Original Research Article

AI Powered Crop Rotation: Optimizing Plant Selection and Timing for Sustainable Farming

Tharun Mukesh

Manthan International School, Mandal, Tellapur, Ramachandrapuram, Hyderabad, Telangana 502032, India

ABSTRACT

As fears of worldwide water shortage rise, finding sustainable solutions for water usage in agricultural fields is of utmost importance. The agricultural sector is the cause of approximately 70% of freshwater use globally, which is why the adaptation of techniques such crop rotation would potentially be pivotal in ensuring sustainable utility of the water consumption by encouraging soil health, water conservation and enhancing long-term resilience of the farmland. However, deducing appropriate crop rotations and optimal planting times is a challenge for farmers because it depends on a myriad of factors such as soil composition, rainfall patterns, temperature fluctuations and the life cycles of pests. This study investigated how data-driven tools can be used to aid improved crop rotation planning through the use of a broad set of agricultural and environmental variables. Based on a dataset from the FAO and World Data Bank of more than 28,000 entries spanning multiple countries and years, this project examined how rainfall, temperature, pesticide application, and other variables affect crop yield. The approach involved data processing by removing redundant columns, one-hot encoding of the categorical variables, numerical feature scaling by standardization and missing value handling. This was followed by the development of predictive regression models aimed at estimating crop yield and suggesting suitable crops and planting schedules. Model performance was measured in terms of standard metrics including mean absolute error, mean squared error, and R^2. The tuned K-Nearest Neighbours model achieved the best performance with $R^2 = 0.99$, MAE = 3257 hg/ha, and MSE = 78.5 million, showcasing high accuracy when predicting crop yield. The aim of this study was to help farmers seamlessly integrate crop rotation into their practice through a tool that provides straightforward insights to aid in maintaining soil health, controlling pest cycles, and reducing water consumption.

Keywords: Crop rotation; Yield Protection; Agricultural optimization; Environmental Sustainability; Machine Learning

INTRODUCTION

Water scarcity represents one of the most critical challenges faced by humanity. In fact, four billion people already suffer water stress for at least one month annually according to the Organization for Economic Co-operation and Development (OECD) (1). Furthermore, issues such as population growth, climate change and rapid urbanization will exacerbate

Corresponding author: Tharun Mukesh, E-mail: tharunmukesh.tm@gmail.com.

Accepted October 6, 2025
https://doi.org/10.70251/HYJR2348.35717728

this problem and lead to nearly three billion people living under severe water stress heeds The Economist (2). Agriculture accounts for around 70% of global freshwater consumption (3), positioning it at the heart of this crisis. With the issue only continuing to worsen, sustainable solutions are needed to combat this problem.

A notable and effective way to remedy this crisis is through promoting the practice of crop rotation, which involves the diversification of crops to accommodate for seasonal changes in climatic conditions. Crop rotation is an established way through which farmers can elevate the soil health, pest resilience, and water sustainability of their farmlands (4). Traditionally, farmers have primarily relied on a combination of firsthand experience and prior knowledge to make decisions on crop rotation. While this manual process may benefit farmers in the short run, it is time-consuming and highly subjective as well. Moreover, it may prove inadequate when responding to the ever-changing environmental conditions or when applied at scale.

Recent studies, however, have investigated how artificial intelligence can be leveraged to overcome these problems. Fenz et al. (2023a) (5) used a deep Q-network (DQN) reinforcement learning model to produce adaptive rotation sequences. The model, though, was based on a simplified farmland model and hence could not fully translate into real world applications. In a follow-up study, Fenz et al. (2023b) (6) combined Normalized Difference Vegetation Index (NDVI) estimated through satellite images with meteorological and soil data to enhance model responsiveness. Even though there was improvement, the model could not differentiate weeds from crops due to limitations in NDVI. Additionally, Liang et al. (2023) (7) developed a decision-support tool based on multi-objective optimization to improve sustainability and agronomic feasibility. However, their approach struggled to take into account uncertainty over the long-run.

Unlike previous studies that have centered mostly on rotation strategies or sustainability, this study stands out because of its straightforward model of crop yield prediction with machine learning. Since the yield is now the measurable output and input features including temperature, rainfall, use of pesticides and year, the model can be trained to make more or less the best decision by learning how to adapt to the changes in the conditions, and provide farmers with insights that are direct and actionable. The methodology comprises data pre-processing, one-hot encoding for categorical information, normalization steps for numerical values and the identification of outliers through Cook's Distance. Several regression models, such as linear regression, random forest, support vector regression, and K-nearest neighbor bands, were trained and compared to identify the best model in terms of accuracy and interpretability. Ultimately, this study seeks to develop a data-driven tool that will be able to guide farmers make planting decisions that are agriculturally viable and environmentally sustainable.

METHODS AND MATERIALS

Dataset

The objective of this project was to train machine learning algorithms that could predict crop yield using environmental and agriculture variables. For this purpose, a dataset (8) was collected with 28,242 observations from a variety of countries, crops, and years. The dataset was preprocessed so that rows with missing or inconsistent data would be eliminated. All entries consisted of the following features: crop yield in hectograms per hectare (hg/ha), average annual precipitation (mm/year), pesticide use (tonnes), and average temperature (°C). The dataset also included identification variables such as country, crop, and year.

Visualisations

Exploratory data analysis was conducted to gain insights into the dataset structure, look at trends by crop and country, and evaluate the input feature and crop yield in relationships. Given the complexity of the 28,000 records and 101 countries, visual analysis was crucial in both model development and explanation.

Firstly, a histogram as shown in Figure 1 was created to display the distribution of the crop yields for the whole dataset. The data was right-skewed, with a majority of the crops having average yields, while fewer achieved extremely high outputs. This distribution indicated the existence of outliers and framed expectations for model performance, particularly where error metrics are sensitive to skewed data.

Furthermore, directly comparing the average yield of every crop type in the data set revealed in Figure 2. More common crops such as maize, rice, and wheat tended to have higher yields, whereas rarer crops such as millet or barley had lower averages. These distinctions are important for model training, as the model might be able to handle crops with greater data

or more standardized yield ranges better.

Another bar chart visualization Figure 3 revealed which nations had the highest mean crop yields. Nations like the United Kingdom and Belgium were near the top, presumably because of more sophisticated farm infrastructure and uniform growing conditions. This served to emphasize regional patterns in performance and exposed the unbalanced inclusion of nations in the dataset.

Additionally, three scatter plots were used to analyze the relationships between crop yield and the main numerical input features. The yield vs rainfall plot Figure 4 indicated a generally positive relationship, meaning that higher levels of rainfall correlate with

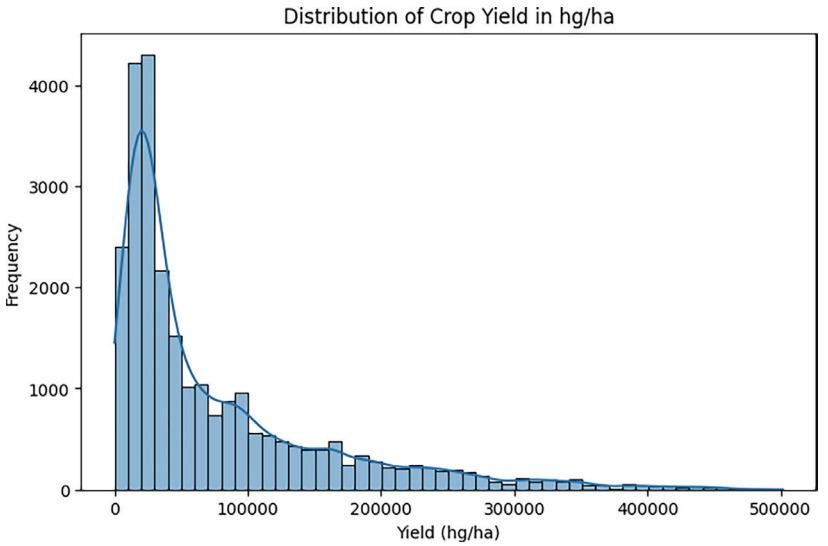

Figure 1. Right-skewed distribution of crop yield across 101 countries and 10 crops showing most countries and crops achieve modest yields.

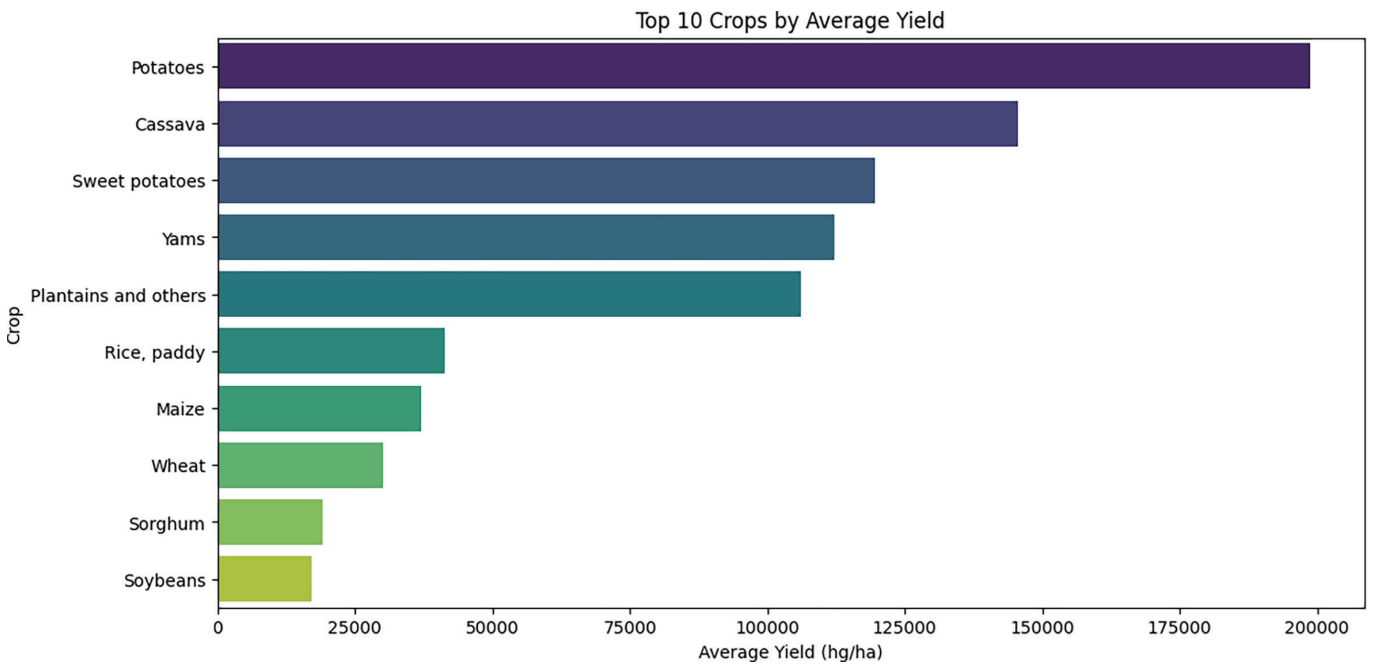

Figure 2. Top 10 crops by average yield showing variation in productivity across crop types.

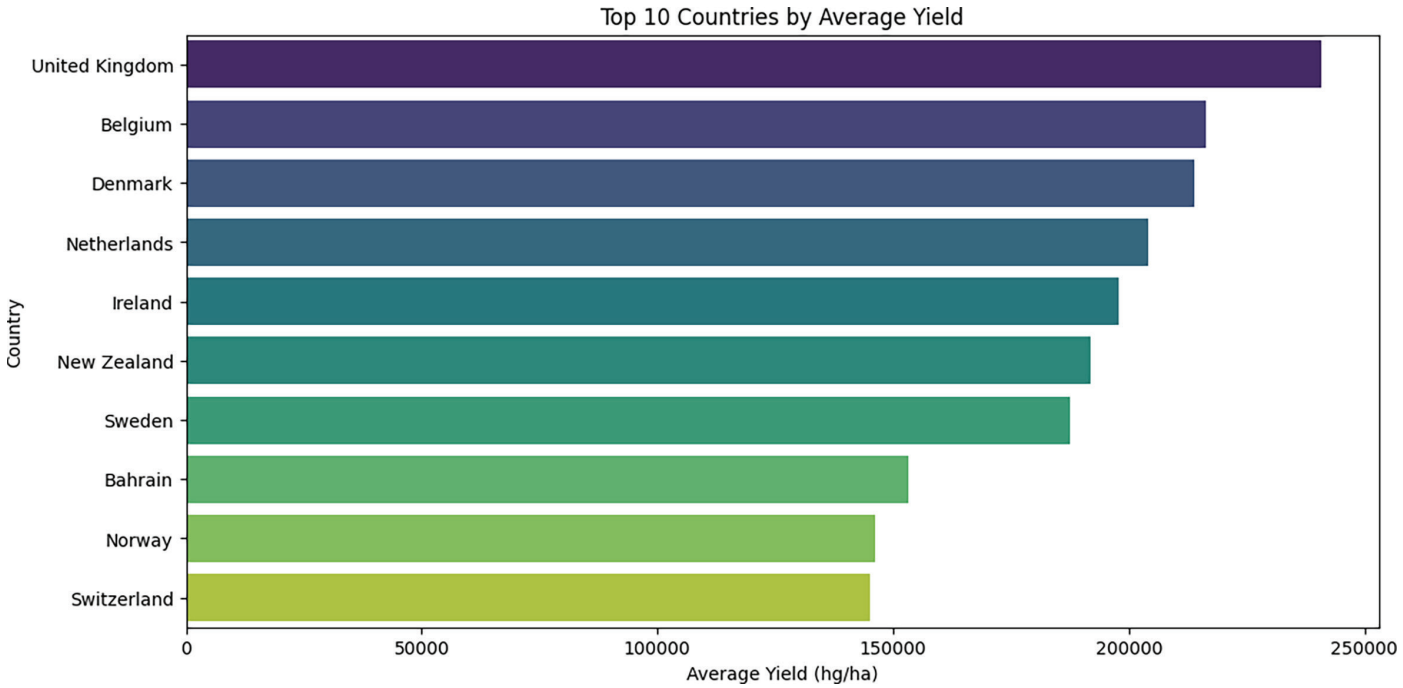

Figure 3. Top 10 countries by average crop yield demonstrating how regional factors and farming strategies contribute to yield.

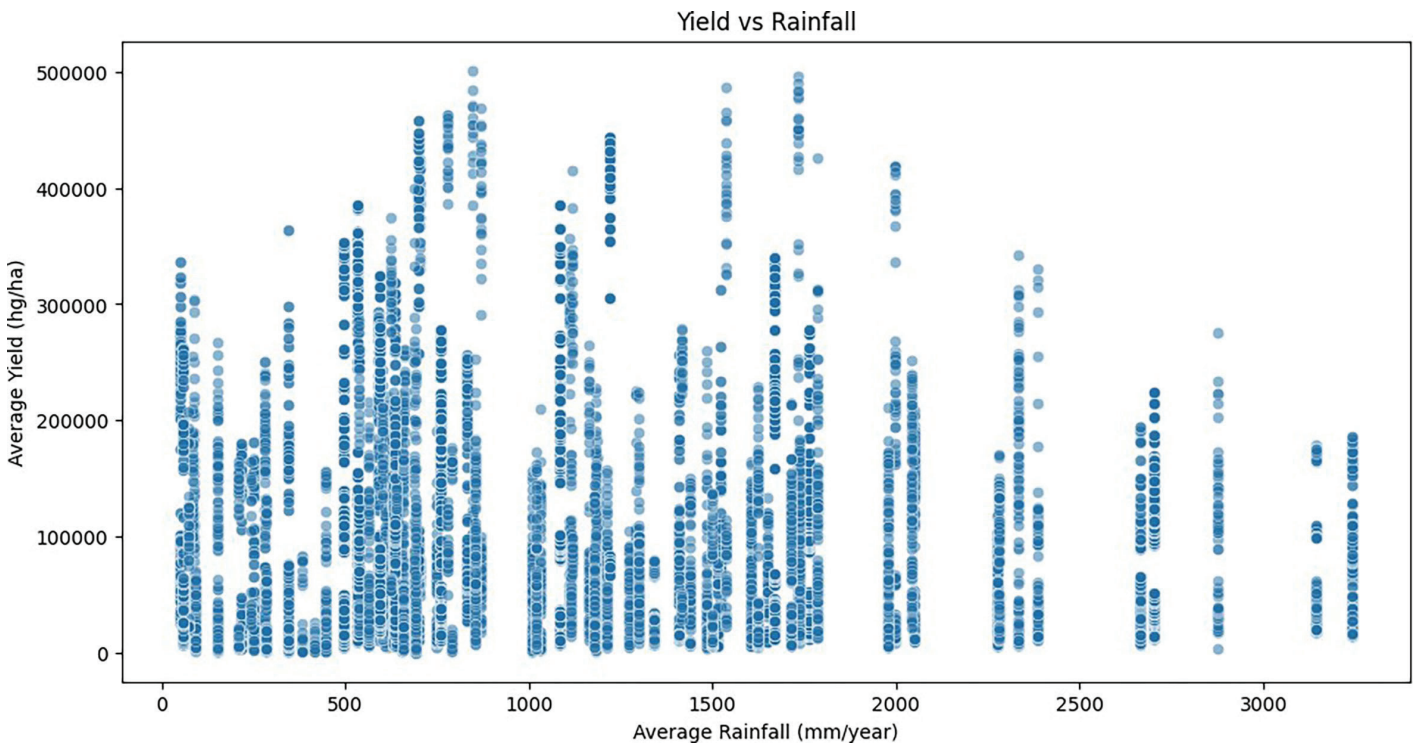

Figure 4. Relationship between average crop yield and average annual rainfall suggests that yields are consistent up to 2000 mm of annual rainfall then decline indicating excessive rainfall can reduce productivity.

higher yields to some specific level.

The yield vs pesticide plot Figure 5 also varied more, suggesting that the use of pesticides can be useful in some instances to improve yield but does not have a straightforward linear impact. The yield vs temperature plot Figure 6 indicated a moderate relationship, potentially representing the optimal ranges of temperature for crop growth.

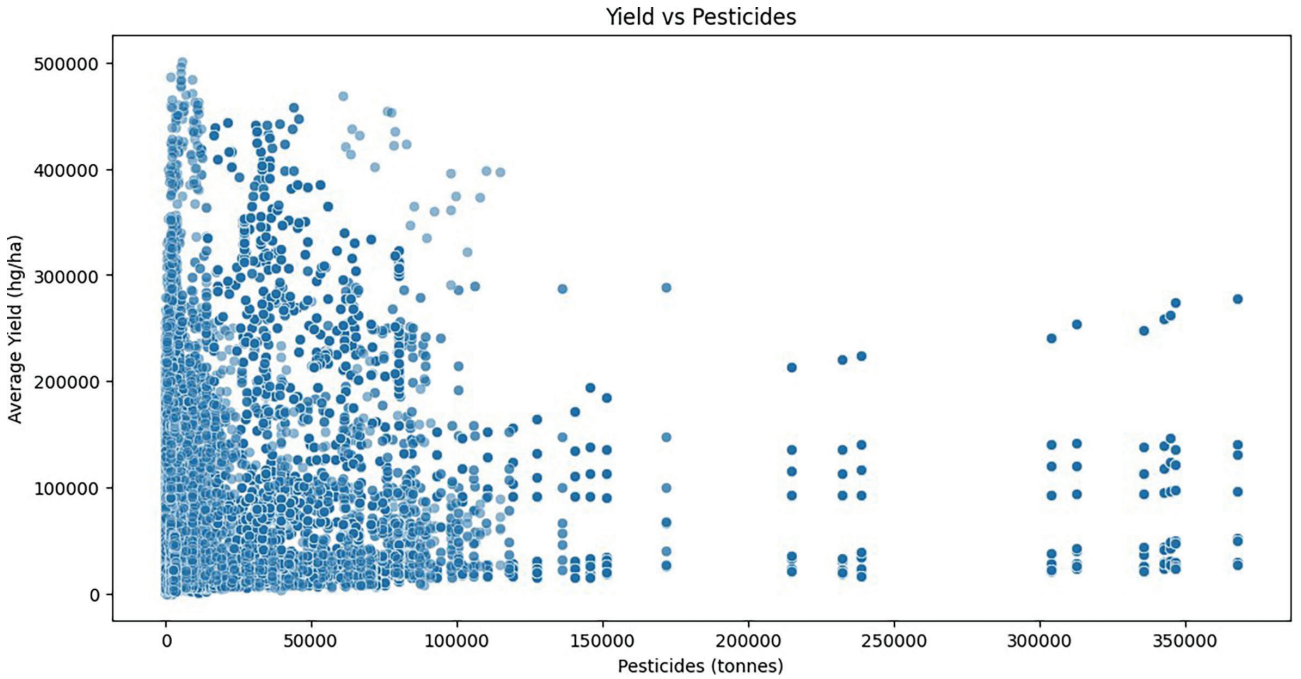

Figure 5. Relationship between average crop yield and pesticide use suggests that pesticides can sometimes improve yield, but do not have a straightforward linear impact.

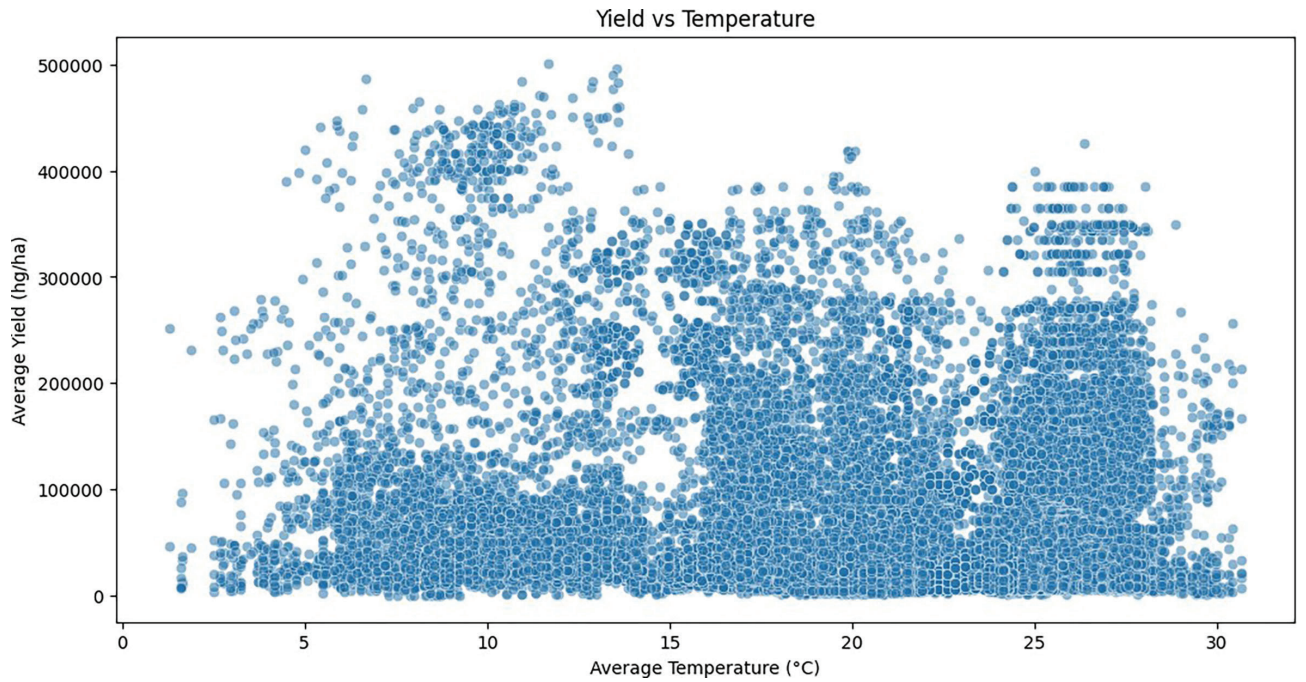

Figure 6. Relationship between average crop yield and average temperature suggests that yields are higher within certain temperature ranges, possibly representing optimal conditions for crop growth.

A correlation matrix (Figure 7) was also generated and represented in the form of a heat-map to investigate linear relationships between numeric variables. Temperature exhibited minor negative correlation with crop yield, whereas pesticide application showed a moderate positive correlation. Additionally, rainfall demonstrated no correlation with crop yield. This reinforced the fact that chosen input features possessed reasonable relations with the target variable and were suitable for model development.

Models

Four regression models were trained and evaluated: Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), and K-Nearest Neighbors (KNN). These models provided a range of algorithmic approaches, from simple linear models to complex ensemble and distance-based techniques. The dataset was randomly split into an 80% training set and a 20% test set, to assess how the model performs when encountered with unseen data. For algorithms that are sensitive to feature scaling (such as SVR and KNN), input features were scaled. Hyper-parameters of SVR, RF, and KNN were adjusted using grid search with cross-validation on training data to maximize their accuracy when predicting the yield.

Model Evaluation

After training, models were evaluated on the test data using three performance metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). Among the models, both Random Forest (untuned) and K-Nearest Neighbours (tuned) achieved the highest accuracy, though KNN demonstrated better performance with lower error metrics. In addition to its predictive performance, Random Forest also provided feature importance scores; the scores aligned with agricultural principles and supported the model's validity.

To ensure that the models built were reliable, further diagnostic tests were run. Cook's Distance was employed for the linear regression model to determine high-leverage outliers. Plots of residuals were inspected to evaluate model fit and check

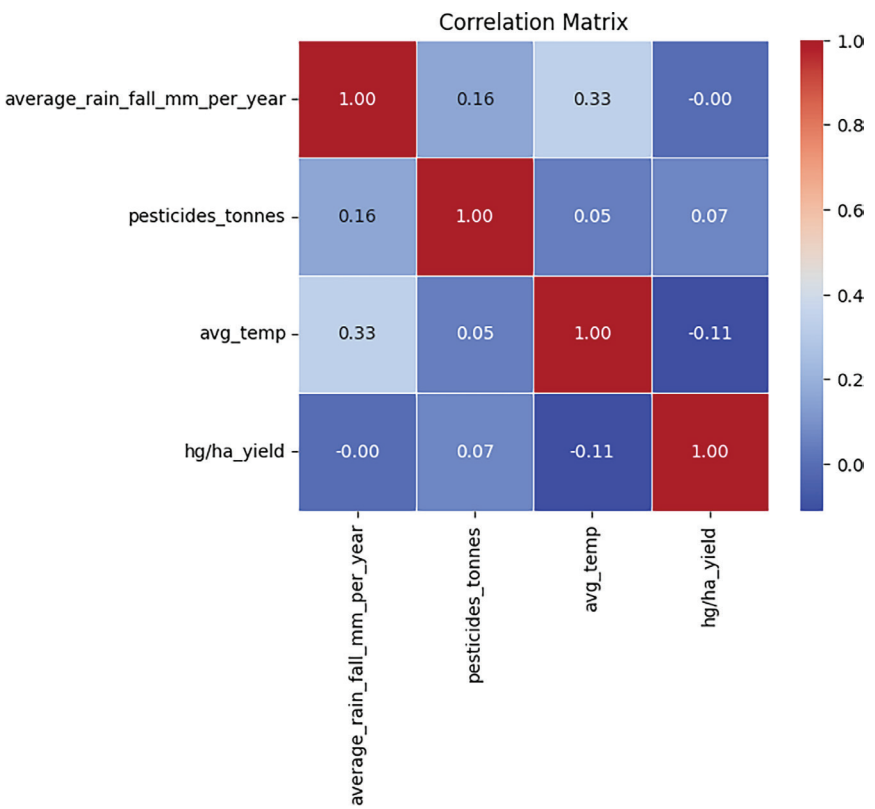

Figure 7. Correlation Matrix of numeric variables. The heatmap shows minor negative correlations between crop yield and rainfall, pesticide use, and temperature.

assumptions of homoscedasticity and linearity. Alongside these, learning curves were also produced to verify underfitting or overfitting of the training data by the models. In the end, the diagnostics validated that the Random Forest model was generalizable, further bolstering the potential of this model to be viable in increasing accessibility of simple technology for farmers.

RESULTS AND DISCUSSION

Model Performance

Model performance was evaluated based on three standard metrics: Mean Absolute Error (MAE), Mean Squared, Mean Absolute Error (MAE), and R^2 score. MAE takes the absolute values of the differences in predicted and actual values, treating any errors equally. Whereas MSE calculates the square of the differences, giving weightage larger errors. The R^2 score is an indication of how well the data can be approximated by the regression line, with a score of 1.0 representing a perfect prediction.

Table 1 indicates that the best overall performance was achieved by the tuned K-Nearest Neighbors (KNN) model, with the minimum MAE (3,257), minimum MSE (78.5 million), and an R^2 of 0.99. This success indicates that crops cultivated under similar conditions have a propensity to yield similar quantities, so KNN's local pattern-matching strategy is extremely successful.

The untuned Random Forest also fared very well, identical to KNN's R^2 but with marginally higher error values. Tuning this model resulted in a dip in performance (R^2: 0.98), indicating the default settings were already very close to the ideal. This is an important observation — over-tuning at times can decrease generalizability even when validation scores seem reliable.

Table 1. Model performance comparison using MAE, MSE and R^2

Model	MAE	MSE	R^2
Linear Regression	29,302	1,770,619,035	0.75
Support Vector Machine (Untuned)	57,187	8,597,983,016	-0.21
Support Vector Machine (Tuned)	28,273	2,812,999,343	0.61
Random Forest (Untuned)	3,384	104,180,110	0.99
Random Forest (Tuned)	4,329	112,297,465	0.98
KNN (Untuned)	4,313	108,947,951	0.98
KNN (Tuned)	3,257	78,501,924	0.99

KNN (Tuned) and Random Forest (Untuned) received the highest accuracy, whereas Linear Regression and Support Vector Machine models are less effective.

Consequently, Support Vector Regression (SVR) performed suboptimally. The untuned model yielded a negative R^2, and tuning it at 0.61 did little to improve. This indicates that SVR might not be best placed to deal with complex, irregular agricultural data unless subject to heavy preprocessing and fine-tuning.

Tuning was beneficial in two of the three models. For KNN and SVR, tuning increased accuracy by choosing superior parameters, like neighbor number or kernel choice. Random Forest, however, performed well without tuning, demonstrating its stability to hyperparameters as well as its natural robustness to high-dimensional data.

These findings underscore that one model is not best for all. KNN performed best in identifying local patterns, Random Forest was best when it came to finding a balance between accuracy and reliability, and SVR showed how imperative it is to allocate algorithms based on data nature. Employing a range of models and stringent testing through various metrics provided a better insight into what performs optimally for yield prediction in agriculture.

Model Diagnostics

To ensure our models were not only accurate but also reliable and interpretable, we used several diagnostic tools to better understand their behavior and limitations.

Learning curves Figure 8 were graphed for the models to see how the performance changed with increasing training data. In both Random Forest and KNN, training and validation scores stayed close together, which suggests good generalization. The R^2 values of the two models are consistent across a wide range of training sizes, suggesting that the models avoided overfitting and underfitting. This consistency suggests the models would maintain relatively high R^2 values with additional training data.

Residual plots Figure 9 and 10 gave us a sense of the distribution of prediction errors. In the models that performed better, residuals randomly dispersed around zero with no strong patterns evident. This is a strong sign as it indicates that the models were not biased to

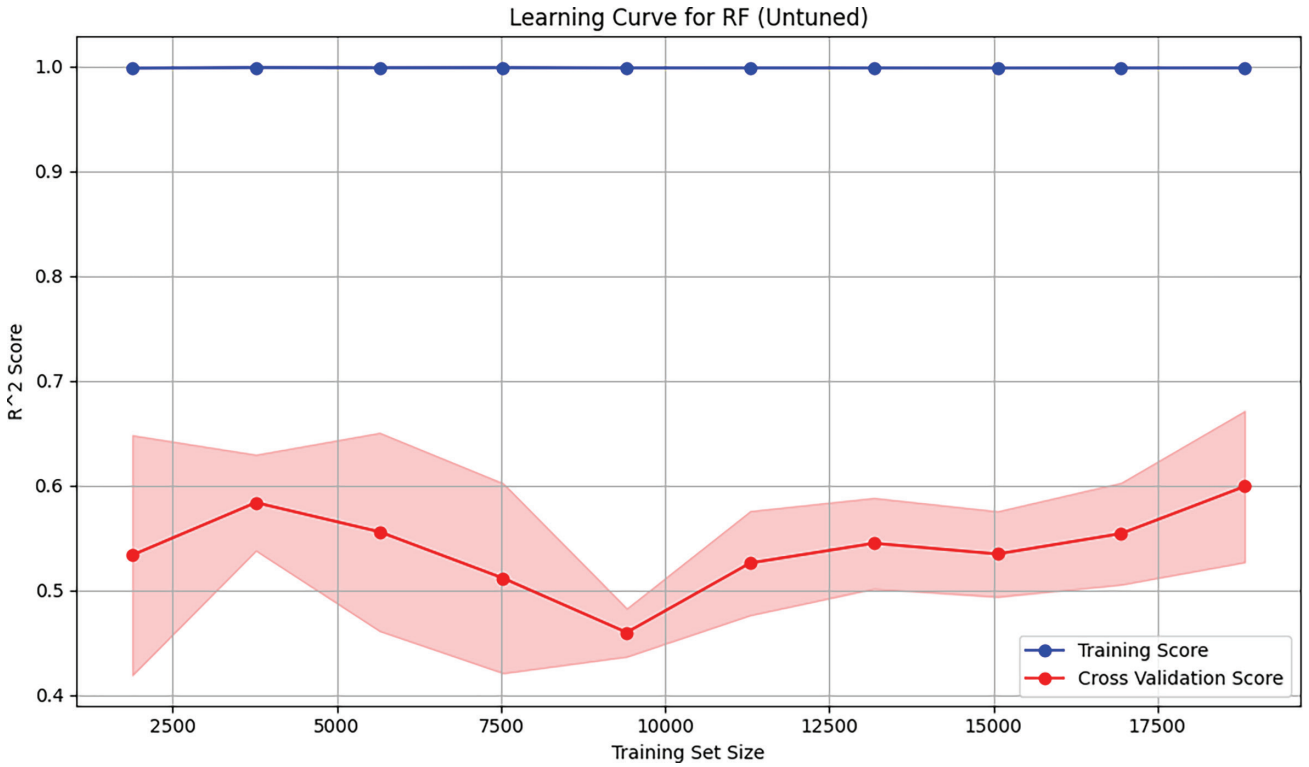

Figure 8. Learning curve for Random Forest (Untuned) model. R^2 stay close across different training sizes, indicating strong generalisation and minimal risk of overfitting and underfitting.

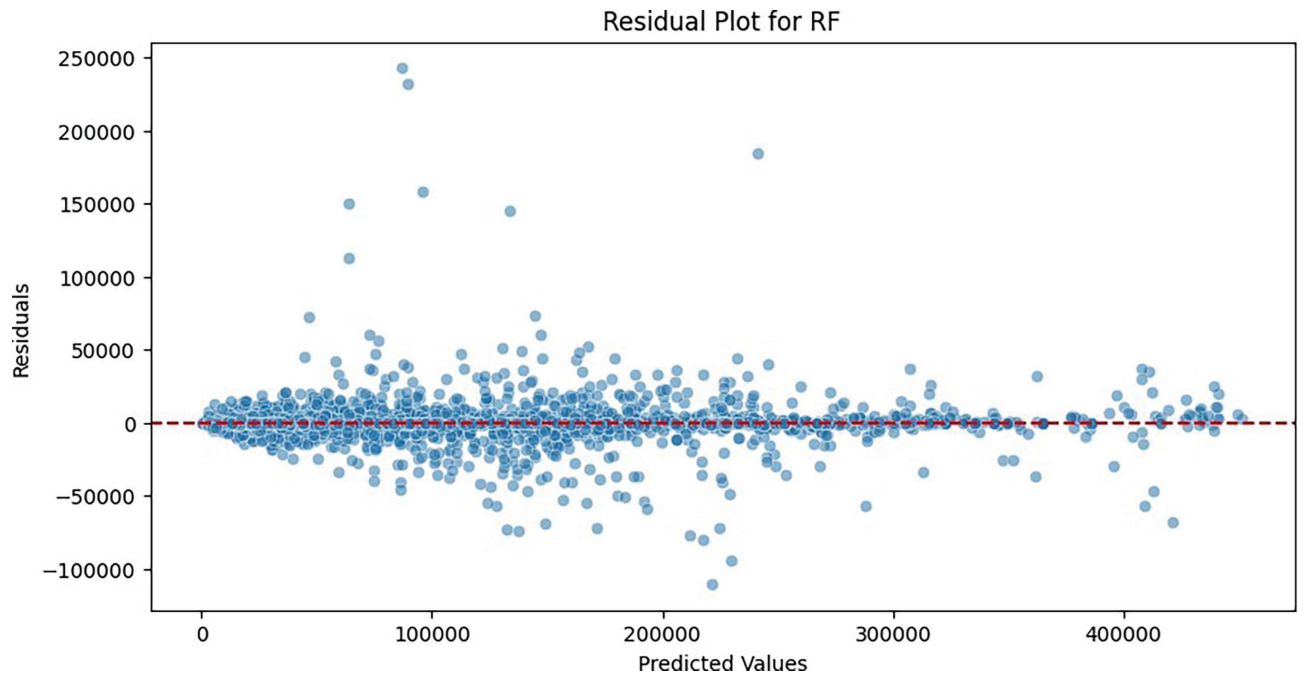

Figure 9. Residual plot for Random Forest model. Prediction errors are randomly scattered around zero with no obvious pattern, suggesting the model captures most of the data structure without systematic bias.

over or underpredict yields within some range. Rather, errors seemed to be evenly spread, indicating that the models were picking up most of the interesting structure in the data.

To look for outlier data points, we applied Cook's Distance Figure 11 to the linear regression model. This will identify individual observations that could have an abnormal influence on the regression fit.

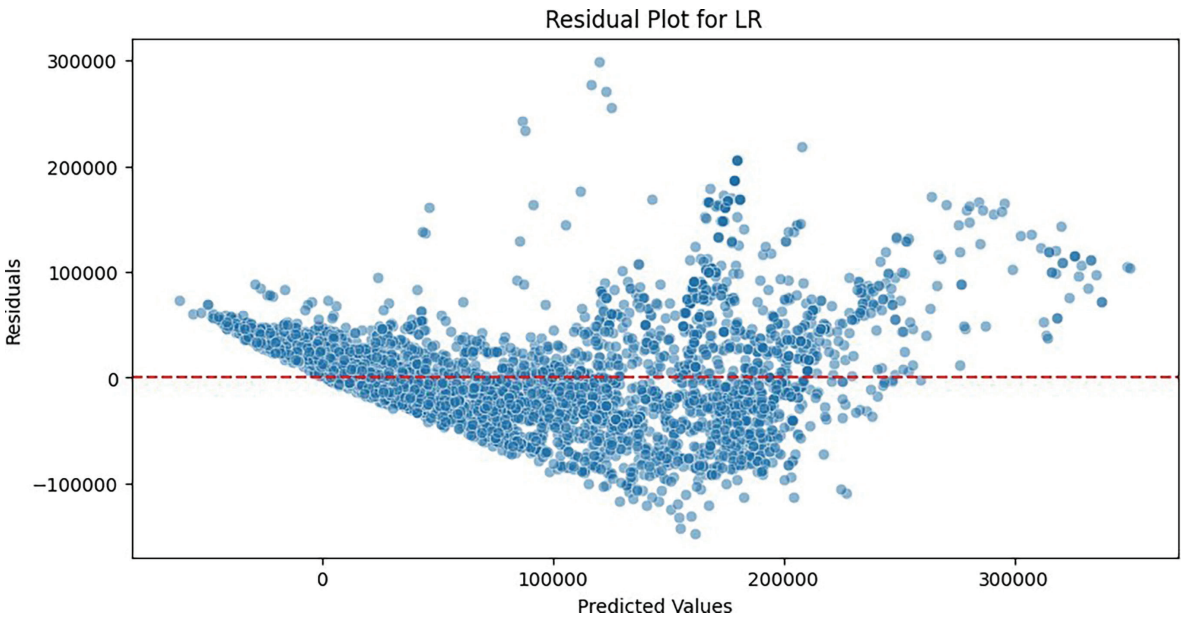

Figure 10. Residual plot for Linear Regression model. Errors are somewhat evenly distributed around zero, although performance is less accurate than more complex models.

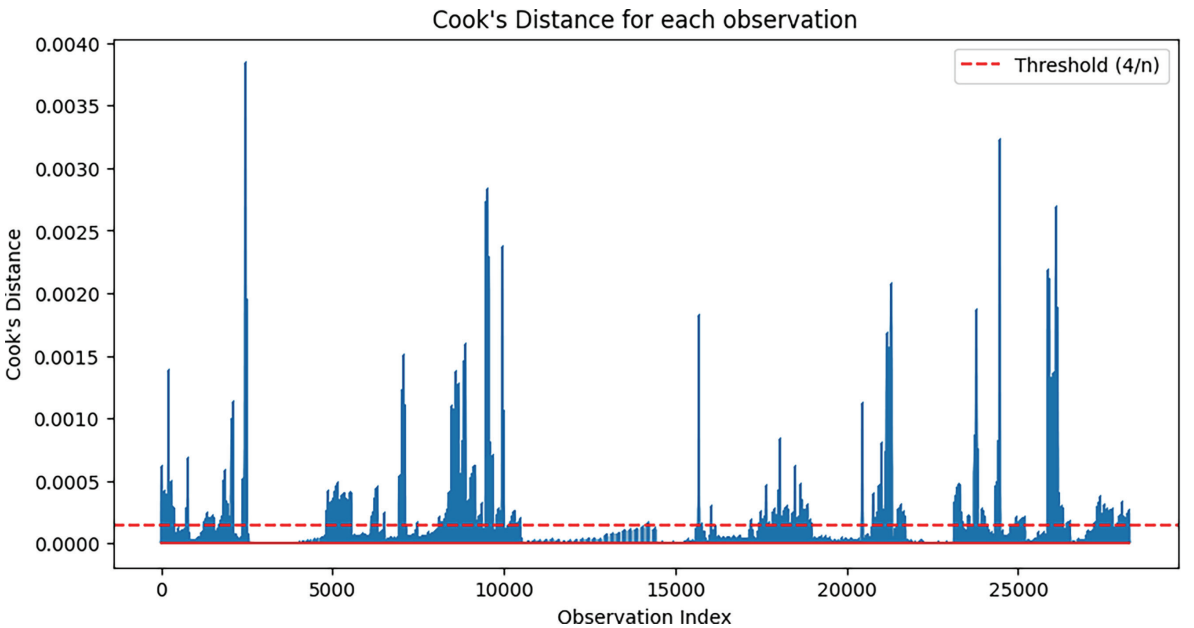

Figure 11. Cook's distance values for all observations in the Linear Regression model. All points fall below the threshold of 1, indicating no single observation has undue influence on the regression fit.

In our situation, all of the values fell far below the standard threshold value of 1, and thus no data point skewed the model abnormally. This is an endorsement of the reliability of the data set and the strength of the linear model, although it trailed more sophisticated approaches.

To gain insight into which variables were most significant in yield prediction, we looked at Random Forest model feature importance scores Figure 12. In doing so, certain crops such as potatoes and cassava emerged to have high importance factors. This may be due to the dataset containing high variance in yields for these crops or these crops display strong patterns connected to other factors. In terms of ecological variables, pesticides were indicated to have a significant role in yield determination. This is possibly due to their direct application onto crops as well as their role of defending crops against harsh climatic conditions. Furthermore, rainfall and temperature had relatively equal and high importance among all the categorical variables. These findings further validated the significance of water availability in the development of crops and were consistent with the role temperature plays in the metabolism of plants.

This ordering not only conforms with common agricultural knowledge but also makes the model more interpretable for both farmers and scientists. Once users know the conditions that are most important, they can concentrate on monitoring and enhancing those. As an example, under conditions with uneven rain, farmers could prioritise the need for enhanced irrigation tactics. This form of interpretability can narrow the gap between machine learning models and precise decision-making in agriculture.

LIMITATIONS

While the models performed well overall, there are some limitations to be highlighted. The extremely high R^2 values, particularly for Random Forest and KNN, bring the risk of overfitting into question. Though diagnostics such as learning curves and residual plots did not flag any instability, such near-perfect scores might not reflect similarly on new or different data. Which is why, external testing in the future could be necessary to verify how well the model generalizes.

Additionally, the Support Vector Regression was also lacking, even when it was tuned. This indicates that

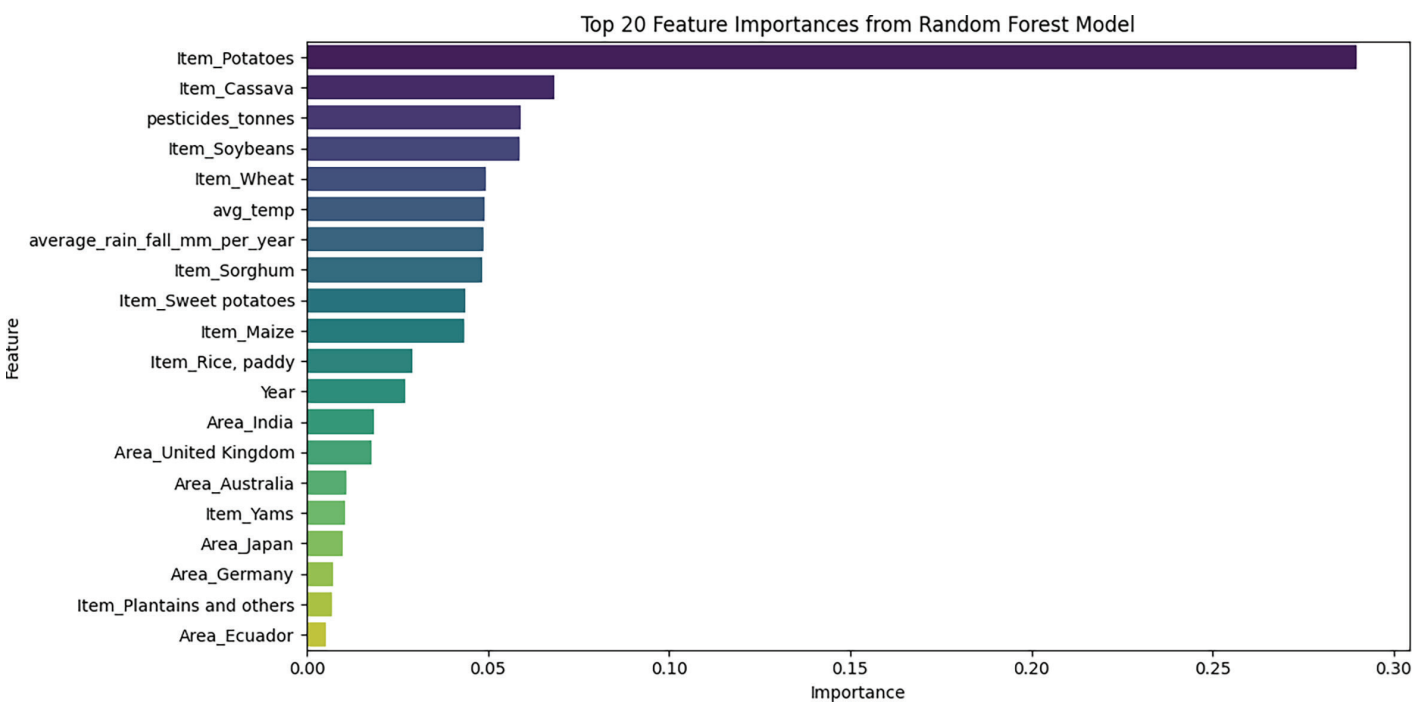

Figure 12. Top 20 feature importance from the Random Forest model. Potatoes, cassava, and pesticides contributed most to the model.

not every algorithm is appropriate for this form of data and reinforces the need for appropriate model selection, especially when attempting to solve agricultural problems due to the non-uniform nature of agricultural data.

Furthermore, the dataset itself was incomplete. Some of the most important variables such as soil quality, certain crop types, and cultivation practices were missing. These omitted characteristics are major factors that influence yield and would make the model less precise or less applicable across different areas. Taking these variables into consideration when building upon this work would vastly improve the model's validity and usability in day to day farming practices.

Finally, even though Cook's Distance did not detect any outliers, actual agricultural data in the real world tend to be irregular. Naturally, inconsistencies in reporting the data, regional biases, or data entry mistakes may still be present. It is vital that these discrepancies must be considered as the model is refined and readied for actual application.

COMPARISON WITH PREVIOUS WORK

Most current research in the field of farmland planning, and more specifically crop rotation use AI to focus on rule-based systems. Although methods like reinforcement learning and satellite data processing are extremely interesting paths to evolve agricultural practices, they stem from reduced logic and do not directly help farmers with actionable predictions that can be adopted with minimal forethought and maximum user-friendliness.

This study is unique from other studies in that it predicts crop yield from measurable variables such as rainfall, temperature, and the application of pesticides. It compares tuned and untuned versions of several machine learning algorithms to see which algorithms perform best with real data. This broader comparison is sometimes missing from other studies, which tend to look only at one algorithm.

One of the positive aspects of this project is that it places high value on model transparency and robustness. Feature importance, residual plots, Cook's Distance, and learning curves were utilized to characterize how the models work and their stability. These diagnostics, as well as visualizations like scatter plots and a correlation heatmap, help make the results more interpretable and understandable.

One area in which prior research can build is by using more advanced data sources such as NDVI or soil type. While this project focuses on practicable features, it provides a solid foundation upon which more exact inputs can build in further research. Generally, this research stands out with its applied focus, solid performance.

CONCLUSION

This project aimed to determine if machine learning algorithms could predict crop yield with accuracy from environmental and input-based variables such as temperature, rainfall, and pesticides. The results were positive and reflected the suitability of the machine learning models for predicting crop yield. Both KNN and Random Forest models performed exceptionally high, accompanied by low MAE and MSE, for highly accurate and reliable performance. The analysis also delivered insights into feature importances, which could go one step further in assisting farmers understand the needs of their farmlands.

Though, there are some scopes for improvement. The high R^2 values indicate potential overfitting, and SVR models underperformed despite hyperparameter tuning, indicating that the model choice is a factor that cannot be overlooked. Examination of multicollinearity between features would make the models more reliable, particularly if there are highly correlated variables. Also, it is important to acknowledge did not include key agronomic variables such as soil quality, crop type, and farming methods. These factors are essential for ecological validity, since they strongly influence yield outcomes in real-world settings. Their omission limits the extent to which the models can capture agricultural complexity and future research that incorporates these features will enhance accuracy and practical applicability.

In the future, the project can be scaled up to a web application that has these machine learning tools embedded. The aim is to have a real-time platform where local data can be entered by farmers and they can receive customized information on yield prediction, crop rotation, and planning. This will fill the gap between sophisticated analytics and decision-making for farmers' daily lives, giving farmers direct assistance.

ACKNOWLEDGEMENTS

I would like to acknowledge my mentor John Basbagil from Lumiere Education.

CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. OECD, Why does the financial sector need to think about water risks? Available from: https://www.oecd-ilibrary.org/en/blogs/2024/05/why-does-the-financial-sector-need-to-think-about-water-risks.html (accessed on 2025-3-14)
2. The Economist. Climate change and population growth are making the world's water woes more urgent. Available from: https://www.economist.com/special-report/2019/02/28/climate-change-and-population-growth-are-making-the-worlds-water-woes-more-urgent (accessed on 2025-3-19)
3. UNESCO, The United Nations World Water Development Report 2023. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000380733 (accessed on 2025-3-20)
4. Agriculture Institute. Crop rotation for soil and water conservation. Available from: https://agriculture.institute/rain-fed-farming/crop-rotation-soil-water-conservation/ (accessed on 2025-3-15)
5. Fenz S, Neubauer T, Friedel JK & Wohlmuth M-L. AI- and data-driven crop rotation planning. *Computers and Electronics in Agriculture*. 2023; 212: 108160. https://doi.org/10.1016/j.compag.2023.108160
6. Fenz S, Neubauer T, Heurix J, Friedel JK & Wohlmuth M-L. AI- and data-driven pre-crop values and crop rotation matrices. *European Journal of Agronomy*. 2023; 150: 126949. https://doi.org/10.1016/j.eja.2023.126949
7. Liang Z, Xu Z, Cheng J, Ma B, et al. Designing diversified crop rotations to advance sustainability: A method and an application. *Sustainable Production and Consumption*. 2023; 40: 532–544. https://doi.org/10.1016/j.spc.2023.07.018
8. Patel R. Crop Yield Prediction Dataset. Available from: https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset (accessed on 2025-3-25)