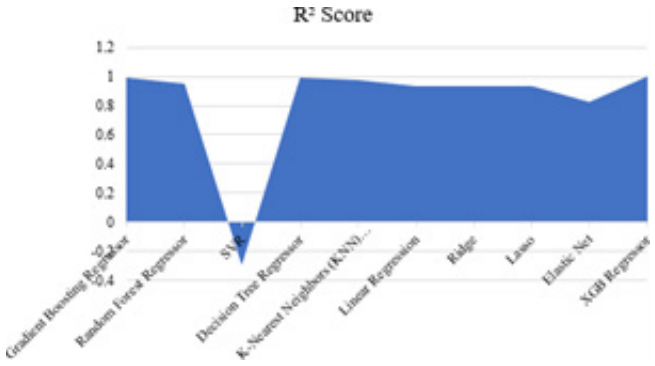


The **Table 5** presents the R^2 scores for various regression models, a key metric used to assess model performance. R^2 measures the proportion of variance in the target variable that is explained by the model. A higher R^2 score indicates a better fit of the model to the data, with 1 being a perfect fit and values less than 0 indicating poor performance.

Table 5. Model vs R2 Error.

Model	R^2 Score
Gradient Boosting Regressor	0.989514681
Random Forest Regressor	0.948021303
Support Vector Regression (SVR)	-0.297308084
Decision Tree Regressor	0.995110049
K-Nearest Neighbors (KNN) Regression	0.976663955
Linear Regression	0.93222587
Ridge Regression	0.932185488
Lasso Regression	0.932225283
Elastic Net	0.828621411
XGBoostRegression	0.997980118

Figure 2. Model vs R^2 Error.



The analysis of the regression models based on their R^2 scores reveals significant variations in their performance. XGBoost Regressor leads with an impressive R^2 score of 0.99798, indicating that it captures nearly 99.8% of the variance in the dataset, making it the most accurate model. Decision Tree Regressor follows closely with a score of 0.99511, also reflecting a strong fit, though slightly less accurate than XGBoost Regressor. Gradient Boosting Regressor also performs well, with an R^2 score of 0.98951, showing it can explain about 98.95% of the data's variance. Models like K-Neighbors Regressor (R^2 = 0.97666) and Random Forest Regressor (R^2 = 0.94802) offer good performance but fall behind the top tree-based models in terms of accuracy. The linear models, including Linear Regression, Lasso, and Ridge, each with an R^2 around 0.93, perform reasonably but are outperformed by more advanced non-linear models. Elastic Net performs the worst among the linear models, with a score of 0.82862, indicating that it struggles to capture the data's complexity. SVR stands out as the least effective, with a negative R^2 score of -0.2973, showing that it performs worse than simply predicting the mean value of the target variable. In conclusion, XGBoost Regressor is the best model for this dataset, providing the most accurate predictions, while SVR is not suitable for the task at hand.

CONCLUSION AND FUTURE WORK

In conclusion, when selecting a model for yield prediction, it is essential to consider both Mean Squared Error (MSE) and R^2 scores to determine the most accurate and reliable model. Based on the analysis of both MSE and R^2 values, XGBRegressor stands out as the top-performing model. With the lowest MSE of 15,615,822.00 and an R^2 score of 0.99798, it demonstrates excellent predictive accuracy and effectively captures the variance in the data. This makes XGBRegressor the best choice for predicting yield, as it provides both high accuracy and low prediction error.

DecisionTreeRegressor and GradientBoostingRegressor also perform well, with low MSE and high R^2 scores, indicating strong model fits. These models, while not as effective as XGBRegressor, still provide reliable predictions and are strong alternatives.

On the other hand, models such as SVR, which has a significantly higher MSE and a negative R^2 score, perform poorly in yield prediction. The high MSE of SVR (10,029,525,063.91) and its negative R^2 score (-0.2973) indicate that it is not suitable for this task, as it fails to predict yield effectively.

Linear models such as Linear Regression, Ridge, and Lasso have decent R^2 scores (around 0.93) but exhibit higher MSE compared to tree-based models, indicating that they may not capture the complexities of the data as effectively.

Overall, models like XGBRegressor and DecisionTreeRegressor should be prioritized for accurate yield prediction, as they not only achieve the best R^2 scores but also minimize prediction errors with low MSE. Linear models and SVR should be reconsidered due to their higher error rates and less reliable performance.

In our future work, we plan to focus on exploring ensemble methods as a way to further enhance yield prediction accuracy and robustness. Ensemble methods combine the strengths of multiple models, making the overall system more effective than any individual model alone. By leveraging various algorithms with complementary strengths, ensemble techniques can improve prediction performance and reduce the risk of overfitting or bias, ensuring that the final predictions are more reliable. By implementing these ensemble techniques, we can combine models like XGBRegressor, RandomForestRegressor, and DecisionTreeRegressor,

which each have unique advantages, resulting in improved accuracy, better generalization, and increased robustness in yield prediction.

REFERENCES

1. Muhammad Haziq Anwar et al, Crop Yield Prediction Using Machine Learning, Journal of Innovative Computing and Emerging Technologies, vol. 04, No.02, 2024.

2. Patil et al, Crop Selection and Yield Prediction using Machine Learning Approach, Current Agriculture Research Journal, Vol.11, No.03, 2023.

3. Nikhil, U.V. et al, Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models. Computers 2024, 13, 137.

4. Amna Ikram et al, Crop Yield Maximization Using an IoT-Based Smart Decision, Journal of Sensors, 2022.

5. Latha Jothi. V et al, Crop Yield Prediction using KNN Model, International Journal of Engineering Research & Technology, vol. 8.No. 12, 2020.

6. Sana Alam et al, Optimizing Agricultural Outcomes: Machine Learning in Crop Yield Prediction, The Asian Bulletin of Big Data Management, vol 04, Issue 04, 2024.

7. Madhuri Shripathi Rao et al, Crop prediction using machine learning, J. Phys.: Conf. Se J. Phys.: Conf. Se, 2022.

8. Shilpa Mangesh Pande et al, Crop Recommender System Using Machine Learning Approach, Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021) IEEE Xplore Part Number: CFP21K25-ART.

9. Meenakshi. G et al, Support Vector Machine for Crop Yield Prediction Towards Smart Agriculture, First International Conference on Artificial Intelligence for Internet of things (AI4IOT): Accelerating Innovation in Industry and Consumer Electronics, 2023.