Original Research Article

AN APPLICATION OF K-NEAREST-NEIGHBOR REGRESSION IN MAIZE YIELD PREDICTION

ABSTRACT

Predictive analytics utilizes historical data and knowledge to predict future outcomes and provides a method for evaluating the accuracy and reliability of these forecasts. Artificial intelligence is a tool of predictive analytics. AI trains computers to learn human behaviors like learning, judgment, and decision-making while simulating intelligent human behavior using computers and has received a lot of attention in almost all areas of research. Machine learning is a branch of Artificial Intelligence that has been used to solve classification and regression problems. Machine learning advancements have aided in boosting agricultural gains. Yield prediction is one of the agricultural sectors that has embraced machine learning. K Nearest Neighbor (KNN) Regression is a regression algorithm used in machine learning for prediction tasks. KNN Regression is like KNN Classification, except that KNN Regression predicts a constant output value for a given input instead of predicting a class label. The basic idea behind KNN Regression is to find the K nearest neighbors to a given input data point based on a distance metric and then use the average (or weighted average) of the output values of these K neighbors as the predicted output for the input data point. The distance metric used in KNN Regression can vary depending on the data type being analyzed, but common distance metrics include Euclidean distance, Manhattan distance, and Minkowski distance. This paper presents the application of KNN regression in maize yield prediction in Uasin Gishu county, in north rift region of Kenya. Questionnaires were distributed to 900 randomly selected maize farmers across the thirty wards to obtain primary data. With a Train Test split ration of 80:20, KNN regression algorithm was able to predict maize yield and its prediction performance was evaluated using Root Mean Squared error-RMSE=0.4948, Mean Squared Error-MSE =0.2803, Mean Absolute Error-MAE = 0.4591 and Mean Absolute Percentage Error-MAPE = 36.17. According to the study findings, the algorithm was able to predict maize yield in the maize producing county.

*Key Words: Algorithm**,** Artificial Intelligence, Machine Learning, Algorithm, K Nearest neighbor, Regression, Classification, Prediction,*

1.INTRODUCTION

Predictive analytics mines the information from existing data to establish patterns to forecast future conclusions and trends. It predicts "what might happen in the future." It utilizes historical data and knowledge to predict future outcomes and provides a method for evaluating the accuracy and reliability of these forecasts. Predictive analytics is a mathematical process that uses historical data to make future pr Artificial Intelligence is a subset of Predictive Analytics. Artificial Intelligence trains computers to learn human behaviors like learning, judgment, and decision-making while simulating intelligent human behavior using computers. It has received a

lot of attention in almost all areas of research. Statistical models are typically used to predict agricultural production, which is time-consuming and labor-intensive. (Bali & Singla, 2021).

AI is completely autonomous, which is the most significant distinction between it and predictive analytics. In contrast, predictive analytics requires human interaction to query data, recognize trends, and test hypotheses. AI has much more scope and more applicability than predictive analytics alone. In addition to perpetually expanding multivariable algorithms, artificial intelligence also employs the strict forecasting model of predictive analytics. (Rutty, M. (2021, June 19).

Machine learning is a branch of Artificial Intelligence that has been to solve classification and regression problems. Machine learning steps involve Data collection, Data preparation which is the process of organizing and preparing data for machine learning training. Data exploration is used to comprehend the nature, format, and quality of the data that will be utilized. Data Wrangling is the process of cleaning and transforming unprocessed data into a format that can be utilized. Data Analysis comes before data cleansing. Data analytics entails the selection of analytical techniques, the construction of models, and the evaluation of the result. The subsequent phase is to train the resulting model to enhance its performance and produce a better solution to the problem and the model is evaluated using the test data. Model deployment is the final stage of the machine learning lifecycle.K Nearest Neighbor (KNN) Regression is a regression algorithm used in machine learning for prediction. KNN Regression is like KNN Classification, except that KNN Regression predicts a constant output value for a given input instead of predicting a class label. The basic idea behind KNN Regression is to find the K nearest neighbors to a given input data point based on a distance metric and then use the average (or weighted average) of the output values of these K neighbors as the predicted output for the input data point.

The distance metric used in KNN Regression can vary depending on the data type being analyzed, but common distance metrics include Euclidean distance, Manhattan distance, and Minkowski distance. Regression is a simple and intuitive algorithm, but it can be computationally expensive, especially when working with large datasets. It is also sensitive to the choice of K and the distance metric used. Therefore, it is important to tune these hyperparameters carefully for optimal performance.

A study conducted a comparative evaluation of the random forest and K-nearest neighbors (KNN) algorithms for the purpose of classifying gulley erosion in Iran. A total of 242 locations were examined to identify and extract twelve characteristics that are associated with gulley erosion. The algorithms successfully discerned significant parameters associated with gulley erosion, exhibiting commendable predictive capabilities of 87 % and 80.9 % respectively. (Avand *et al.,* 2019).

The evaluation of consumer demand prediction performance in 2019 involved the utilization of the K-Nearest Neighbors (KNN) algorithm using industrial time series data from the M3-Competition. The study included a comparison of various k parameter configurations and model selection techniques. The models of locally constant mean and locally linear ridge regression demonstrated optimal balance between forecast accuracy and computation durations, resulting in good prediction performance and little forecasting error. The research findings indicate that K-nearest neighbors (KNN) can be considered a viable substitute for demand forecasting in an industrial setting. This approach demonstrates a notable level of prediction accuracy while also requiring minimal calculation time. (Kück, M., & Freitag, M. (2020).

In 2020, Kohli, Godwin, and Urolagin researched sales prediction, given that it could lead businesses to identify potential risks and make accurate predictions. According to the authors, sales prediction when data is lacking or is marred with missing values and outliers becomes challenging and is graded more as a regression problem than a time series. They analyzed Rossmann sales data using machine learning predictive models to identify distinct sales patterns with risk variables. Specifically, they evaluated Linear regression and k-nearest neighbor regression models. The model performances were importantly evaluated using statistical methods like RMSE and MAPE. The outcome of the study allowed more suitable sales classifiers for prediction purposes. (Kohli, Godwin & Urolagin, 2020).

In 2020, machine learning egression algorithms were evaluated for the yield prediction of Mustard crop. Using soil samples data from Jammu region was evaluated on Random Forest, Naïve Bayes, Multinomial Logistic Regression, K-Nearest Neighbor and Artificial Neural Network algorithms. KNN and ANN outperformed the other algorithms in yield prediction. (Pandith et al.,2020).

In light of the pivotal role that agriculture plays in the Indian economy, a research investigation was conducted in the year 2021 to explore the application of machine learning algorithms for the purpose of predicting soil fertility. This prediction was based on an analysis of several soil characteristics. The decision tree and K-nearest neighbors (KNN) algorithms demonstrated superior predictive ability. (R, J., & M, S. D. (2021).

Due to the rapid change in climate patterns and the high number of livestock deaths caused by protracted droughts and unpredictable rainfall, the community leadership of Kumahumato area had determined that irrigation-assisted agriculture should be prioritized. They utilized eight algorithms, including Neural Networks, Nave Bayes, Support Vector Machines, Logistic Regression, Decision Trees, K-Nearest Neighbor, and Random forests. During in-depth data analysis, the final three algorithms mentioned in this article demonstrated the highest levels of accuracy, ranging from 95 to 100 percent. (Duale, Munene & Njogu,2021).

In 2022, there was rising concern, globally and regionally in African agricultural production is of increasing concern to the world's leading nutrition-focused international organizations. In addition, farmers and agricultural decision-makers require cutting-edge instruments to help them make rapid decisions that affect the quality of agricultural yields. Globally, climate change has been a significant phenomenon in recent decades. The character of agricultural production has been impacted by climate change. The advent of big data technology has led to the development of new, highly effective analytical tools, such as machine learning. The research was undertaken to predict a system based on machine learning to determine the yield of rice, maize, cassava, seed cotton, yams, and bananas, throughout the year in West African countries. It analyzed climatic data, weather data, agricultural yields, and chemical data to assist decision-makers and producers in predicting their country's annual crop yields. Utilizing a decision tree, multivariate logistic regression, and k-nearest neighbor models, they created their system. When using three machine learning models, both models produced favorable results. In addition, they found that the decision tree model performs the best, followed by the K-Nearest Neighbor model and logistic regression, and that the prediction results of the decision tree model and the K-Nearest Neighbor model correlate with the expected data, demonstrating the model's effectiveness. (Cedric et al., 2022).

Oil palm yield prediction was required for field management, import and export, and global food security. Oil palm yield forecast using meteorological and soil data using k-nearest neighbor (KNN) algorithm. This research made use of 35 years' worth of Pahang state, Malaysia's yield, soil, and weather records. The findings indicated that machine learning models may produce accurate forecasts with a sizable amount of data from various sources. As a crucial component of precision agriculture, machine learning is a fantastic potential tool for predicting oil palm productivity. (Khan et al,.2022).

In the field of medical research, a recent study was conducted by Zainab to compare the performance of the K-Nearest Neighbors (KNN) and decision tree machine learning algorithms in predicting cardiac disease. The study revealed that the K-nearest neighbors (KNN) algorithm demonstrated favorable performance in comparison to other methods. (lone, zainab & Hod.(2023)

In Kenya, KNN Regression can be applied to various fields such as agriculture, healthcare, finance, and education. For example, KNN Regression can predict crop yields in agriculture based on weather data, soil conditions, and other relevant variables. In healthcare, it can be used to predict patient outcomes according to the patient's medical history and other health-related factors. To apply KNN Regression in Kenya, one would first need to gather relevant data and preprocess it to ensure it is suitable for the algorithm.

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.

These findings are expanded upon in this study, which also strengthens the theoretical underpinnings of the use of machine learning to maize yield prediction in Uasin Gishu County using a multitude of features or variables. The research predicted maize yield in Uasin Gishu County using K Nearest Neighbor Regression algorithm.

2. Methodology:

2.1 Study area

The study area was in Uasin Gishu County. It is in the North Rift region of Kenya. Uasin Gishu County covers an area of 3346 km^2 (2995 km^2 arable, 333 km^2 non-arable, 23km^2 water masses, and 196 km^2 urban area). The map of the study area is attached in Appendix 1.

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).

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	Sub-county	Sub county location 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.

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.

2.4 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

Packages used in KNN regression were library(caret), library (MASS) and library(ggplot2).

KNN regression output was

```
> knn_reg <- train(Y.avg ~ ., data = train, method = "knn",
+                  trControl = trainControl(method = "cv", number = 20),
+                  tuneGrid = expand.grid(k = 1:100))
```

Permutation feature importance was computed using the R package "caret."loess r-squared variable importance and only 20 most important variables shown (out of 30).

A plot of training error versus validation error for different values of K is displayed in figure 1.

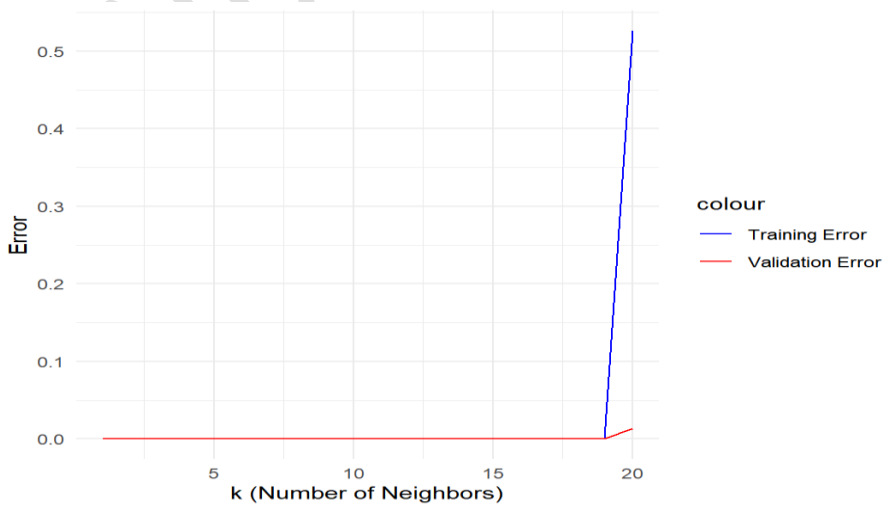

Figure 1: Plot of training error versus validation error for different values of K.

In figure 1, X-axis represents different values of the hyperparameter "k." In KNN, "k" represents the number of nearest neighbors to consider when making predictions while the y-axis represent the mean squared error rate. The training error rate (blue line) in the plot indicates that the value of "k" declines, indicating a reduction in the number of neighbors examined, the training error tends to decrease. This can be attributed to the increased complexity of the model, which allows for a closer fit to the training data. The validation error (red line) obtained from the test data, an evaluation metric serves to approximate the extent to which your model will exhibit generalizability when applied to novel, unobserved data. The point where the two lines meet is the desired point that represents the optimal compromise between bias and variation. The concept of generalization refers to the ability of a model to perform effectively on unseen data, striking a balance between underfitting and overfitting.

Table 3: Table of Display of Important Variables as per KNN Regression.

Variable	Importance
Point of sale	0.03534
Time of sale	0.01665
Full time farmer	0.00925
Household size	0.00732
Maize farming practices	0.00678
Source of labor	0.00579
Marital status of farmer	0.00509
Method of tillage	0.00478
Decision maker	0.004412
Cropping system	0.004285
Stage of sale	0.004016
Machinery ownership	0.0036605
Fertilizer type	0.0035020

Soil testing	0.0034774
Land size allocated to maize	0.0024184
Fertilizer use	0.0023329
Land ownership	0.0013227
Maize yield factors	0.0013162
Age of the farmer	0.0012681
Maize production process	0.0008543

Table 4 below displays the prediction performance evaluation of KNN Regression.

Table 4. Model Prediction Performance Evaluation

Metric	MSE	MAE	RMSE	MAPE
Index	0.2803	0.4592	0.4948	36.17

4.CONCLUSIONS

The study finds that machine learning has a lot of potential as a predictive tool in smart agriculture. KNN regression was able to predict maize yield from the presented thirty variables revolving around characteristics of maize famers to general factors affecting maize yield in Uasin Gishu county.

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. Finally, the everchanging developments in the tech space indicate a need to fast implement machine learning techniques to make Agriculture more sustainable for future generations in

Kenya. The findings and insights presented here serve as a steppingstone for further inquiry and can inform decision-making processes in various domains.

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.

DEFINITIONS, ACRONYMS, ABBREVIATIONS

AI Artificial Intelligence

KNN K Nearest Neighbor

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

RMSE Root Mean Squared Error

REFERENCES

1. Avand, M., Janizadeh, S., Naghibi, S. A., Pourghasemi, H. R., Bozchaloei, S. K., & Blaschke, T. B. (2019). A comparative assessment of random forest and K-nearest neighbor classifiers for gully erosion susceptibility mapping. *Water*, *11*(10), 2076. https://doi.org/10.3390/w11102076
2. Bali, N., & Singla, A. (2021). Deep learning-based wheat crop yield prediction model in Punjab region of North India. *Applied Artificial Intelligence*, *35*(15), 1304–1328. https://doi.org/10.1080/08839514.2021.1976091

3. Cedric, L. S., Adoni, W. Y., Aworka, R., Zoueu, J. T., Mutombo, F. K., Krichen, M., & Kimpolo, C. L. (2022). Crops yield prediction based on machine learning models: Case of West African countries. *Smart Agricultural Technology*, *2*, 100049. https://doi.org/10.1016/j.atech.2022.100049

4. Duale, M., Munene, E., & Njogu, M. (2021). Historical and political contestations in the Dadaab refugee camps and north-eastern Kenya. *Borderless Higher Education for Refugees*. https://doi.org/10.5040/9781350151277.ch-001

5. Ione, zainab F., & Hod, saurabh. (2023). *Heart Disease Prediction System Using Machine Learning Algorithms (KNN and Decision Tree Algorithm)*. https://doi.org/10.21203/rs.3.rs-3222970/v1)

6. Khan, N., Kamaruddin, M. A., Sheikh, U. U., Yusup, Y., & Bakht, M. P. (2022). Environment-based oil palm yield prediction using K-nearest neighbour regression. *2022 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. https://doi.org/10.1109/iicaiet55139.2022.9936752

7. Kohli, S., Godwin, G. T., & Urolagin, S. (2020). Sales prediction using linear and KNN regression. *Algorithms for Intelligent Systems*, 321–329. https://doi.org/10.1007/978-981-15-5243-4_29

8. Kück, M., & Freitag, M. (2020b). Forecasting of customer demands for production planning by local k-nearest neighbor models. *International Journal of Production Economics*, *231*, 107837. https://doi.org/10.1016/j.ijpe.2020.107837

9. Pandith, V., Kour, H., Singh, S., Manhas, J., & Sharma, V. (2020). Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. *Journal of Scientific Research*, *64*(02), 394–398. https://doi.org/10.37398/jsr.2020.640254

10. R, J., & M, S. D. (2021a). Predictive model construction for prediction of soil fertility using decision tree machine learning algorithm. *Kongunadu Research Journal*, *8*(1), 30–35. https://doi.org/10.26524/krj.2021.5

11. 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

APPENDIX

1. Map of Uasin Gishu county displaying the 30 wards.
   [Source: Author]