Original Research Article

APPLICATION OFXGBOOST REGRESSION IN MAIZE YIELD PREDICTION

ABSTRACT

Artificial Intelligence (AI) is the human-like intelligence imbued in machines so that they can perform tasks that normally require human intelligence. Machine learning is an AI technique which carries on the concepts of predictive analytics with one important distinction: the AI system can make assumptions, test hypotheses, and learn independently. XGBoost, Extreme gradient boosting, is a popular machine-learning library for regression tasks. It implements the gradient-boosting decision tree algorithm, which combines several feeble decision trees to produce a robust predictive model. In Boosted Trees, boosting is the process of transforming poor learners into strong learners. It is an ensemble method; a weak learner is a classifier with a low correlation with classification, whereas a strong learner has a high correlation. Maize is a staple food in Kenya and having it in sufficient amounts in the country assures the farmers' food security and economic stability. Crop yield measures the seeds or grains produced by a particular plot of land. Typically, it is expressed in kilograms per hectare, bushels per acre, or sacks per acre. This study predicted maize yield in Uasin Gishu, a county in Kenya, using XGBOOST regression algorithm of machine learning. The regression model used the mixed-methods research design, the survey employed well-structured questionnaires comprising of quantitative and qualitative variables, directly administered to selected representative farmers from 30 clustered wards.The questionnaire comprised 30 variables related to maize production from 900 randomly selected maize farmers distributed across 30 wards. XGBOOST machine learning ~~Regression~~regression model was fitted, and it could predict maize yield and identify the top features or variables that affect maize yield. The model was evaluated using regression metrics ~~Root Mean Squared error-RMSE=0.4563, Mean Squared Error-MSE =0.2082, and Mean Absolute Error-MAE = 0.3532~~

Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Mean Absolute Error (MAE), which values were 0.4563, 0.2082, and 0.3532, respectively. This algorithm was recommended for maize yield prediction.

Key Words:Artificial Intelligence, Machine Learning, Algorithm, XGBOOST, Boosted, Prediction, Ensemble Learning

1.INTRODUCTION

Artificial Intelligence (AI) is the human-like intelligence imbued in machines so that they can perform tasks that normally require human intelligence. Therefore, AI systems can think like humans, behave like humans, reason, and act rationally. AI systems are not only more potent and practical than conventional computers, but they also address complex human issues that are emerging. Mvurya discusses the degree to which AI has been implemented in the specific sectors associated with the Kenyan Big Four Agenda. The contribution of this document is a framework for the Kenyan government's use of AI's world-changing technologies to advance its development agenda. (Mvurya, 2020)

Machine learning, an AI technique, carries on the concepts of predictive analytics with one important distinction: the AI system can make assumptions, test hypotheses, and learn independently. (Rutty 2021, June 19).In Boosted Trees, boosting is the process of transforming poor learners into strong learners. It is an ensemble method; a weak learner is a classifier with a low correlation with classification, whereas a strong learner has a high correlation.XGBoost, Extreme gradient boosting, is a popular machine-learning library for regression tasks. It implements the gradient-boosting decision tree algorithm, which combines several feeble decision trees to produce a robust predictive model. XGBoost is classified as a Boosting technique within Ensemble Learning. Ensemble learning is a compilation of predictors, which are multiple models that provide more accurate predictions. By adding weights to succeeding models, the Boosting technique attempts to correct the errors made by earlier models. The algorithm begins with a basic model, such as a decision tree, and then iteratively enhances the model by adding additional trees to the ensemble. During each iteration, the algorithm calculates the error

between the predicted and actual values for the training data. It then uses this error to train a new tree that will help reduce the error in the next iteration. The process is repeated until the error is minimized, or a maximum number of iterations is reached. Additionally, XGBoost uses a technique called gradient boosting to reduce the bias and variance of the model.

In Kenya, XGBoost regression has been used in various industries, including finance, healthcare, and agriculture. For example, it has been used to predict crop yields based on weather patterns, market demand, and other factors. It can also be used in healthcare to predict the risk of diseases based on patient data.Overall, XGBoost regression is a robust machine learning algorithm that can be used in various industries in Kenya to make accurate predictions and improve decision-making.

Crop yield measures the seeds or grains produced by a particular plot of land. Typically, it is expressed in kilograms per hectare, bushels per acre, or sacks per acre. An indicator such as the average crop yield per acre evaluates a farmer's agricultural production on a specific field during a given time period. It is the most important indicator of a farmer's performance, as it represents the result of all the time and resources spent cultivating vegetation in their fields.The objective of the yield forecast is to provide an accurate, scientific, reliable, and independent forecast of crop yield as early as feasible during the crop-growing season, taking into account the impact of weather and climate.

In addition, the Kenya Maize Development Program-KMDP reports that the average Kenyan consumes 98 kilograms of maize per year. Maize is the staple food in Kenya. In addition, maize prices in Kenya are among the highest in sub-Saharan Africa, and the poorest 25% of the population uses 28% of their income on the crop.The relationship between maize yield and harvest area is positive, as demonstrated by Epule in 2022. According to him, maize yields and harvest area in Africa have increased by 71.35 percent and 60.12 percent, respectively. (Epule, Chehbouni&Dhiba, 2022).

In 2021, a study was done to investigate relationships between plant growth and development and apply maize plant growth models. It compared sigmoid, light GBM, and XGBOOST plant models on a dataset of the Leibniz Institute. ~~AIC~~ Akaike's Information (AIC), ~~BIC~~ Bayesian Information Criterion (BIC), RMSE, and R Squared metrics were utilized to evaluate model performance. (Sheth, D. (2021). Given the factors influencing the prediction worth of houses, such as neighborhood, location, and available amenities, a study was done to estimate house prices in 2021. It had been observed that price fluctuations were observed due to these factors.(Et.al. & J. A. 2021). In 2022, Wang et al., developed a method for predicting mortality using XGBoost and Logistic regression algorithms on a dataset of three hundred and sixty-eight patients at a train test ratio of 7:3. The algorithm proved beneficial to physicians in patients' evaluation of TB at high risk. (Wang., *et al,* 2022).

The fact that maize is the primary cereal produce in Africa and Kenya and a significant staple crop demonstrates the significance of maize. To ensure food security for a swiftly expanding population in the face of climate variability, numerous studies on maize, ranging from climate's effect on maize to yield predictions, have been conducted. Numerous researchers from around the world have used machine learning to forecast maize yield and have demonstrated its reliability.

This study builds on these results and contributes to the foundations of the application of machine learning in maize yield prediction in Uasin Gishu County using a myriad of features or variables. Moreover, the significance of crop modeling as a decision tool for farmers and other agricultural decision-makers to increase production efficiency has increased in recent years. The research predicted maize yield in Uasin Gishu County using XGBOOST regression algorithm.

2. MATERIALS AND METHODS

2.1 Study Area

The study area was in Uasin Gishu Countyin the North Rift region of Kenya. Uasin Gishu 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 Uasin Gishu 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 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).

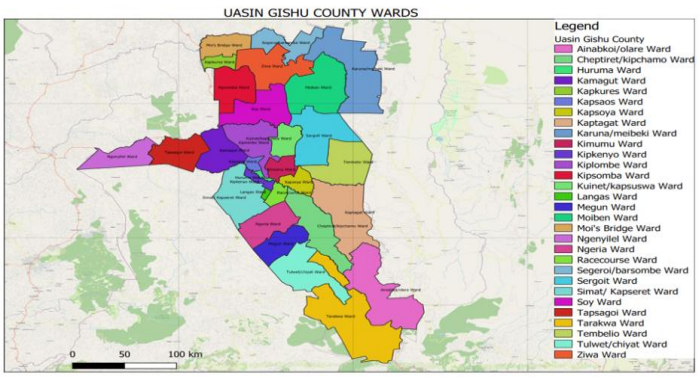

Figure 1: Map of Uasin Gishu County Displaying the 30 Wards.

[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

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

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%.

Comment [MF1]: Please give reasons.

2.3 Data Analysis

All the questionnaires were collected and labeled with unique identification numbers before data entry. Database creation, coding, and entry were done on Epi Data software.

3. RESULTS AND DISCUSSION

In XGBoost regression, the algorithm tries to fit a model to the training data by minimizing the error between the actual and predicted values. The algorithm begins with a basic model, such as a decision tree, and then iteratively enhances the model by adding additional trees to the ensemble.

xgbc = xgboost(data = xgb_train, max.depth = 1, nrounds = 20).

Table 3: Output of XGBOOST Basic Model Rmse in Train Data

Nround	Train-rmse	Nround	Train-rmse
1	0.782695	11	0.468363
2	0.645654	12	0.467327
3	0.565455	13	0.466418
4	0.521087	14	0.465588
5	0.497294	15	0.464838
6	0.484557	16	0.464137
7	0.477629	17	0.463468

8	0.473679	18	0.462782
9	0.471233	19	0.462090
10	0.469580	20	0.461443

Table 4: Feature Importance as per XGBOOST Regression.

Feature	Gain	Cover	Frequency
x12	0.1928	0.1	1
x24	0.1731	0.1	0.1
a	0.166	0.2	0.2
f	0.0753	0.1	0.1
x15	0.0743	0.0667	0.0667
x7	0.0657	0.0667	0.0667
d	0.0412	0.0667	0.0667
c	0.0317	0.0667	0.0667
x18	0.0298	0.0333	0.0333
x20	0.0269	0.0333	0.0333
e	0.0254	0.0333	0.0333
x19	0.0248	0.0333	0.0333
x17	0.0195	0.0333	0.0333
x22	0.0194	0.0333	0.0333
h	0.0183	0.0333	0.0333
g	0.0156	0.0333	0.0333

It indicated the relevant significance of each variable towards the prediction of maize yield. These were the most influential variables in the maize yield prediction model.

The plot of Feature Importance is displayed in Figure 2.

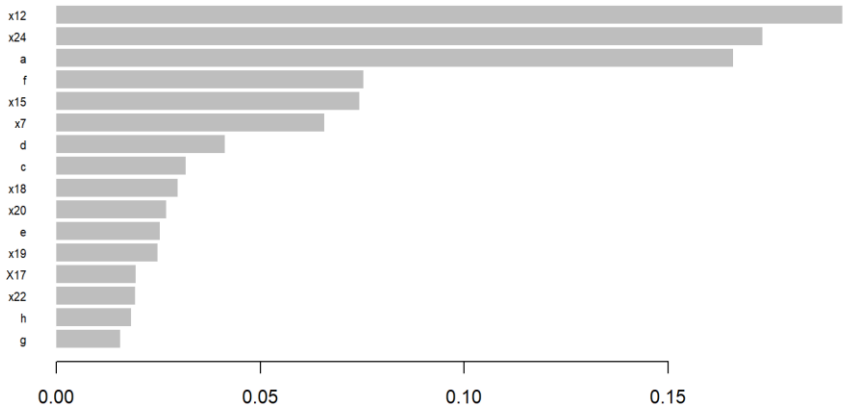

Figure 2: Feature Importance Plot of XGBOOST Regression.

A high feature importance score indicates that the feature is highly predictive of the target variable and should be considered important when making predictions. Conversely, a low feature importance score indicates that the feature is less important in predicting the target variable.

Defining the XGBOOST parameters.The model parameters used were max. depth = 1, eta = 0.3. nthread = 4 (was set according to the number of physical CPU cores available on the machine used, which was 4 and nrounds = 30. The max. depth controls the maximum depth of each individual tree in the boosting process. A decision tree is a fundamental component of XGBoost's boosting process, and It is constructed by recursively breaking down the data into smaller and smaller subsets based on the input feature values. The max. depth parameter limits the number of such partitions that can be made, and it is one way to control the complexity of the tree and prevent overfitting. Its low value makes the tree will simpler and have fewer splits, making it more interpretable and less prone to overfitting, while if max. depth is set to a high value; the tree can become very complex, with many splits and leaves, leading to overfitting and poor generalization performance.

The study used nrounds = 30. This parameter determines the total number of trees to be grown, which must be large enough to ensure accurate predictions while remaining small enough to prevent overfitting. Through cross-validation, a suitable value can be determined.

Feature contributions/importance was the relative importance of each feature in predicting the target variable.

Model performance was evaluated based MSE, ~~Mean squared error (MSE),~~MAE, ~~Mean absolute error (MAE), Root mean squared error (RMSE)~~RMSE, and Mean Absolute Percentage ~~error~~Error, (MAPE).

Table 5: Model performance based on MSE, MAE, RMSE and MAPE

Metric	MSE	MAE	RMSE	MAPE
Index	0.2082~~372~~	0.3532~~667~~	0.4563~~302~~	25.2700~~1~~

Comment [MF2]: Why are MAPE and its values not included in the Abstract Section? We recommend that MAPE and its values be added to the Abstract Section.

Figure 3 is the plot of original versus predicted, where red is the actual while blue is the predicted plot.

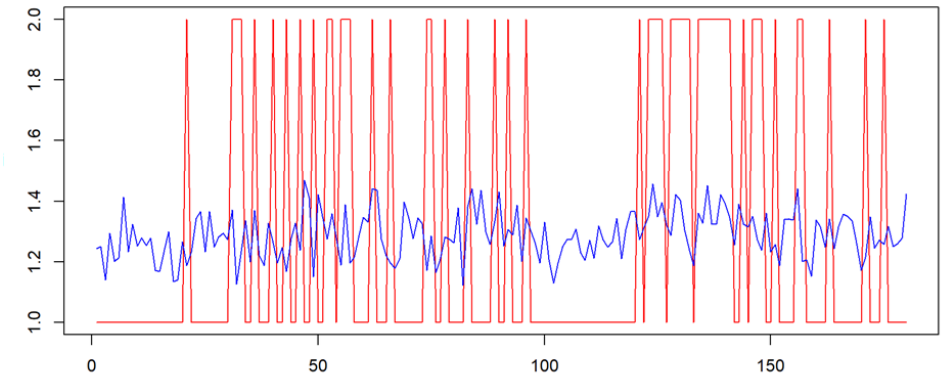

Figure 3: Plot of Original versus Predicted In XGBOOST

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 Uasin Gishu county. Further, the final data set was able to predict maize yield with a good performance of MSE = 0.2082372, MAE = 0.35321667, RMSE = 0.4563302 and MAPE = 25.27001.

Conclusively, this study has presented XGBOOST to have performed well in maize yield regression model in Uasin Gishu county. It may not necessarily mean that it is sufficient in maize yield prediction. Future research recommends a comparative study on GBDT's advanced algorithms LightGBM – (Light gradient boosting machine) and Catboost-(Category boosting) and artificial neural network models in order to analyze XGBoost profoundly and find the most suitable model algorithm for maize yield prediction.

Advanced machine learning techniques, such as deep learning, have demonstrated promising potential for managing massive data sets. However, their use has high-cost implications. Policymakers should give funding for research studies and investment in new technology priority to support agricultural research to achieve food security and economic stability for farmers

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. Abrougui, K., Gabsi, K., Mercatoris, B., Khemis, C., Amami, R., &Chehaibi, S. (2019). Prediction of organic potato yield using tillage systems and soil properties by Artificial Neural Network (ANN) and multiple linear regressions (MLR). *Soil and Tillage Research*, *190*, 202–208. https://doi.org/10.1016/j.still.2019.01.011
2. Basso, B., Cammarano, D., &Carfagna, E. (2013). Review of Crop Yield Forecasting Methods and Early Warning Systems.

3. Et. al., J. A. (2021). Prediction of house price using XGBoost regression algorithm. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(2). https://doi.org/10.17762/turcomat.v12i2.1870
4. Epule, E. T., Peng, C., Lepage, L., & Chen, Z. (2011). Forest loss triggers in Cameroon: A quantitative assessment using multiple linear regression approach. *Journal of Geography and Geology*, *3*(1). https://doi.org/10.5539/jgg.v3n1p30
5. Epule, T. E., Chehbouni, A., & Dhiba, D. (2022). Recent Patterns in Maize Yield and Harvest Area across Africa. *Agronomy*, *12*(2), 374. https://doi.org/10.3390/agronomy12020374
6. Mvurya, M. (2020). The Extent and Use of Artificial Intelligence to Achieve the Big Four Agenda in Kenya. *Multidisciplinary Journal Of Technical University Of Mombasa*, *1*(1), 1-7. doi: 10.48039/mjtum.v1i1.9
7. Pardey, P.G., Beddow, J.M., Hurley, T.M., Beatty, T.K.M. and Eidman, V.R. (2014), A Bounds Analysis of World Food Futures: Global Agriculture Through to 2050. Aust J Agric Resour Econ, 58: 571-589. https://doi.org/10.1111/1467-8489.12072
8. Quinto, B., &Zhang,Xiaokun. (2021). Chapter 1. In *Ji Yu spark de xiayi Dai Ji Qi Xue XI: Cover xgboost,LightGBM,SparkNLP,distributed deep learning with keras,and more = next-generation machine learning with Spark*. essay, Ji xie gong ye chu ban she.
9. Rutty, M. (2021, June 19). *Predictive analytics vs. AI: Why the difference matters*. TechBeacon. Retrieved November 2, 2022, from https://techbeacon.com/enterprise-it/predictive-analytics-vs-ai-why-difference-matters
10. Tilman, D., Balzer, C., Hill, J., &Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(50), 20260–20264. https://doi.org/10.1073/pnas.1116437108
11. Sheth, D. (2021). Plant growth and LAI estimation using quantized embedded regression models for high throughput phenotyping. https://doi.org/10.21203/rs.3.rs-1060088/v1

12. Wang, R., Wang, L., Zhang, J., He, M., & Xu, J. (2022). XGBoost machine learning algorism performed better than regression models in predicting mortality of moderate-to-severe traumatic brain injury. *World Neurosurgery*, *163*. https://doi.org/10.1016/j.wneu.2022.04.044

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
NACOSTI	National Commission for Science, Technology and Innovation
QGIS	Quantum Geographic Information System
R	Programming Language named R
RF	Random Forests
RFC	Random Forest Classifier
RMSE	Root Mean Squared Error
WI	Wilmott's Index
XGBOOST	Extreme Gradient Boosted Regression