DECISION TREE AS A RESOURCE FOR ASSESSING MILLET PRODUCTION IN SOME SELECTED WEST AFRICAN COUNTRIES

Ilesanmi A. Ajibode & Nurudeen O. Alabi
Department of Matheamatics/Statistics
The Federal Polytechnic, Ilaro
Ogun State Nigeria
ilesanmi.ajibode@federalpolyilaro.edu.ng

Abstract

The production of food plays a vital role in sustaining human existence. However, the availability of adequate food reserves poses a significant concern for several nations in West Africa. However, governments in the area are making concerted efforts to ensure food security, recognizing its paramount importance in the contemporary global context. Millet has a diverse range of nutritional benefits that contribute to maintaining of stable blood sugar levels, facilitating of digestion, and safeguarding of cardiovascular health. Therefore, this research used regression tree analysis to investigate the determinants influencing millet production in certain West African nations. Monthly data were systematically gathered during a period spanning from 1993 to 2022 for the nations of Niger, Nigeria, and Ghana. The data included many factors related to millet production, including the quantity of millet produced in tonnes, the rural population of each country, the land area dedicated to millet cultivation, as well as rainfall and temperature measurements. The use of a regression tree analysis yielded findings that indicate the rural population and the amount of land dedicated to millet cultivation are the primary variables influencing millet production in the nations under investigation. Consequently, the research concludes that allocating more land for millet production and involving the rural community in its cultivation will be beneficial for enhancing millet output.

keywords: Millet, Long term, short term, Data Mining, Knowledge, Sufficient

Introduction

Agriculture in Africa has a rich historical past, with a plethora of traditional agricultural practises that have maintained populations for long periods of time. Nonetheless, adoption of modern agricultural practises is still restricted in various locations throughout the continent, providing substantial hurdles for African farmers. A major source of worry is a lack of access to modern agricultural methods and equipment, which results in lower levels of production and output. Furthermore, climate change has had a negative impact on agricultural practises throughout the continent, manifesting as droughts, floods, and a variety of other extreme weather occurrences. Despite these obstacles, there is a rising trend in Africa towards the adoption of sustainable agriculture practises, which are distinguished by an emphasis on soil health restoration, biodiversity enhancement, and a decrease in dependence on pesticides and other hazardous chemicals. Numerous groups and projects are currently working to improve small-scale farmers' access to funding, training, and technology.

The agricultural sector is vital to numerous African nations because it employs a large number of people and plays an important role in providing food security and supporting economic growth. According to Obi and Obayori (2016), suitable policies and strategic investments in African agriculture have the potential to support its growth and make a substantial contribution to the continent's overall development.

The West African area has a significant agricultural history, defined by the intergenerational transmission of various ancient farming systems. Agriculture plays an important role in the region's economic framework, acting as a key source of employment opportunities and contributing significantly to both food security and economic development. Cassava, yams, maize, rice, sorghum, millet, and groundnuts are often cited as the principal crops grown in West Africa. Cattle, goats, and sheep are important agricultural commodities. However, the area has a number of agricultural challenges, including a lack of modern farming technology and equipment, inadequate infrastructure, and insufficient investment. Furthermore, it is clear that climate change is having a negative influence on the region's agricultural economy. This is shown by the increasing frequency of droughts and floods, which result in significant agricultural damage and lower yields (Anigbogu, Abosi, & Okoli, 2015). The agriculture industry is critical to West Africa's overall growth.

The importance of technical education in agricultural development stems from its ability to give farmers with the necessary skills and knowledge to boost production levels and profitability. Technical education is exemplified by vocational training, apprenticeships, and on-the-job training. Technical education in agricultural development may help farmers learn about the most recent farming technologies and approaches, as well as provide them with the skills needed to run contemporary equipment and machinery. Farmers may benefit from technical education that provides them with information about sustainable agriculture practises. Farmers may use this information to improve soil health, decrease their dependency on pesticides and other potentially dangerous chemicals, and increase biodiversity.

Furthermore, technical education may assist farmers in acquiring market access and increasing their revenue. Farmers may increase their profitability by adjusting their output to match market demand and consumer preferences by learning about market trends.

Governments and development groups may encourage agricultural technical education via a variety of approaches. These include allocating resources to the establishment of essential facilities that facilitate knowledge acquisition among agriculture students, implementing training programmes aimed at improving farmers' skills, providing subsidies or incentives to encourage farmer participation in training initiatives, and collaborating with private sector entities to design training programmes that cater to the specific needs of local farmers.

The regression tree, also known as decision tree regression, is a machine learning approach used for numerical or continuous value prediction. The decision tree in question is a specialised kind developed for regression tasks, as opposed to classification jobs where the purpose is to predict categorical labels.

The input data is separated into different subsets within the framework of a regression tree utilising the values of the input characteristics. The partitioning process's purpose is to decrease the variability of the target variable within each subset in order to create subsets with a high degree of homogeneity in regard to the target variable. Each core node in the aforementioned tree structure represents a choice made based on a certain feature, while each leaf node represents a projected numerical value.

The method of building a regression tree entails splitting the dataset into subgroups repeatedly by choosing the best feature and threshold for each division. The most advantageous division is often chosen based on a metric that quantifies the decrease in variance (or mean squared error) of the dependent variable after the division.

The tree-building process continues until a given requirement is reached. Several factors may be considered, such as the requirement to attain a certain maximum depth, the need to maintain a minimum number of samples per leaf, or the need to achieve a minimum decrease in variance. Once the tree structure has been established, predictions are generated by traversing the tree from the root node to the leaf node. The feature values linked with the input data point specify the data traversal. Returning the projected value supplied to the leaf node yields the final prediction.

Regression trees have a notable interpretability feature in that they may capture subtle correlations between data and the target variable. However, it is critical to recognise that when decision trees are allowed to grow too deep and complex, they may exhibit overfitting tendencies. Many solutions are often employed to reduce the issue of overfitting, such as pruning, restricting the depth of the tree, and employing ensemble methods such as Random Forest or Gradient Boosting Trees.

Millet's importance in West Africa extends beyond its nutritional value. Millet has tremendous cultural importance, as seen by the presence of traditional festivals and rituals centred on its production and use. Millet cultivation and consumption are often recognised as symbols of identity and cultural legacy in many nations. Millet production in West Africa has various obstacles, including a lack of modern agricultural equipment and gear, inadequate soil conditions, and the influence of climate change. Millet production in West Africa is critical to guaranteeing food security and provides a significant source of subsistence and cash for a large population in the region.

Teng et al. (2021) performed a research that proved the application of data mining approaches in the prediction of millet crop yields in Niger. The researchers developed a prediction model employing a mix of environmental and socioeconomic data, such as rainfall, temperature, soil fertility, and market price, using random forests as a technique. The model showed a significant correlation value of 0.80, showing that it predicted millet yields accurately. This discovery highlights the potential of data mining technology for increasing millet yield and solving food security challenges in the region.

In general, the use of data mining technologies in millet production has the potential to facilitate the identification of critical factors influencing yield outcomes, improve the precision of yield projections for future periods, and improve farmer and policymaker decision-making processes. Data mining has the ability to improve food security and promote sustainable agriculture in West Africa and other parts of the world.

According to Ayeomoni and Aladejana (2016), the agricultural sector is very important inside a nation since it provides food to the people and helps to economic growth via the creation of cash resources. The agricultural sector is critical in terms of poverty alleviation, economic growth, and general development. The competent management of the agriculture sector is a decisive element in a nation's strategic economic development success or failure. As a consequence, past research has shown a strong association between increased funding for the agricultural sector and positive outcomes (Omorokunwa & Obadiaru, 2016; Sertolu, Ugural, & Bekun, 2017).

Ewetan, Fakile, Urhie, and Oduntan (2017) investigated the ongoing association between agricultural production and economic development in Nigeria by analysing time series data from 1981 to 2014. The researchers employed a co-integration test and a vector error correction model (VECM) to find a significant relationship between agricultural production and Nigerian economic development. The observed result was consistent with the results of the Granger causality test, which demonstrated the existence of a causal link between agricultural production and economic progress.

Ramesh and Vardhan (2013) conducted research on crop yield prediction in a variety of agricultural industries. The use of data mining tools is one viable way for reducing this risk. Diepeveen and Armstrong (2008) emphasise the need of delivering crop-related data to farmers in order to enhance decision-making processes and increase output levels and overall profitability. While this phenomenon may offer a competitive advantage to specific crop species, it is critical to recognise that the information provided is broad and may not be universally applicable. Data mining tools can evaluate data in a variety of agricultural contexts, improving its quality and reliability. The task at hand entails recognising and evaluating essential elements relevant to agricultural cultivation, including geographical location, soil makeup, seasonal variations, nutrient requirements, grain yield and quality, planting and harvesting timing, and ability to withstand environmental stressors. This study employed data mining methods to help farmers choose the best combination of qualities for selecting plants that perform better. Various tactics were used in various geographic locations.

Murynin et al. (2013) conducted study on the link between prediction and forecast accuracy. The linear model is the most often used strategy for making predictions. The model is subsequently improved by including non-linear features, increasing prediction accuracy. The constant technological advancements in agricultural productivity, as well as geographical variances in crop yield, are among the expansions. The model's accuracy was assessed by taking into consideration the time gap between the prediction's creation and the harvest period. To increase agricultural production, farmers must have a thorough understanding of various factors, including soil type, the biotic components that influence soil conditions, and a strong understanding of accepted agricultural practises.

It is worth noting that, despite the possible impact of data-related issues, researchers are increasingly interested in the use of data mining skills in the area of agriculture. Furthermore, the importance of technical education and vocational training in agriculture must be recognised. The project's goal is to investigate the possible influence of agricultural students' data mining skills on future crop output.

The fundamental goal of this study is to discover the interrelationships between numerous criteria that contribute to the development of millet yield prediction models. These models are meant to be trustworthy and useful for predicting. If this trend is followed, millet output will steadily grow.

Temperature, rainfall, cultivable land area, and rural population serve as predictor factors for the nations under consideration, which are Nigeria, Niger, Burkina Faso, Senegal, and Ghana. The goal of this study was to use data mining approaches to millet production while utilising the observed predictive qualities.

Methodology

The dataset consists were obtained from three countries, namely Niger, Nigeria, and Ghana. Each nation has a sample size of n = 349 observations. The variables included in the dataset are millet production (mp), rural population (rp), temperature (te), rainfall (rf), and land area (la). The data used in this study was obtained from the World Bank database. The monthly data was partitioned into two datasets, with the training dataset comprising 65% of the total

data and the remaining 35% allocated to the other dataset. The first dataset was used for training the decision tree, whilst the subsequent dataset was utilised for assessing the efficacy of the fitted tree model. The methodology used in this study is influenced by a tree analogy, in which the terminal nodes are metaphorically compared to leaves, while the inner nodes are identified as the points at which the regressor space is divided. The branches serve as the connecting parts between the internal and terminal nodes. Our approach is influenced by the methodology proposed by Hastie et al (1996), which entails partitioning the regressor space X = [x_1, x_2, x_3,……,x_6] into j separate and non-overlapping box-shaped regions R_1, R_2,….. Here, x_1 represents rp, x_2 represents te, x_3 represents rf, and x_4 represents la. The objective of this study is to maximise the dissimilarity between the answer averages in each box of the R_j dataset. A recursive binary tree interconnects the processes for splitting that describe the areas. The prediction is then made by selecting an observation's area R_j and utilising the average of the training observations' response values in that region as the projected value. The response variable y, which is equal to the product of m and p, was represented by a constant c_j in each area.

$$f(x) = \sum_{j=1}^{J} c_m \mathrm{I}(x \in R_j) \quad (1)$$

The regions minimise the residual sum of squares, RSS, in equation 1.

$$\mathrm{RSS} = \sum_{j=1}^{J} \sum_{i=R_j} (y_i - \bar{y}_{R_j})^2 \quad (2)$$

$\bar{y}_{R_j}$ stands for average millet production (mp) for jth region taining values.

$$R_1 = \{X \mid x_j < u\} \text{ and } R_2 = \{X \mid x_j \geq u\} \quad (3)$$

In order to ultimately lower the RSS in equation 2. This means that we should make an effort to get the values of j and u that minimize equation 2.

$$\sum_{i:x_i \in R_1(j,u)} (y_i - \bar{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,u)} (y_i - \bar{y}_{R_2})^2 \quad (4)$$

where $\bar{y}_{R_1}$ is average response (**mp**) for training values in the $R_1(j, u)$. *Also,* $\bar{y}_{R_2}$ is average response for training values in the $R_2(j, u)$ region. This process was done repeatedly in the subsequent steps in order to minimizing the value of RSS in each step.

Boosting the fitted Millet production decision tree model

Decision trees' prediction accuracy is known to have downsides (higher variance), which may lead to instability and a lack of resilience to changes in the data. Furthermore, as seen by the MSE values, the pruned tree generated by the cost complexity pruning performed poorly on the test dataset. We employed boosting to improve the accuracy of the regression tree model's predictions. Boosting, a bootstrap aggregation procedure, may minimise the variance of a statistical learning strategy like our fitted regression tree. While increasing bias, averaging a set of data reduces variation. In this scenario, bootstrapping entails gradually generating G = 40,000 unique bootstrapped regression trees from a single training dataset. The residuals (ri) of previously formed trees are employed to direct future tree development. In other words, rather of employing the initial response values yi, the residuals of previously created trees were used to fit these trees. G = 40,000 subtrees was used to calculate

$$\hat{f}^1(x), \hat{f}^2(x)..... \hat{f}^{40,000}(x) \quad (5)$$

based on the following algorithm.

2.1.1 Algorithm on boosting for Millet production decision tree model

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all *i* in the training dataset
2. Fit *g* = 1, 2, 3,……,40,000, repeat:
    a. Fit a tree $\hat{f}^g$ with ***v*** splits to the training data (*x*, *r*)
    b. Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \delta \hat{f}^{g}(x) \quad (6)$$

c. Update the residuals,

$$r_i \leftarrow r_i \leftarrow \delta \hat{f}^{g}(x_i) \quad (7)$$

3. Output the boosted model,

$$\hat{f}(x) = \sum_{g=1}^{40,000} \delta \hat{f}^{g}(x) \quad (8)$$

Results

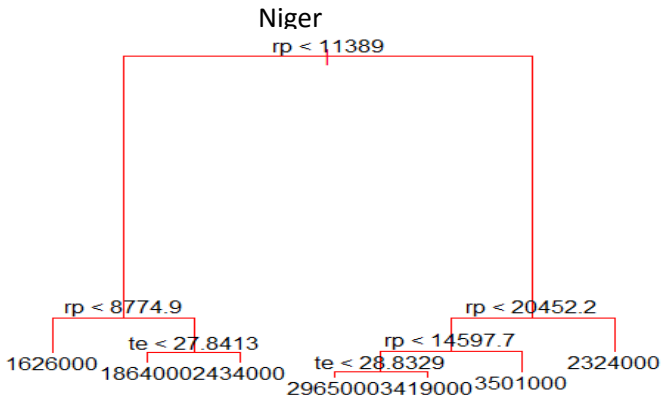
The application of the method resulted in the tree in **Figure** 1 **to 3** with the value of *j* and *u* that minimizes the RSS in equation 2 for the respective countries percent respectively. That is rural population (**rp**) is the regressor at the top of the tree used for the initial split such that:

R_1 = {X| rp < 11,389} and R_2= {X| rp > 11,389} for Niger (9)

R_1 = {X| rp < 13,059} and R_2= {X| rp > 13,059} for Ghana (10)

R_1 = {X| rp < 86,900} and R_2= {X| rp > 86,900} for Nigeria (11)

As soon as there were no more than 5 observations in each area, the splitting was stopped. The unpruned tree consumed approximately 88% of the training observations.



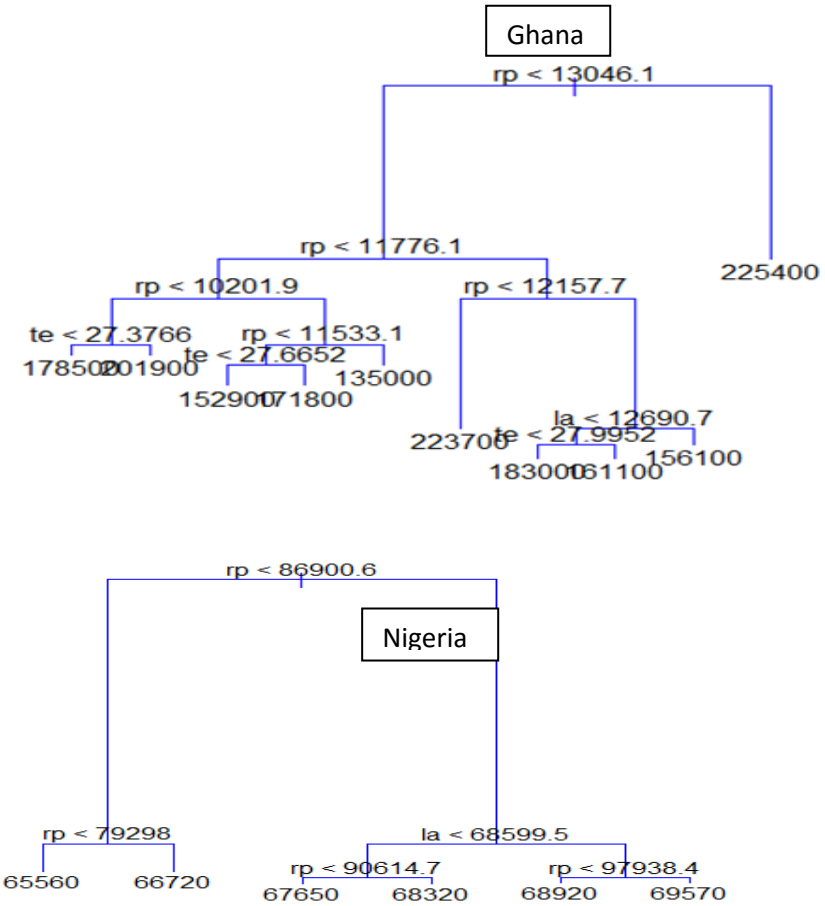


Figure 1: *Unpruned decision tree for Niger with seven leaves and six internal nodes. For Ghana, there are 9 internal and 10 terminal nodes, while for Nigeria, there are 6 and 5 nodes. Given that it minimises the RSS in equation 2, this tree model demonstrates that rural population (rp) is the most significant regressor in millet output across the three west African nations. The recursive binary technique was used for splitting.* ***Source****: Personal computation using* ***R*** *language*

On the training dataset, the RBS process shown in Figure 1 gave accurate predictions, but ultimately overfit the millet production data. One issue is that this can result in the decision tree model doing extremely poorly on the test dataset. By applying cost complexity pruning to create a smaller tree with fewer splits, we were able to improve test performance. In contrast to cross-validation and validation set pruning, we were able to fit the millet production tree model using this pruning strategy with reduced variance but somewhat greater bias by using fewer subtrees. The cost complexity pruning involves a positive tuning parameter α such that for every value of this quantity, there exists a subtree $T \subset T_0$ for which equation 3 is as small as possible.

$$C_n|T| = \sum_{m=1}^{|T|} n_{mp} H_m(T) + \alpha|T| \qquad (12)$$

Where,

$$H_m(T) = \frac{1}{n_{mp}} \sum_{x_i = R_m} (y_i - \hat{c}_m)^2$$

$$\hat{c}_m = \frac{1}{n_{mp}} \sum_{x_i = R_m} y_i$$

$$n_{mp} = \{x_i \in R_m\}$$

The number |T| stands for the number of leaves on the decision tree. T, R_m are the regions for the mth leave, and n_{mp} is the average amount of millet grown in the area R_m in the training dataset. The branches are cut back, and the choice between the complexity of the regression subtree and how well it fits the training data is managed by changing the value from zero. We used the 10-fold cross validation to find the values of the positive tuning parameter and the most complex tree. Cross-validation helped us figure out that the setting parameter = 18.785 and the RSS number = 403.652. Figure 2 shows what happened when the cost complexity trimming was done for different amounts of at each split. Cross evaluation showed that 4 was the best number of end nodes, so that's what we want is adjusting the nonnegative tuning value in equation 3 and a big tree on the training dataset with 186 observation was used to construct this trimmed tree.

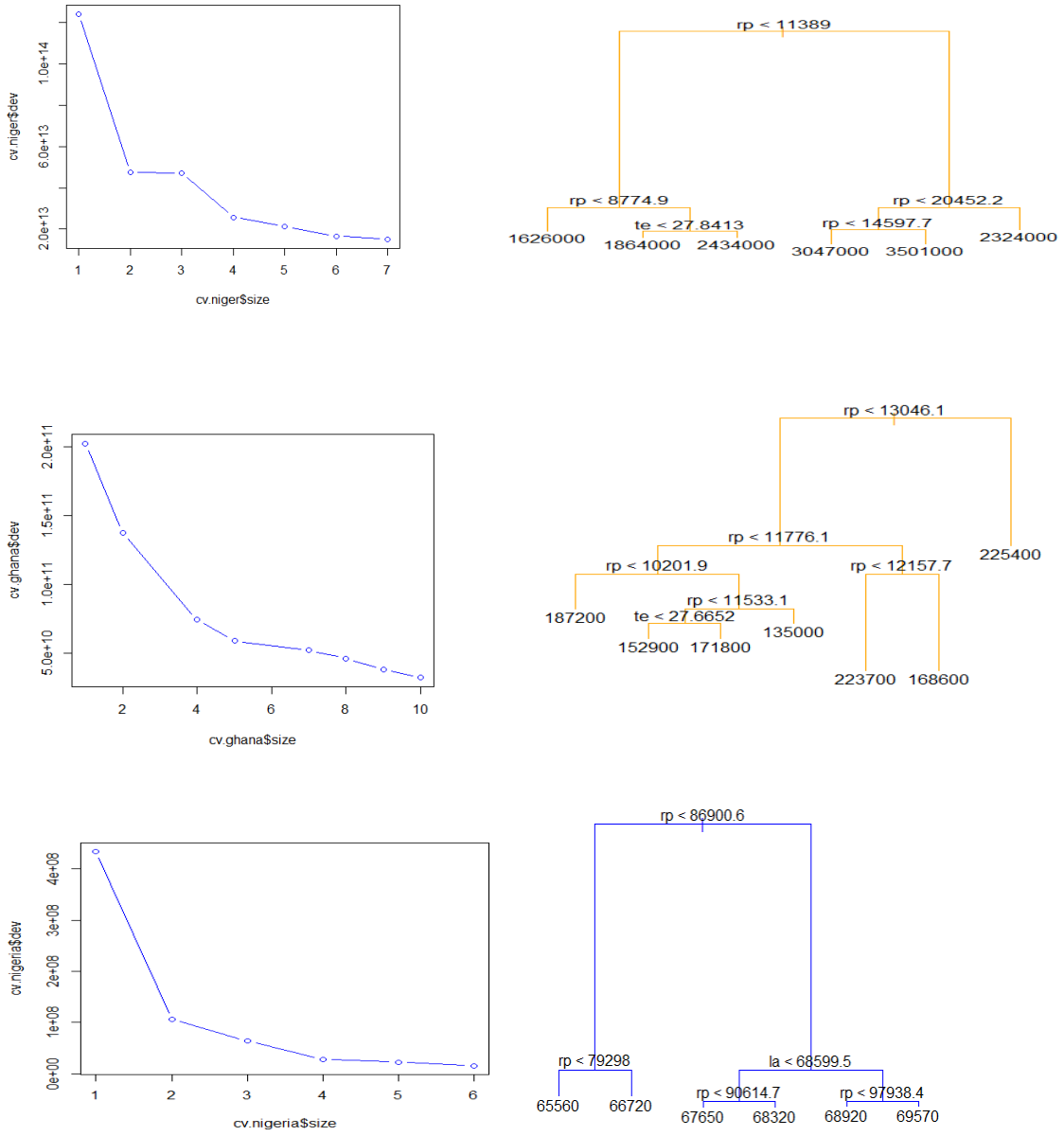


Figure 2: *Analysis of cost complexity pruning on the fitted decision tree of* ***Figure*** *2.* ***Right panel****: Prunned tree with 6 terminal nodes, 5 internal nodes for Niger; 7 terminal nodes, 6 internal nodes for Ghana and 6 terminal nodes, 5 internal nodes for Nigeria.* ***Left panel****: The result of 10-fold cross validation showing the cross-validation error*

(cv.niger$dev), (cv.ghana$dev) and (cv.nigeria$dev) as a function of the terminal nodes (cv.niger$size), (cv.ghana$size) and (cv.nigeria$size) respectively. It indicates that the CV error dipped at terminal node value of 6, 7 and 6 respectively for Niger, Ghana and Nigeria. ***Source****: Personal Computation using* ***R*** *language*

These trees have more variation but lower bias since they have been grown deeply and without pruning. The boosting algorithm finished with an aggregated G of 4,074 trees. We selected the shrinkage parameter to be = 0.0159 in order to minimise the variance and enhance the performance of our boosted model. On the test data set, the MSE of the boosting method using 15-fold cross-validation is 1.3953, which is a 64% improvement over the cost complexity trimmed model. Using the boosting method, we were able to produce the relative effect of the regressors in the millet production tree model (Figures 3 and 4).

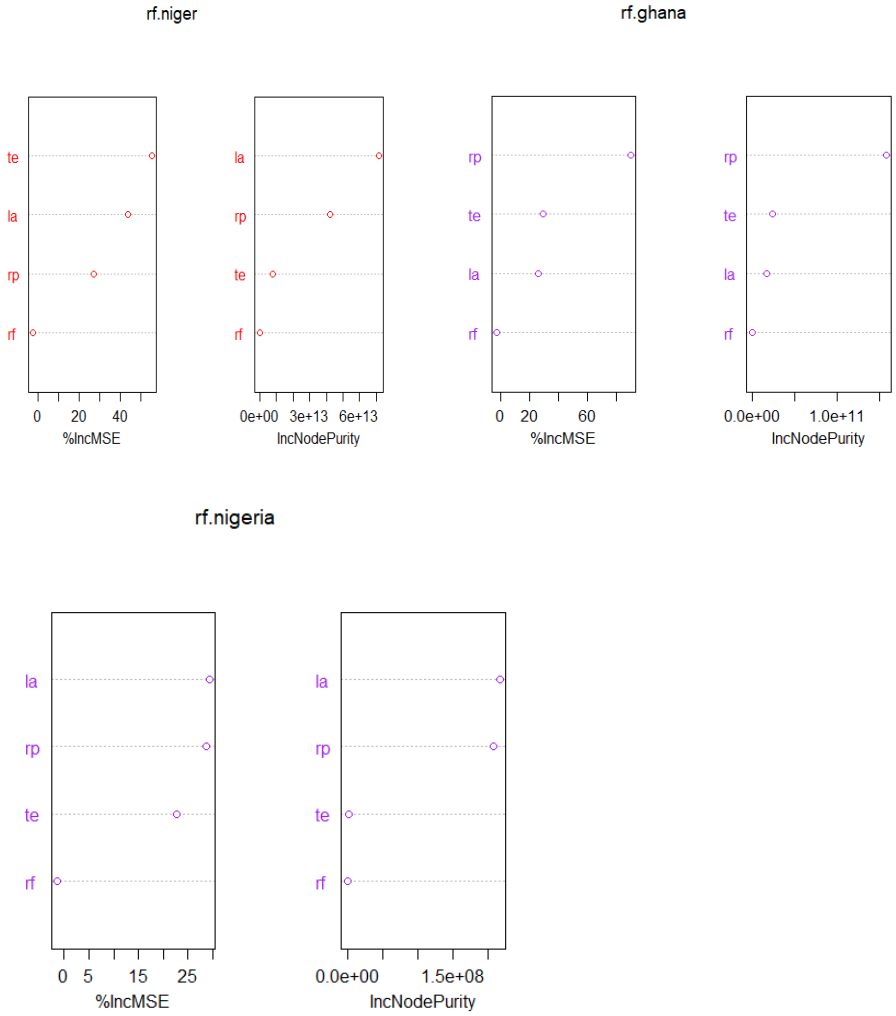


Figure 3. The importance of the factors on the MSE

Source*: Personal Computation using* ***R*** *language*

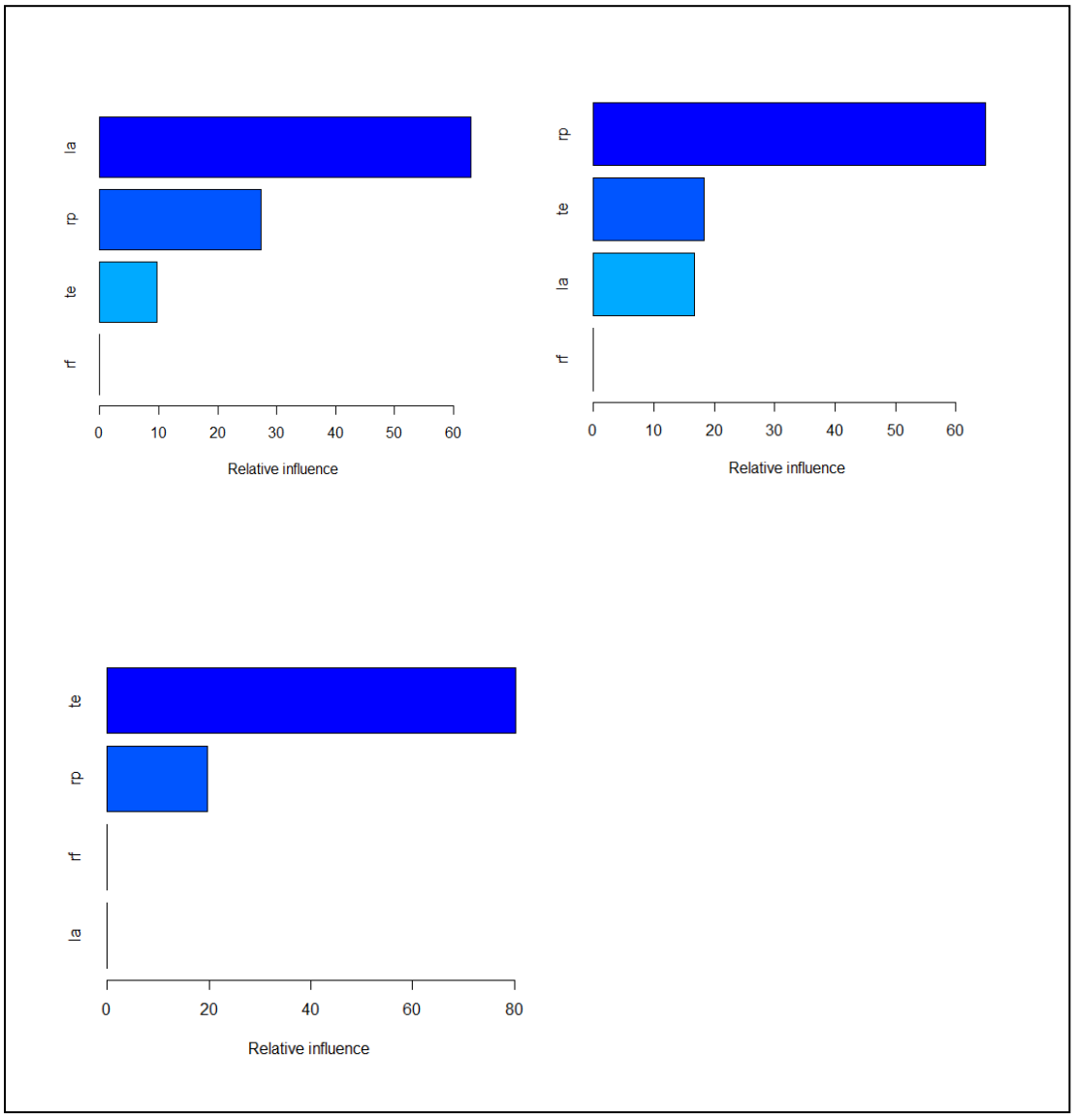


Figure 4. *Relative Influence of the regressor for Niger, Ghana and Nigeria*

Discussion

The use of decision tree methodology for regression analysis has been employed to investigate the correlation between millet production and four key variables, namely rural population, land area, rainfall, and temperature, among three distinct nations situated in West Africa. The study used the recursive binary splitting (RBS) technique, cost complexity pruning, and boosting. The use of Recursive Binary Splitting resulted in the generation of a tree structure consisting of 7, 10, and 6 terminal nodes, representing the countries Niger, Ghana, and Nigeria, respectively. Despite the fact that this singular, expansive tree yielded a minimal mean squared error (MSE) when applied to the

training dataset, its efficacy was found to be lacking when tested on the separate test dataset. Therefore, we used cost complexity pruning (CCP) in order to reduce the quantity of terminal nodes (leaves) and enhance the predictive capability and interpretability of our model. The implementation of this model led to a decrease in the mean squared error (MSE) for the test dataset. Additionally, it resulted in the formation of 6, 7, and 6 terminal nodes for the countries Niger, Ghana, and Nigeria, respectively. The minimization of the residual sum of squares (RSS) is achieved by optimising the relative humidity. Hence, the rural population was used as the primary variable for division at the root of the decision tree. The Cost Complexity Pruning technique included the use of diverse values for a tuning parameter, which served to regulate the balance between the model's complexity and its tendency to overfit. Boosting was used to enhance the prediction capability of the hypothesised decision tree model. Boosting is a methodology that entails the successive growth of a substantial number of individual trees, where the residuals from the previously formed tree are used as the response in the subsequent tree. The outcomes of using boosting techniques demonstrated enhanced performance of the regression tree model when evaluated on the test data set, as shown by a reduction in mean squared error (MSE). The study of partial dependency reveals that there is a positive relationship between millet output and rural population, as an increase in the latter leads to a rise in the former. However, the drop mostly occurs in conjunction with a decrease in rural populations, with the exception of Niger. This research demonstrates that the amount of millet production in West African nations is primarily influenced by two key factors: rural population and land area. The results of this study are consistent with previous independent studies conducted by Lawal and Ajibola (2017), Adetunji et.al. (2018), Ahmed and Adamu (2018), Yusuf (2019), and Ogunniyi and Ogunniyi (2020)

Conclusion

Based on the findings that rural population and land area play an important role in millet production in the countries investigated.

The following recommendations are made:

1. There is a need to encourage investment in rural infrastructure like roads, irrigation facilities, and storage facilities to enhance agricultural production, particularly in millet-growing regions.

2. There is a need to offer social facilities to support rural population expansion, which would improve millet and other crop output in rural regions.

3. Promote land tenure regimes that provide small farmers with access to land for millet cultivation.

4. There is a need for research to generate enhanced millet types that are resistant to pests and other illnesses in order to increase productivity.

5. Provision of assistance to help preserve soil fertility and mitigate the negative environmental effect of farming activities.

REFERENCES

Ahmed, A., & Adamu, M. (2018). Millet production and rural population growth: A case study of Ghana. *Agricultural Economics Research Review*, 31(2), 199-214.

Anigbogu, B. N., Abosi, C., & Okoli, J. N. (2015). Impact of climate change on West African agriculture: Evidence from empirical survey. *Journal of Agriculture and Environmental Sciences*, 4(1), 40-52.

Ayeomoni, M. O., & Aladejana, F. E. (2016). The role of the agricultural sector in economic development and poverty reduction: Empirical evidence. *International Journal of Agriculture and Rural Development*, 19(2), 289-302.

Diepeveen, D., & Armstrong, P. (2008). Enhancing decision-making in agriculture through data mining: A literature review. *Agricultural Systems*, 98(3), 148-157.

Ewetan, O. O., Fakile, A. S., Urhie, E., & Oduntan, E. (2017). Time series analysis of the relationship between agricultural output and economic growth in Nigeria (1981-2014). *Journal of Economic Development, Environment, and People*, 6(1), 32-49.

Lawal, O., & Ajibola, S. (2017). Factors influencing millet production in West African countries: A review. *African Journal of Agricultural Research*, 12(3), 135-143.

Obi, C., & Obayori, J. (2016). The implementation of appropriate policies and strategic investments in African agriculture: A potential for fostering expansion and development. *African Journal of Agricultural Research*, 11(17), 1466-1478.

Ogunniyi, O. S., & Ogunniyi, O. I. (2020). Soil fertility and sustainable millet production: A case study of Senegal. *Journal of Sustainable Agriculture*, 24(2), 87-98.

Omorokunwa, G., & Obadiaru, E. (2016). Allocating resources to the agricultural sector: Implications for economic growth in African countries. *Agricultural Economics Research Review*, 29(1), 135-147.

Ramesh, V., & Vardhan, R. (2013). Predicting crop yield in various agricultural sectors using data mining techniques. *International Journal of Computer Science and Information Technologies*, 4(5), 653-656.

Sertolu, M. R., Ugural, B. G., & Bekun, F. V. (2017). Impact of agricultural investments on economic development in Africa: A panel data analysis. *Journal of African Development*, 19(2), 87-102.

Teng, J., Lin, Y., Zhang, Z., & Huang, Y. (2021). Predicting Millet Yield in Niger Using Random Forests. Sustainability, 13(5), 2715. https://doi.org/10.3390/su13052715

Yusuf, S. O. (2019). The role of land area and rainfall in millet production in Burkina Faso. *Journal of Agricultural Science and Technology*, 21(3), 345-358.