Original Research Article

RANDOM FOREST REGRESSION IN MAIZE YIELD PREDICTION

UNDER PEER REVIEW

ABSTRACT

Artificial Intelligence is the discipline of making computers behave without explicit programming. Machine learning is a subset of artificial Intelligence that enables machines to learn autonomously from previous data without explicit programming. The purpose of machine learning in agriculture is to increase crop yield and quality in the agricultural sector. It is driven by the emergence of big data technologies and high-performance computation, which provide new opportunities to unravel, quantify, and comprehend data-intensive agricultural operational processes. Random Forest is an ensemble technique that reduces the result'soverfitting. This algorithm is primarily utilized for forecasting. It generates a forest with numerous trees. The random forest classifier predicts that the model's accuracy will increase as the number of trees in the forest increases. All through the training phase, multiple decision trees are constructed. It generates subsets of data from randomly selected training samples with replacement. Each data subset is employed to train decision trees. It utilizes multiple trees to reduce the possibility of overfitting. Maize is a staple food in Kenya and having it in sufficient amounts in the country assures the farmers' food security and economic stability. This study predicted maize yield in the Kenyan county of UasinGishu using the machine learning algorithm Random Forest regression. The regression model employed a mixed-methods research design, and the survey employed well-structured questionnaires containing quantitative and qualitative variables, which were directly administered to 30 clustered wards' representative farmers. The questionnaire encompassed 30 maize production-related variables from 900 randomly selected maize producers in 30 wards. The model was able to identify important variables from the dataset and predicted maize yield. The prediction evaluation used machine learning regression metrics, Root Mean Squared error-RMSE=0.52199, Mean Squared Error-MSE =0.27248, and Mean Absolute Error-MAE =

0.471722. The model predicted maize yield and indicated the contribution of each variable to the overall prediction.

Key words: Artificial Intelligence, Machine Learning, Random Forests, Big Data, Decision Tree(s),Algorithm

1.INTRODUCTION

Artificial intelligence (AI) is defined as "a system's ability to interpret external data accurately, learn from such data, and use these insights to achieve defined objectives and duties through flexible adaptation."1 As a field of study in the 1950s, AI remained relatively scientifically obscure and of limited practical relevance for over half a century. Today, as a result of the increasing popularity of Big Data and advances in computing capacity, it is now part of the business world and public discourse. (Haenlein& Kaplan, 2019).

Machine learning is at the intersection of cybernetics (control science) and computer science and is currently attracting a great deal of professional and public interest. From the many decades of the history of transition and learning, it can be seen that during the first few decades, these two answers are very near to one another, and productive cooperation can be observed between them. (Alexander,2020)
Today, as a result of the increasing popularity of Big Data and advances in computing capacity, it is now part of the business world and public discourse. (Haenlein& Kaplan, 2019).Random Forest is an ensemble technique that reduces the result's overfitting. This algorithm is primarily utilized for forecasting. Voting is applied to the training data to derive the prediction. Regression is the technique for predicting the relationship between dependent and independent variables. Regression serves a crucial role in data modeling and analysis. The relationship between the dependent and independent variables in linear regression is linear.

On the African continent, the deployment of artificial intelligence (AI) technologies is proliferating, but policy responses are still in their infancy. It is argued that in order for artificial intelligence to promote rather than undermine socioeconomic inclusion in African contexts, policymakers must be aware of the following critical dimensions: gender equity, cultural and linguistic diversity, and labor market shifts.

UNDER PEER REVIEW

(Gwagwa et al, .2021)

To determine a course of action, the decision tree algorithm uses a tree-shaped diagram. Each limb of a tree represents a potential choice. The algorithm for a random forest is a collection of decision trees. The purpose of machine learning in agriculture is to increase crop yield and quality in the agricultural sector. This agriculture technology is used by seed retailers to analyze data and produce improved crops.

Recent research has demonstrated that machine learning (ML) can provide accurate forecasts more quickly and flexibly than crop simulation modeling. However, a "committee" of machine learning models (machine learning ensembles) that can reduce prediction bias, variance, or both and better reflect the data's underlying distribution can outperform a single machine learning model.

In 2021, Brown et al,utilized interpretable random forest models that could yield estimates of a set of (potentially correlated) malnutrition and poverty prevalence measures utilizing freely available, regularly updated, georeferenced data. In 2020, V Geetha used a random forest algorithm to predict yield in various crop datasets.These datasets were utilized for both training and evaluation purposes. Random Forest classifiers demonstrated a tremendous capacity to forecast crop yield. Various results indicate that Random Forest is an effective learning algorithm for analyzing crops under current climatic conditions and has a high degree of data investigation precision. (Geetha et al., 2020)

A study conducted in 2020 aimed to assess the effectiveness of random forest regression in predicting maize production. To enhance the algorithm's performance, a ranking-based technique was proposed and implemented. The ranking of individual vegetative indices (VIs) was determined based on a merit metric that relied on the correlation parameter. The ranking revealed the key vegetation indices to be utilized in the model. The results of the evaluation indicated that the ranking process resulted in a random forest model for maize yield prediction that exhibited a relatively high level of performance.
(Marques.,2020)

A research investigation was conducted with the objective of forecasting the yield of bio-corn through the utilization of four distinct machine learning algorithms, one of which was the random forest technique. The study employed a genetic algorithm to assess the relative significance of variables in terms of feature importance. The authors developed a software package for predictive purposes, after the outstanding performance of the Random Forest method compared to the other three algorithms. The investigation provided valuable insights into the pyrolysis

UNDER PEER REVIEW

process, leading to the development of a more simplified technique for predicting yield. (Ullah.,2021)

Crop yield forecasting was conducted by considering climatic and edaphic parameters of production, such as rainfall, temperature, pH, and humidity, for multiple crops to assess the crop with the highest yield potential. The analysis was conducted using the random forest regression machine learning algorithm. The algorithm had the highest rates of prediction accuracy, making it the most suitable choice for yield prediction recommendations. (Kumar et al., 2020)

The decision-making subphase of agricultural management that generates profits is crop yield prediction. Predicting yield is dependent on the soil, the numerous atmospheric conditions, and the climate. In addition to sowing and yield prediction, the crop is also nourished and maintained physically, which may be more efficient. In other terms, crop yield production is a measurement of the amount of food grains. With the aid of ML techniques, several models have been developed that take as input variables such as soils, weather conditions, crop name, insect information, and water level and perform analyses with the crop yield as the output. (Bhansali, Saxena& Bandhu, 2021).

Literature reports regression, simulation, expert systems, and artificial neural networks (ANN) as yield prediction methodologies. Regression models have been extensively utilized in numerous studies, especially for forecasting purposes. (Santin et al.,2019)

According to the 2014 Kenya Demographic and Health Survey, infant mortality in Kenya. They investigated the Logistic regression, K-neighbor, and random forest algorithms. Comparatively, Random Forest performed admirably with high levels of accuracy of approximately 97.1%, followed by the Logistic Regression model with 86.1% and K-nearest neighbor with 85.5%. The study determined that the random forest model performed the best. (Kioko, C., 2022).

According to Kenya Agricultural and Livestock Research Organization-KALRO, millions of Kenyans rely on maize as a staple food. The total area devoted to maize production is approximately 1.5 million hectares, with an estimated annual average production of 3.0 million metric tons, resulting in a national average yield of 2.0 tons per hectare. In the high-potential highlands of Kenya, harvests typically range from 4 to 8 T/Ha, representing 50% (or less) of the genetic potentials of the hybrids

UNDER PEER REVIEW

Several Kenyan cash commodities, including Tea, flowers, maize, wheat, cotton, coffee, and pyrethrum, can be evaluated for their ability to predict yield. Given that maize is a staple crop that is cultivated in the majority of Kenya, this study was based on maize yield estimations.

This research expands on these findings and contributes to the foundations of the application of machine learning to maize yield prediction in UasinGishu County using a multitude of features or variables.The research predicted maize yield in UasinGishu County using Random Forest regression algorithm.

2.Materials and Methods

2.1 Study Area

The study area was in UasinGishu County. It is in the North Rift region of Kenya. UasinGishu County covers an area of 3346 km2 (2995 km2 arable, 333 km2 non-arable, 23km2 water masses, and 196 km2 urban area).

Table 1: List of Wards Per Sub County of UasinGishu County.

Sub-county	Wards
Ainabkoi	Ainabkoi/ Olare, Kaptagat, Kapsoya,
Kapseret	Kipkenyo, Simat, Ngeria, Megun, Langas,
Kesses	Tulwet, Cheptiret, Racecourse, Tarakwa
Moiben	Kimumu, Moiben, Karuna/Meibeki, Sergoit, Tembelio,
Soy	Soy, Kuinet, Ziwa, Kipsomba,Moisbridge, Kapkures, Segero,
Turbo	Huruma, Ng'enyilel, Tapsagoi, Kiplombe, Kapsaos, Kamagut.

The County has a population of 1,163,186 persons: 301110 households and 213,982 farm families. The average household size is 4, with a farm holding size of 10 acres, 65% of which is titled. On average, 42% of the land holding is put under commercial crop (mainly maize), 12% under subsistence crop,105 under improved pastures and forage, 16% under natural pastures, 6% underwood lot, 9% is unusable, and 5% under homestead. The per capita land holding in the County is estimated at 2.5 acres, which is economically low for the major crop enterprises. This is one of the reasons for the high poverty level in the

UNDER PEER REVIEW

County, which averages 46%, and 32 % of the population experience food insecurity 3-4 months in a year. Of the total population, only 20% fall in the high food diversity group (3 food groups).

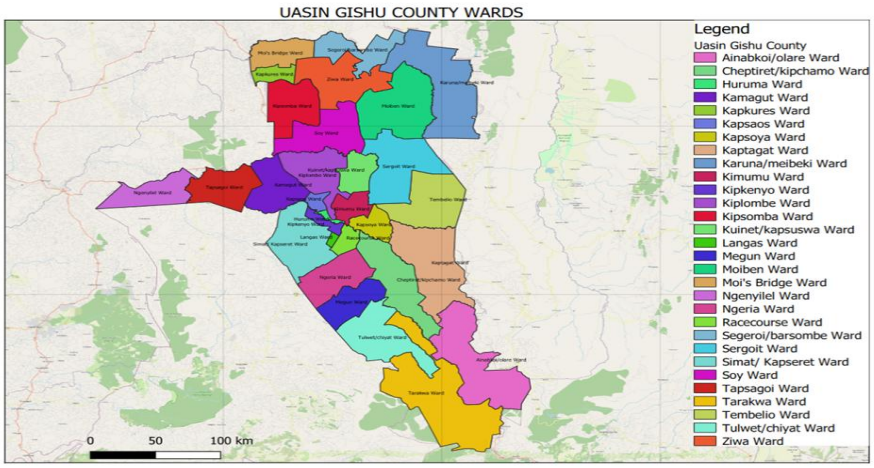

Figure 1: Map of UasinGishu County Displaying the 30 Wards.

[Source: Author]

2.2 Study Research Design

In this study, the mixed-methods research design, the survey employed well-structured questionnaires comprising of quantitative and qualitative variables, directly administered to selected representative farmers at a specific period.

To get a representative sample from the County, a survey-monkey online sample size calculator was used to compute the sample size. At a 99.9% confidence level,5% margin of error, % population proportion of 50%, and a total county population of 1,163,186, the appropriate sample size was estimated at 1082 farmers for this study. One thousand eighty-two farmers spread across 30 wards implied thirty-six-36 farmers per ward. This study targeted a mix of both small- and large-scale maize farmers. Due to financial constraints, the study managed 30 farmers per ward, presenting a final sample size of 900 farmers from the County. Each ward was designated as a stratum, and ten farmers from three villages within each ward were selected at random to form the sample. This was done to account for any potential similarities in the perspectives or agricultural practices of county maize producers.

Table 2: Brief Description of Questionnaire Variables

UNDER PEER REVIEW

No.	Variable	Brief Description
1	Sub-county	Sub countylocation of the farmer
2	Ward	Ward location within the sub county of the farmer
3	Gender	Gender of the farmer
4	Age	Age bracket of the farmer
5	Education	Highest completed level of education of the farmer
6	HH size	Household size of the farmers' family
7	Full-time	If the farmer is a full-time farmer
8	Marital	Marital status of the farmer
9	Decision M	Who oversees decision-making in the maize farm
10	Credit	If the farmer seeks credit/loans to facilitate maize farming
11	Years	Length of time in years the farmer has been farming maize
12	Variety	Latest maize variety the farmer planted
13	Fertilizer	Type of fertilizer the farmer uses in maize production
14	Subsidy	If the farmer was a beneficiary of fertilizer subsidy program
15	Soil test	If the farmer had undertaken soil testing and analysis
16	Tillage	Method of land tillage used by the farmer in maize production
17	Ownership	who owns the land used in maize production
18	Size	Size of land allocated for maize production
19	Machinery	Ownership of machinery used in maize production
20	Cropping system	Cropping system used in maize production
21	Labor	Source of labor used in maize farm
22	Sale stage	Stage of sale of maize
23	Sale time	Time of sale of maize
24	Sale point	Point of sale of maize
25	Yield	Maize yield in number of bags per acre.

UNDER PEER REVIEW

2.3 Assumptions of the study

- Land types, physiographic factors such as topography, altitude and exposure to light and wind, changing weather patterns, and farmers' economic and socio-economic status were held constant.
- Edaphic factors (soil) such as moisture content, mineral and organic content, presence of other organisms, and PH levels were held constant.
- Climatic factors and the not mentioned agronomic practices adopted by maize farmers at various plant growth stages were held constant.
- The errors were independent and identically distributed with zero mean and common variance.
- The farmers were assumed to be independent of each other.

3.Data Analysis

Database creation, coding, and entry were done on Epi Data software. After data entry, it was exported to Microsoft Excel, SPSS, and R programming language for analysis. QGIS (Quantum Geographic Information System) was used in creation of the study area map.

The study involved working with categorical variables in coding and analysis. Before implementing machine learning regression algorithms on categorical variables. This study used categorical features as factors in R using *the As.factor*command in R software.

In this research, the split ratio between Train and Tests was 80:20. The training data set consisted of (720)80% of the 900 observations, while the test data set comprised the remaining (180)20%.

3.Results and Discussion

A random forest model was built in R software using "*randomForest*" and"*Dplyr*" packages.

In this random forest regression model, each decision tree in the forest contributed to the overall prediction. Feature importance was done in the model as the sum of decreased impurity. It measured the

UNDER PEER REVIEW

reduction in the impurity of the nodes in the decision tree. It was able to indicate the contribution of each variable to the overall prediction.

a) Plotting the out-of-bag error versus the number of trees to select the optimal value:

When training a random forest model, a common approach to select the optimal number of trees is to plot the out-of-bag (OOB) error as a function of the number of trees in the forest. The OOB error is the average error rate of each decision tree in the forest on the data points not included in that tree's bootstrap sample.

Therefore, by observing the plot of OOB error versus the number of trees, one can identify where the OOB error begins to level off and choose that as the optimal number of trees for the model. The optimal number of trees is the one that provides the lowest OOB error without overfitting.

Plotting the out-of-bag error (MSE) versus the number of trees to select the optimal value gave the output plot, as shown below in Figure 2, showing that the mean squared error was reducing as the number of trees increased.

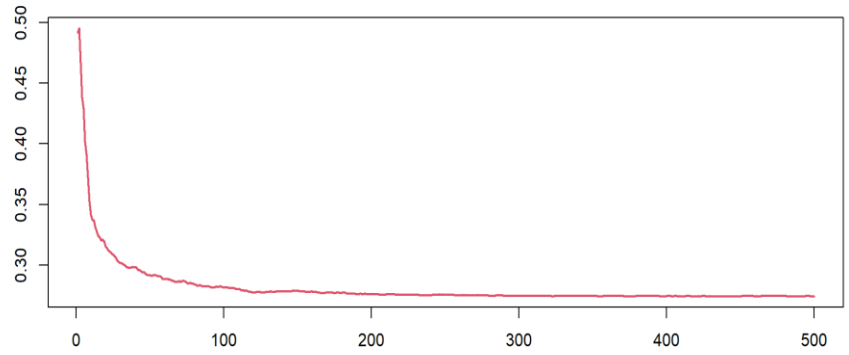

Figure 2: Plot of Out-of-bag Error versus Number of Trees.

An optimal mtry value was set at 6 to optimize the random forest model. Bag (OOB) error is a performance metric calculated using the samples that are not used in the model's training, called out-of-bag samples. These samples provide an unbiased estimate of the model's performance. Out-of-bag error estimates the error rate (1 - accuracy) that this training approach has for new data from the same

distribution. This estimate is based on the predictions obtained for each data point by only averaging those trees for which the record was not in the training data.

The plots for the out-of-bag and test error rates are displayed in Figure 3, where the x-axis has the mtry values while the y- axis has the out of bag error rates.

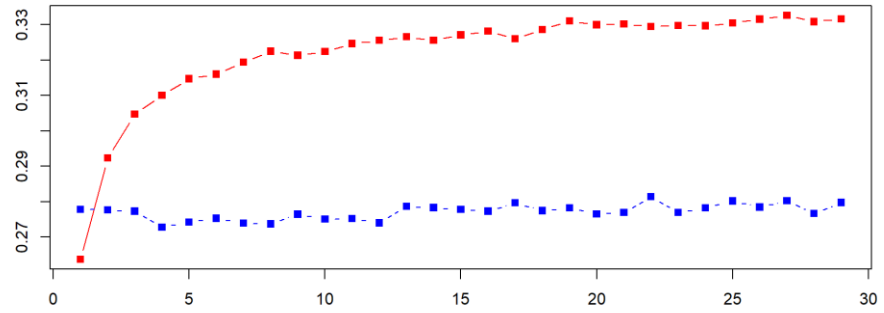

Figure 3: Plot of Out of bag Error Rates and Mtry Values.

Running a final model on the train dataset gave the output below.

Call:

$$Randomforest(formula = Y.avg \sim x3 + x4 + x5 + x6 + x8 + x9 + x10 + x11 + x12 + x13 + x14 + x15 + x16 + x17 + x18 + x19 + x20 + x21 + x22 + x23 + x24 + a + b + c + d + e + f + g + h, data = train\_data, ntree = 500, mtry = 6)$$

Type of random forest: regression

Number of trees: 1500

No. of variables tried at each split: 6

Mean of squared residuals: 0.2858586

% Var explained: 4.12

An optimum value was obtained at *ntree*=1500 and *mtry*=6.Variable importance as per the final model.

UNDER PEER REVIEW

Table 3: Display for Variable Importance for the Final Model.

Variable	Importance
Variety grown	14.166297
subcounty	13.482586
Post harvest process	11.113343
Weather patterns	10.985592
Maize yield factors	10.684125
Maize seeds	10.474045
Fertilizer use	10.365862
Cost of production	10.2663
Production process	9.994013
Farming practices	9.922929
Labor source	8.81613
Age of farmer	8.353711
Decision maker	7.424476
Fertilizer	6.014735

UNDER PEER REVIEW

type	
Years in farming	5.818921
Time of sale	5.323495
Education level	5.075838
Point of sale	4.251471
Cropping system	4.113893
Stage of sale	3.751986

The model performance evaluation is as displayed in the table below alongside respective evaluation metrices.

Table 4: Model Performance Evaluation

Metric	MSE	MAE	RMSE	MAPE
Index	0.2724817	0.4717222	0.5219978	38.13801

4.CONCLUSIONS

Through rigorous research and critical analysis, several key findings and insights have been obtained, shedding light on various aspects of relations between factors affecting maize yield predictions in UasinGishucounty. Further, the final data set was able to predict maize yield and indicate the essential variables in each model, using Random Forest regression algorithm.

Given the availability of big data in agricultural production, future studies can be done to explore other machine learning regression techniques from single tools to hybrid tools (combined linear and non-linear models) to evaluate their efficiency in predictive modeling and predicting maize yield in specific Kenyan counties based on aggregated price data from all counties and or based on the factors affecting maize production in the respective area of study.Finally, the everchanging developments in the tech space

UNDER PEER REVIEW

indicate a need to fast implement machine learning techniques to make Agriculture more sustainable for future generations in Kenya.aize is one of many export earners for Kenya. Yield predictions using artificial intelligence or machine learning techniques can be evaluated on Tea, coffee, and cut flowers.

ETHICAL APPROVAL

The study received ethical licenses and was permitted through the National Commission for Science, Technology, and Innovation-NACOSTI on 21ST February 2022 and from the Board of Post-Graduate Studies of the University of Eldoret.

REFERENCES

1. Alexander L. Fradkov, Early History of Machine Learning, IFAC-Papers Online,Volume 53, Issue 2, 2020, Pages 1385-1390, ISSN 24058963, https://doi.org/10.1016/j.ifacol.2020.12.1888.
2. Bhansali, A., Saxena, S. &Bandhu, K. (2021). Chapter 8 Machine learning for sustainable agriculture. In K. Kant Hiran, D. Khazanchi, A. Kumar Vyas& S. Padmanaban (Ed.), *Machine Learning for Sustainable Development* (pp. 129-146). Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110702514-008
3. Browne, C., Matteson, D., McBride, L., Hu, L., Liu, Y., & Sun, Y. et al. (2021). Multivariate random forest prediction of poverty and malnutrition prevalence. *PLOS ONE*, *16*(9), e0255519. doi:10.1371/journal.pone.0255519
4. Geetha, V., Punitha, A., Abarna, M., Akshaya, M., Illakiya, S., &Janani, A. (2020). An Effective Crop Prediction Using Random Forest Algorithm. *2020 International Conference On System, Computation, Automation And Networking (ICSCAN)*. doi: 10.1109/icscan49426.2020.9262311
5. Gwagwa, A., Kraemer-Mbula, E., Rizk, N., Rutenberg, I., & De Beer, J. (2015). Artificial Intelligence (AI) Deployments in Africa: Benefits, Challenges and Policy Dimensions. *The African Journal Of Information And Communication*, *26*(26), 1-28. doi: 10.23962/10539/30361
6. Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5-14. doi: 10.1177/0008125619864925
7. Kioko, C., 2022. Supervised machine learning approaches to predict infant mortality: a case study

UNDER PEER REVIEW

of the 2014 Kenya Demographic and Health Survey. [online] Erepository.uonbi.ac.ke. Available at: <http://erepository.uonbi.ac.ke/handle/11295/155900> [Accessed 11 May 2022].

8. Marques Ramos, A. P., Prado Osco, L., Elis Garcia Furuya, D., NunesGonçalves, W., Cordeiro Santana, D., Pereira RibeiroTeodoro, L., Antonio da Silva Junior, C., Fernando Capristo-Silva, G., Li, J., Henrique RojoBaio, F., Marcato Junior, J., Eduardo Teodoro, P., &Pistori, H. (2020). A random forest ranking approach to predict yield in maize with UAV-based vegetation spectral indices. *Computers and Electronics in Agriculture*, *178*, 105791. https://doi.org/10.1016/j.compag.2020.105791
9. Santin, M., Ranieri, A., Hauser, M., Miras-Moreno, B., Rocchetti, G., &Lucini, L. et al. (2021). The outer influences the inner: Postharvest UV-B irradiation modulates peach flesh metabolome although shielded by the skin. *Food Chemistry*, *338*, 127782. doi: 10.1016/j.foodchem.2020.127782
10. Kumar, Y. J., Spandana, V., Vaishnavi, V. S., Neha, K., & Devi, V. G. R. R. (2020). Supervised machine learning approach for crop yield prediction in agriculture sector. *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. https://doi.org/10.1109/icces48766.2020.9137868
11. Ullah, Z., khan, M., RazaNaqvi, S., Farooq, W., Yang, H., Wang, S., &Vo, D.-V. N. (2021). A comparative study of machine learning methods for bio-oil yield prediction – a genetic algorithm-based features selection. *Bioresource Technology*, *335*, 125292. https://doi.org/10.1016/j.biortech.2021.125292

APPENDIX

DEFINITIONS, ACRONYMS, ABBREVIATIONS

AI Artificial Intelligence

DT Decision Tree

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

ML Machine Learning

MSE Mean Squared Error

UNDER PEER REVIEW

NACOSTI National Commission for Science, Technology, and Innovation

R Programming Language named R

RF Random Forests

RFC Random Forest Classifier

RMSE Root Mean Squared Error

UNDER PEER REVIEW