AFRICAN JOURNAL OF APPLIED ECONOMICS (AJAE)
ISSN 3057-3335

Mzumbe University
Department of Economics

Determining the effectiveness of the sample split approach on supervised classification model on maize yield production

Justine N. Mbukwa
Department of Mathematics and Statistics Studies
P.O. Box 87, Mzumbe University-Tanzania
Correspondence: jnmbukwa@mzumbe.ac.tz

Abstract

The study was motivated by the challenges of yield volatility in developing countries, which in turn affects people's livelihoods and slows economic development. Maize is a staple food in Tanzania, consumed widely across both rural and urban areas. It is vital for national food security, providing a major share of daily caloric intake. Economically, it supports millions of smallholder farmers through subsistence and income. This study aimed to predict maize yield in Tanzania using discriminant supervised classification model. Data were collected using a structured questionnaire from 421 smallholder farmers in the Mbozi and Mvomero districts in Tanzania. Data analysis was performed using R programming 4.2.3. The results showed 0.867 classifier accuracy on the training sample, indicating a likelihood of the studied units being classified as low-yielding producers, with 13.3 percent of the expected cost of misclassification. Using the sample spilt approach, the study results on out-of-sample discovered the highest probability of farmers was classified as below average with 0.873 model performance and 12.7 expected costs of misclassification. Out of 100 cases (small farmers), 13 are misclassified, slightly fewer than what has been correctly classified. Applying the sample division approach, out of the 100 cases, 12.7 are misclassified. The classification model results indicated that the out-of-sample improves the model accuracy compared to the training sample, suggesting the intervention in resource allocation in terms of subsidies, training programs, and access to better seeds and fertilizers to the producers below the average.

Keywords: Supervised classification, Sample split, Quadratic discriminant model

Citation:

Mbukwa, J. (2025). Determining the effectiveness of the sample split approach on supervised classification model on maize yield prediction. *AJAE*, 1(1), 13-35

Manuscript info:
Received: 11th February 2025
Revised: 18th July 2025
Accepted: 6th August 2025
Published online: 14th August 2025

1. Introduction

Maize stands as a cornerstone of global agriculture, widely cultivated and essential for food, feed, and industry (FAO, 2023). Its adaptability and high demand make it vital to food security and rural livelihoods worldwide. This is because it plays a role in food security, livestock feed, and industrial applications. Its significance is evaluated in the following dimensions: Firstly, maize is a food security and staple food supply for millions of people in Sub-Saharan Africa and Latin America. It is one of the top three cereals worldwide (together with rice and wheat) that contribute significantly to daily calorie intake. Secondly, it reduces costs for both rich and poor nations in terms of food security (Erenstein *et al.,* 2022). Maize is a key staple in Africa, accounting for 30-50% of total caloric intake in countries such as Zambia and Malawi, where 90% of the population depends on it. Maize plays a vital role in global food security as a staple for millions. Beyond human consumption, it is a major component of livestock feed, accounting for over 60% of global production and supporting the meat, dairy, and poultry industries. Additionally, maize serves as a key input in industrial applications, including biofuels and processed goods.

Thirdly, maize is one of the most productive crops, allowing nations to dramatically boost their export earnings. The United States of America, Brazil, and Argentina are significant exporters, accounting for the majority of the maize consumed globally. In 2022, worldwide maize commerce reached 181 million metric tons, with maize exports adding billions to these economies (FAO, 2023; Ribeiro-Duthie *et al.*, 2021).Maize (*Zea mays*), commonly known as corn, is believed to have originated in southern Mexico over 9,000 years ago through the domestication of a wild grass called *teosinte* (Matsuoka et al., 2002). Indigenous peoples in the region played a crucial role in transforming this wild plant into a reliable food crop through generations of selective cultivation. As maize spread throughout the Americas, it became a central component of many cultures and diets. During the Columbian Exchange in the 15th and 16th centuries, it was introduced to Europe, Africa, and Asia, where it rapidly adapted to various climates and agricultural systems (Piperno & Flannery, 2001). This remarkable transformation from a wild grass in Mexico to a globally cultivated staple established maize as a cornerstone of modern agriculture, vital for food security, livestock feed, and industrial use. This crop contains nearly 72 percent starch, 10 percent protein, and 4 percent fat, supplying an energy density of 365 Kcal/100g.

Moreover, it is grown throughout the world, with the United States, China, and Brazil being the top three maize-producing countries globally. Together, they produce approximately 563 million of the 717 million metric tons of maize annually (Ranum *et al.,* 2014). In this regard, annual maize production stands at approximately 384 million metric tons for the United States, 273 million metric tons for China, and 108 million metric tons for Brazil. According to FAO statistics, it has been noted that in 2022, America was the topmost-producing region for maize globally. In this viewpoint, the United States of America and Brazil have been accounting for 39 percent of the world's production. Furthermore, statistics indicate China to be the second largest producer, accounting for 24 percent (FAO, 2023).Since its adaptation approximately 9,000 years ago, maize has contributed to an increasing and varied role in universal agri-food systems. In the past decades, maize production has suddenly increased universally. This contributed to an increase in a combination of demand and technological development. This has pushed the yield increases and area expansion. It has

been further noted that maize is the leading cereal in terms of production bulk, and it will become the most cultivated crop in the next decade. Again, globally, it has been noted that maize is adapted to be used multi-purposely to feed the population. Also, it is an imperative food crop, China, especially in sub-Saharan Africa, and Latin America, apart from other food crops in the regions (Erenstein *et al.*, 2022). In east Africa, especially; Kenya, Uganda, Tanzania, and Ethiopia; yields have been rising slowly, from ~1.03 t/ha in 1961 to ~1.75 t/ha in 2019 (+69%) (Epule *et al.,* 2022).

However, in developing countries, crop yields are lower than expected, which in turn affects production and economic growth (Akudugu *et al.,* 2012; Wiggins. & Keats, 2013). Despite that, both production and productivity are still low in all of Sub-Saharan Africa, and agriculture continues to be an important economic sector (Ahmed *et al.,* 2013). In most African countries, soil fertility has been declining due to low levels of fertilizer usage and limited access to water resources, which contribute to a decrease in maize productivity (Shi & Tao, 2014). On the other hand, the growing population in Africa puts pressure on agricultural resources together including maize yield and harvest area. Transversely, in Africa, access to advanced options for increasing the yield is insufficient (Epule *et al.,* 2022). Africa is one of the region's most severely affected by climatic and non-climatic factors. These drivers continue to limit maize and other crop yields, posing a significant threat to the continent's food security (Tesfaye *et al.,* 2015). In this response, many farmers have attempted to increase their harvest area as a strategy to boost maize production (Epule *et al.,* 2011). In West Africa, low yields have been reported among small farmers in Ghana (Akudugu *et al.,* 2012).

In East Africa, particularly in Uganda, Woniala and Nyombi (2014) reported that low corn yield ranges from 150 kg to 1992 kg/acre. Also, the problem of low crop yield has been noted in Tanzania (Haug & Hella, 2013). On the other hand, it has been noted that despite the interventions implemented by the government of Tanzania, such as subsidies and the elimination of unnecessary taxes, agricultural productivity is still low, especially for the marginalized smallholder farmers (Mkonda & He, 2018). Despite Tanzania's implementation of subsidies and tax reforms aimed at improving input access for smallholder farmers, maize yields have seen only limited improvement. Challenges such as inefficient subsidy targeting, delays in input distribution, and limited farmer awareness reduce the effectiveness of these programs.

Additionally, infrastructural constraints, poor market access, and climate-related stresses undermine productivity gains. Addressing these systemic issues alongside policy reforms is essential to translate subsidies into tangible yield improvements for smallholders (Mgonja *et al.,* 2017; Saitoti & Ngalawa, 2020). While many studies have explored the factors limiting maize yields; few have evaluated the probability of smallholder farmers reaching expected yield levels, particularly using discriminant analysis as an analytical approach. This paper contributes to the body of knowledge by revealing the likelihood of smallholder farmers who are at risk of producing low maize yields. This was further analyzed along with the sample split methodology for improving statistical power due to the low cost of the proposed discriminant model.

2. Theoretical framework

The paper has been grounded in three theories, namely, the Production function theory, human capital theory, and Cost-Benefit Analysis (CBA). The Production function theory was first developed (Cobb & Douglas, 1928) and aims to answer the question of how inputs are combined to produce

a particular level of output. The Cobb-Douglas production function is particularly widely used to describe agricultural production. It shows the relationship between inputs such as land, labor, capital, and technology and output.

The Cobb-Douglas production function, while widely used, has key limitations in the context of smallholder farming. It assumes constant returns to scale and perfect competition, which rarely reflect the realities faced by smallholders who often operate under imperfect market conditions with limited access to resources. The model also fixes the elasticity of substitution between inputs at one, ignoring the limited flexibility in substituting factors like land and labor. It overlooks critical factors such as climate variability, risk, and shocks, which significantly affect smallholder yields. Furthermore, it treats inputs as homogeneous and fails to capture the dynamic nature of farming, such as learning and technology adoption over time. As a result, more flexible models like Translog or stochastic frontier analysis may offer better insights for analyzing smallholder agricultural production.

However, the Cobb-Douglas production function and discriminant models serve different purposes in statistical and econometric modeling, so whether the Cobb-Douglas function is "best for variable identification" when using discriminant analysis depends on what you're trying to achieve. The purpose of the Cobb-Douglas function is to model output (continuous) as a function of inputs (e.g., labor, capital), whereas the Discriminant Model aims to classify the observations based on predictor variables into known classes. Several studies have critiqued the applicability of input-based production theories like the Cobb-Douglas production function in the context of developing economies, particularly where informal labor markets and subsistence farming dominate (Pingali & Sunder, 2017). Barrett *et al.* (2010) argues that assuming labor is a homogeneous input ignores variations in skill, experience, and time allocation across household members; Market Imperfections. In the context of this paper, input variables such as improved seeds, labour (hired and domestic labour respectively), fertilizer, pesticides, farm size, and livestock (goats, sheep, and chickens) are crucial for maize productivity, and other control variables have been used as a discriminating variable.

The second important theory is the “pioneer theory of human capital,” originated by Becker (1964). He connected the idea that investments in education, training, and health can increase the productivity and economic value of a person as physical capital. Becker’s theory suggests that education and training enhance productivity, typically measured through years of formal schooling. However, excluding informal education, such as indigenous knowledge, can lead to model misclassification, especially in contexts like farming, where individuals with limited formal education may still possess high productivity due to traditional expertise. In this study, the use of local seed as a variable is justified, as it serves as a proxy for indigenous knowledge, capturing the informal skills and practices that significantly influence agricultural productivity, particularly within classification models like Quadratic Discriminant Analysis (QDA) (Becker, 1993; Schultz, 1961). Also, a positive, significant effect of each additional year of education on maize yield has been noted (Solomon, 2019).

The Cobb-Douglas function and Cost-Benefit Analysis match conceptually and practically when used together. Cobb-Douglas provides the technical efficiency or productivity estimates, while CBA evaluates whether those gains translate into economic value (Chirwa, 2025).

In agriculture, it is indicated that the level of education from the point of view of the number of years of training can determine the capacity of the farmer to adopt new methods of agriculture, to make reasonable decisions, and to manage resources effectively. It has been revealed a positive, significant effect of each additional year of education on maize yield (Solomon, 2019). The age of the respondent often reflects experience, which could influence farming decisions and productivity.

Cost-benefit analysis (CBA), originated by (Coase, 1960). His theory emphasizes the relationship between the cost of resources and gain or production. Farmers need to balance the costs of resources (seeds, fertilizers, labor) with the expected increase in production (profitability) and sales. Farmers seek to utilize resources that will generate the highest return on investment (ROI). If the cost of resources (such as improved seeds) is high and the marginal yield increase is low, it may not be profitable to use these resources. Thus, the transition from identifying productivity drivers to employing discriminant analysis is not abrupt but rather a logical methodological progression. Identifying key productivity determinants is grounded in both theoretical frameworks and empirical evidence. However, while understanding these drivers is crucial, statistical classification techniques such as discriminant analysis offer additional analytical power. They enable the systematic classification of outcomes, such as yield status based on multiple covariates, thereby enhancing the ability to distinguish between different groups or performance levels within the population under study. Beyond the theoretical conceptualization, some empirical studies were carefully scrutinized to capture the relevant covariates and statistical modeling as per section 1.2

3. Empirical review

Since the research problem is based on classification modeling, understanding some of the drivers that affect productivity from the previous studies is indispensable. Obasi (2013) discovered that educational level, farm experience, farm size, extension exposure, and workforce as positively related to productivity. Previous studies have also reported other factors (land-to-work ratio, use of fertilizers, pesticides, manure, and household size) affecting land productivity studies (Urgessa, 2015). Shita *et al.* (2018) revealed that fertilizer and real domestic products affect productivity. Kakar *et al.* (2016) found that rainfall, acreage, fertilizer, and credit had a positive impact on agricultural productivity. Nuno and Baker (2021) found that the agricultural experience of the head of household, the number of economically active family members, and quantity of organic fertilizer applied, the irrigated land area, and arable land fertility significantly affect agricultural productivity. Adimassu and Kessler (2016) have shown that livestock, land tenure, labor, and social capital affect yield productivity as a result of a lack of rainfall. Moreover, Srivastava *et al.* (2007) argued that quadratic discriminant modeling is a Bayesian distribution-based classifier that minimizes the expected Bregman divergence of any class conditional distribution and also minimizes the expected misclassification costs. The study is in line with quadratic classification because it is specified in the context where the outcome is a two-group outcome, where their sample variance-covariances are not equal. Thus, the classification of the cases given several continuous predictor variables, unlike other classification models such as logistic regression, tree-based models etc.

In terms of crop yield, the model has proven to be appropriate because the farmers are likely to produce low, medium, or high yields. These classes are known in advance. Morais and Lima (2018) argued that quadratic discriminant modeling is appropriate for supervised classification problems. It is responsible for predicting the

odds of each class as a Gaussian distribution and uses posterior probability to estimate a maximum-likelihood class. Among other documented statistical prediction methods, Mupangwa *et al.* (2020) found that the discriminant model has the highest predictive power for maize yield. Empirically, Alhassan *et al.* (2016) used the quadratic discriminant model to classify farmers into known risks (low, medium, and high risks) that they encountered earlier in maize production. The results revealed classification rates of 80 percent (low risk), 89 percent (medium risk), and 93 percent (high risk). Although these scholars used the same classification model in question, however, they have failed to refine the estimates, farmers' forecast estimates (probability) of a real product below or above average (FAO, 2010; Urassa, 2010). Agrawal *et al.* (2012), used the time series of wheat yield 30 years (1970-2000) to divide the outcome into three groups (congenially, normal, and unfavorable) based on yield distribution. Using these three classes as the known populations, the discriminant model function was fitted. The scores generated were used as independent variables in the modelling.

Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description (Shmueli, 2010). Model accuracy is the most imperative part of the construction of a supervised model. In this regard, a good generalization performance must have a sensible data-splitting approach, and this is decisive for model authentication (Fall *et al.,* 2015; Xu & Goodacre, 2018). In the present study, the model performance was assessed by splitting the datasets into a training set and a test set without introducing any bias (Joseph, 2022). The first part of the data is meant for fitting (estimation of the unknown parameters) in the model, whereas the second part is for assessing the accuracy. On the other hand, the training data is used to fit the model, while the testing data is used to measure how well the model predicts new and unseen data. For practical purposes, therefore, the overall study sample was divided into a modeling set/training sample (80 percent) and an external evaluation/out-of-sample set (20 percent) (Martin *et al.,* 2012).

4. Methodology

This paper used the secondary data collected from the smallholder farmers (unit of analysis) who are the members of small farmers in Tanzania, namely Mtandao wa Vikundi vya Wakulima Wadogo Tanzania (MVIWATA). The data focused on smallholder farmers responsible for the maize cultivation. The data were collected via a survey between October 2015 to mid of February 2016, where two districts, namely Mbozi and Mvomero were sampled. The protocol for conducting the study was observed and the letter was granted by Mzumbe University. A letter was used to introduce the government officers from the regional level to the local level, where data used to be collected. By having an official letter introducing the researchers, the process of gaining access to data is well streamlined. The letter and formal introduction ensure that the study complies with institutional protocols, ethical guidelines, and permissions from relevant authorities. This helps to demonstrate that the research is being conducted responsibly, which is essential both for the integrity of the study and for protecting the rights of participants.

A standard structured questionnaire was used to collect structured data, responsible for answering specific research questions for this study. Of 430 cases, a random sample size of 421 was met (97.9% response rate), where the analyses were based. To arrive at the lowest unit of analysis, multistage random sampling was used alongside five stages. In stage one, two regions were purposively selected based on agro-climatic

zones. In stage two, one district was randomly selected, from each region. This was followed by selecting the wards each district. Households were then selected randomly from village registers using simple random sampling

Data analysis was performed using the R programming language (version 4.3.2). To validate the performance of the quadratic classifier model, the sample partitioning approach was used (Shmueli, 2010; Martin *et al.,* 2012; Korjus *et al.,* 2016). This means the sample consisting of the response rate was partitioned into two sub-samples: a large sample (training) and a small sample (test) for prediction and performance assessment. The process is governed by two measures or criteria. In this paper sample split ratio (80:20), where 80 percent of the dataset is classified as training and 20 percent is testing, was used. This is done on random split bases, in which cases were randomly divided into training and testing sets based on a specific ratio. Despite the limitations associated with small sample sizes, a stratified 80:20 train-test split was employed to ensure a balanced representation of outcome classes and to minimize sampling bias. This approach is consistent with strategies recommended for moderate-sized datasets. To assess model performance, multiple classification metrics were used, including sensitivity and specificity. A model is considered to perform well if it achieves both sensitivity and specificity of at least 80% (Kohavi, 1995; Yadav *et al.,* 2021).

In this current paper, the measured variables were defined. The outcome variable (maize yield), has been defined as a ratio of the amount of crops harvested in terms of a kilograms or metric tonnes (t) per crop area per hectare (ha) (FAO, 2010). Following similar grounds, the present study used the yield as an outcome defined as output (number of bags in tonnes per acre. Furthermore, the recall approach was used to capture the quantity of maize along the respective units from selected small farmers who participated in the study (Me-Nsope & Larkins, 2016). Given this approach, the target individual smallholder farmers that were interviewed and requested to provide the quantities harvested in his/her field/farm for the previous farming season within six months. The respondents provided the production in terms of bags as well as the known number of tins/buckets which are based on kg or metric tons.

To establish the classes (below or above the average), the categorization process was based on the National Bureau of Statistics report (*The Tanzania National Sample Census of*, 2011). The report noted that the annual maize per acre in Tanzania is 1.3 tonnes (3.75 bags per acre). However, this figure found so far as compared to that of South Africa and the World at large stands at 2.3 tonnes per hectare (7.805 bags per acre) and 4.3 tonnes per hectare (12.42 bags per acre) respectively (FAO, 2010; Urassa, 2010). Therefore, the midpoint of the maize yield between South Africa and the World estimate is given as (7.805 +12.42)/2 = 10.11 bags per acre. Thus, in this study, the middle value has been taken as a benchmark for the smallholder farmers whose productivity is either below (population one-π_1) or above the average (population two-π_2).

The predictor variables were measured in terms of scale. These include: age of the respondent (age), number of year at school (nsyear), household size (hsize), number of goats (ngoat), number of sheep (nsheep), number of chicken (nchicken), quantity of the local seed in kilogram (qtylocalseedmaize), number of hired labor (nhiredlabor), number of home labour (nhomelabor), quantity of improved seed in kilograms (qtyimprseedmzrain), quantity of pesticides in litres (qtypestmzrain), quantity of fertilizer in kilograms (qtyfertmzrain), past harvest of maize in kilograms (pastharv), farm size (fsmzr2), quantity of maize sales in

kilograms (qtysamz_yr2), cost of improved seed maize (cost_imprmaizer), cost of fertilizer (cost_fertmaizer), cost of pesticides (cost_Lpestmaizer). Because, there are many predictors, the best inputs with the power to discriminate cases to the respective classes were selected based on Wilks lambda statistic (El Ouardighi *et al.*, 2007).

Classification modelling

Since the outcome variable is group-based, supervised statistical modeling is appropriate to use. Supervised classification refers to predictive statistical analysis where a model is built in such a way that a new situation is speculated via known facts (Maindonald, 2012). From this standpoint, the classes to which the observational vectors are classified are known in advance. The Linear Discriminant Model is one of the statistical methods used. It predicts the probability of known target groups given the selected features in the form of a linear combination (Härdle & Simar, 2003; Raykov & Marcoulides, 2008). Once a sample has been divided into two classes, it is assumed that the sample variance-covariance matrices for two groups are the same ($S_1 = S_2$) whereas each group is characterized by a multivariate normal distribution, both $f_1(x)$ and $f_2(x)$ are associated densities with the mean vectors and covariance matrices of μ_1, Σ_1 and μ_2, Σ_2 correspondingly. Given the random variable $X^T = [x_1, x_2, ..., x_p]$ and the attached labels π_1 and π_2, then, the joint distribution is represented as follows:

$$f(x \mid \pi_i) = \frac{1}{(2\pi)^{\frac{p}{2}} \mid s \mid^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \bar{x}_i) s^{-1} (x - \bar{x}_i) \right) \quad \text{for } i = 1, 2 \qquad \ldots(1)$$

Nevertheless, for practical purposes, the linear discriminant model is recommended if the variance–covariances are the same. It should be noted that for practical situations, the sample variance-covariance matrix is used in place of the population variance-covariance.

Quadratic classifier

The quadratic classification for classification of the normal population is recommended to be used when the sample variance-covariance are dissimilar $(s_1 \neq s_2)$ for populations one and two, such that: $\boldsymbol{\pi_1} : N(\underset{\sim}{\bar{\mathbf{X}}}_1, s_1)$ and $\boldsymbol{\pi}_2 : N(\underset{\sim}{\bar{\mathbf{X}}}_2, s_2)$ when $s_1 \neq s_2$. Now, if $f_1(\underset{\sim}{\mathbf{x}})$ is the probability density function (p.d.f) of population one ($\boldsymbol{\pi_1}$) and $f_2(\underset{\sim}{\mathbf{x}})$ is the p.d.f of population two ($\boldsymbol{\pi_2}$). The allocation rule that minimizes the ECM is given by:

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi_1}$ if

$$-\frac{1}{2}\underset{\sim}{\mathbf{x}}_0'(s_1^{-1} - s_2^{-1})\underset{\sim}{\mathbf{x}}_0 + (\underset{\sim}{\bar{\mathbf{X}}}_1 s_1^{-1} - \underset{\sim}{\bar{\mathbf{X}}}_2 s_2^{-1})\underset{\sim}{\mathbf{x}}_0 - \lambda \geq \log w \qquad \ldots(2)$$

Where $w = \frac{c(1|2)\,p_2}{c(2|1)\,p_1}$

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$ otherwise.

Where $\lambda = \frac{1}{2}\ln\left(\frac{|s_1|}{|s_2|}\right) + \frac{1}{2}\left(\bar{x}_1 s_1^{-1} \bar{x}_1 - \bar{x}_2 s_2^{-1} \bar{x}_2\right)$

s_1 : sample variance-covariance of the observed data group 1

s_2 : sample variance-covariance of observed data group 2

$\bar{x}_1$: sample mean vector for the observed data group 1

$\bar{x}_2$: sample mean vector for the observed data group 2

p_1 : prior probabilities of the population one ($\boldsymbol{\pi}_1$)

$\underset{\sim}{\mathbf{x}}_0$: observed data matrix

p_2 = prior probabilities of the population two ($\boldsymbol{\pi}_2$)

$c(1|2)\,p_2$ =cost of misclassifying cases to population one ($\boldsymbol{\pi}_1$) multiplied by its prior probability

$c(2|1)\,p_1$ = cost of misclassifying cases to population two ($\boldsymbol{\pi}_1$) multiplied by its prior probability

In practice, before fitting the discriminant model, the equality of covariance matrices between the two groups must be tested. In this article, the Box's M test, an exact test, was used to check the violation of this assumption. The test can be transformed into a statistic helping as an approximate test based on chi-squared and F distributions. If the computed value exceeds the tabulated value at a specified level of significance (Friendly & Sigal, 2020; Jiamwattanapong & Ingadapa, 2021) or reject the null hypothesis if the p-value is smaller than the level of significance (Ahmed *et al.,* 2013; Hair *et al.,* 2014).

5. Results

Results for the equality of the sample covariance matrices

Table 1 indicates the results of the statistical test for the equality of the two groups. The findings discovered that there is a strong, statistically significant evidence to argue that the sample covariance matrices for the group of farmers whose productivity is below the average are dissimilar from those of above the average, as the value is less than 5% alpha (p-value < 2.2e-16) (Ahmed et al., 2013). From this standpoint, the quadratic

classifier is found to be relevant to classifying the cases into the known groups, and the two groups, below and above, yield, respectively. This means the sample variances and covariances are unequal across groups, violating a key LDA assumption. The results provide strong justification for using QDA instead of LDA.

Table 1: Testing the equality of the sample variance-covariance matrices

Chi-square test	Degree of freedom (df)	p-value
526.53	21	p-value < 2.2e-16

Source: Findings (2025)

Feature Selection Results

Table 2 indicates six selected features out of nineteen variables for discriminating the smallholder producers into the respective groups based on the Wilks lambda statistic (El Ouardighi *et al.,* 2007). These variables have been defined in section 3) of the methodology. The selection is of great importance because the input measurements help in discriminating the cases to the respective classes with small noise. Thus, the findings indicated that the retained predictors have the smallest p-values, less than 5% alpha and except for the number of goats judged at 10% level of significance. This means the most significant discriminating variables were selected based on the highest discriminatory power to ensure the accuracy of the classification rule with a low error rate. Wilks' statistics are crucial for determining the importance of a variable in terms of its discriminatory power to substantiate the groups (e.g., above against below-average producers). The smaller values propose that the predictor values are capable of distinguishing between these groups. Ranking significantly, the findings revealed the variables affecting small farmers' productivity, including past harvests (in kilograms), cost of fertilizer in Tanzanian shillings, quantity of maize sales, household size, local seed quantity in kilograms, and number of goats. The findings show that past harvest quantity and fertilizer cost under rainfed conditions were the strongest discriminators (Wilks' Lambda = 0.54 and 0.46; $p < 0.001$). Maize sales (0.45), household size (0.44), and local seed use (0.43; $p < 0.05$) also contributed significantly, while goat ownership showed marginal significance (0.43; $p = 0.0798$). These results suggest that both production-related variables and household characteristics play a key role in classifying maize yield groups, highlighting the value of QDA in identifying factors influencing smallholder productivity.

Table 2: Stepwise Discriminant Analysis Results

Variables	Wilks.lambda	F.statistics.overall	p.value.overall	F.statistics.diff	indep. p.value.diff
pastharv: Past harvest in kgs	0.5401367	356.7296	5.218831e-58	356.729597	0.0000***
cfertmzr: Cost of fertilizer for maize under rainfed in Tanzania in shillings	0.4642074	241.2297	2.202060e-70	68.371190	0.0000***
qtysamz: quantity of maize sales in kilograms	0.4482790	171.0747	2.706007e-72	14.817020	0.0001***
hsize: Household size (hsize)	0.4385046	133.1697	3.996934e-73	9.272816	0.0025***
qlsdmz: quantity of the local seed in kilograms	0.4339466	108.2678	5.695180e-73	4.358971	0.0374**
ngoats: number of goats	0.4307382	91.19010	1.358753e-72	3.083676	0.0798*

Source: Findings (2025 (***) Statistically significant at 1% level of significance; (**) Statistically significant at 5 % a level of significance; (*) Statistically significant at 1 % level of significance

Proportion Estimation by Group

Table 3 indicates the proportion of smallholder farmers in terms of maize productivity. Based on the prior Probabilities, the findings show that 60.33% of the studied smallholder farmers are producing below average. Therefore, the findings propose a higher occurrence of below-average smallholders in the studied sample of producers.

Comparing group means is essential in discriminant classification analysis, as it reveals significant differences between predefined groups in this case, smallholder farmers with below- and above-average maize yields. The QDA results show clear distinctions, indicating non-random yield variation. Above-average farmers spent more on fertilizer (149,063 TZS vs. 43,805 TZS), had higher past harvests (17.12 kg vs. 5.36 kg), and sold more maize (12.23 kg vs. 1.62 kg), highlighting the role of input use and market orientation. Below-average farmers relied more on local seeds, which may limit productivity. Greater goat ownership in higher-yielding households suggests livestock supports resilience. These findings call for policies enhancing input access, market linkages, and mixed farming support.

Table 3: Quadratic Discriminant Classifier

Prior probability of groups						
Below the average 0.6033254				Above the average 0.3966746		
Predictors Variables						
Yield status	Past harvest in kgs (pastharv)	Cost of fertilizer's maize under rainfed (cfertmzr)	Household size (hsize)	quantity of the local seed in kilogram (qlsdmz)	quantity of maize sales in kilograms(qtysamz)	Number of goats reared (ngoat)
Below the average	5.362087	43805.01	5.775591	4.241142	1.622689	1.118110
Above the average	17.116946	149062.87	6.479042	3.526946	12.227545	2.299401

Source: Findings (2025)

Predicated Counts per Group

Table 4 indicates the predicted counts for the studied small farmers. The findings indicated that, out of 421 studied cases, 258 of them were predicted to be below the average, whereas 163 were found above the average maize productivity, respectively. This predicted figure opens up a discussion and intervention on why smallholder farmers are characterized by a high likelihood of producing below the average. The discussion has been well articulated in section four (4) of the paper.

Table 4: Total Predicted Counts per Group

Status	Maize yield	
	Below the average	Above the average
Predicted cases	258	163

Source: Findings (2025)

Actual Group Predicted Membership

Table 5 illustrates the actual group and predicted group membership. The study results disclosed that out of the 421 smallholder producers interrogated, 228 of them whose maize productivity was below the average, were correctly classified with productivity below the average while 26 of the farmers were misclassified above the average given their yield. On the other hand, 137 of the cases were correctly classified with yield above the average while 30 of the cases misclassified. These results may prove that among the studied cases, probably there is a great chance to argue that a high number of cases being below the average, may be correctly placed justifying that they are failing to meet their expectation.

Table 5: Predicted Group Membership (Confusion Matrix)

Actual Group	Predicted group		
	Below the average	Above the average	Total
Below the average	228	26	254
Above the average	30	137	167
Predicted size	258	163	421

Source: Findings (2025)

Quadratic Discriminant Prediction

Table 6 indicates the model results using the quadratic classifier function. In the confusion matrix, the findings indicate that out of 228 producers correctly categorized as below average, 30 producers were incorrectly categorized as above average. Contrarywise, 26 producers classified as above average were below average. This categorization accuracy directs the model's effectiveness in predicting productivity categories. Furthermore, the findings revealed that the overall accuracy of the correctly classified cases was 0.867. The smallholder farmers that were correctly classified to be below the average were 89.76% (sensitivity), whereas about 82.04% (specificity) of the smallholder farmers were classified to be above average. This high accuracy suggests that the proposed quadratic discriminant model can be reliably used to identify farmers who may need support or targeting interventions aimed at improving the performance of these farmers. In this paper, the interpretation of the classification metrics was based on: >=90% (excellent); 80–89% (good); 70–79%(fair); <70% (poor) as proposed (Zhou, Obuchowski & McClish, 2011; Pepe, 2003).

Given the Kappa value of 0.721, the results indicate a substantial agreement between the predicted and actual classifications. This signified that the model provides reliable classifications beyond chance. On the other hand, the findings showed that the maximum posterior distribution (positive predictive value=PPV), the probability of the farmers whose maize yield is below average given that their results are positive, was 0.8834. The results revealed a high likelihood that producers identified as below average truly are in that category. This enhances confidence in targeting support programs effectively. The minimum posterior distribution (the probability of the farmers whose maize yield is above the average to be attained) i.e. negative predictive value (NPV = 0.8405).

Based on positive predictive value results, it is likely to argue that the ratio of smallholder farmers truly classified as below the average they really deserve to be below the established threshold, and hence they fail to meet their expectation on maize productivity. On the other hand, the ratio of the net predictive value may suggest that cases meet their maize yield expectation to some extent.

Table 6: Predictive Membership using original data set Model Validation (Confusion Matrix and Statistics)

Confusion matrix	Below average	Above average
Below average	228	30
Above average	26	137
Metric		**Value**
Accuracy		0.867
95% CI		(0.8308, 0.8979)
No Information Rate		0.6033
P-Value [Acc > NIR]		<2e-16
Kappa		0.721
Sensitivity		0.8976
Specificity		0.8204
Pos Pred Value		0.8837
Neg Pred Value		0.8405
Prevalence		0.6033
Detection Rate		0.5416
Detection Prevalence		0.6128
Balanced Accuracy		0.859
'Positive' Class		below the average

Source: Findings (2025)

Table 7 indicates the out-of-the-sample results based on the sample split approach. This is important to find out if improves the model accuracy. The findings indicated a better accuracy rate of $\left(0.873(95\%\ CI: 0.802-0.9256)\right)$ compared to the predicted results under the training sample $\left(0.867\left(95\%\,\mathrm{CI}: 0.8308, 0.8979\right)\right)$. Since the kappa statistic is close to 1, indicates that the model is very reliable and accurate. Compared with other statistical tests, therefore, there is enough statistical evidence to conclude that the test set provides more consistent outcomes with small errors, since most of the statistical tests are within the reasonable range. Therefore, it is more likely to argue that the expected high maize yield is difficult to achieve given the predicted results, unless some intervention measures are taken to rescue the small farmers. On the assessment of the model fit using the test set, it was discovered that the model indicates the soundness for the counts of the small-holder farmers whose maize yield is below the average (Sensitivity=0.8837), while the ability of the model under this set indicates that the farmers whose maize yield is above the average (Specificity=0.8500).

Table 7: Validation (Test) Set Prediction Results

Confusion matrix	Below average	Above average
Below average	76	6
Above average	10	34
Metric		**Value**
Accuracy		0.873
95% CI		(0.802, 0.9256)
No Information Rate		0.6825
P-Value [Acc > NIR]		6.17E-07
Kappa		0.7146
Sensitivity		0.8837
Specificity		0.85
Pos Pred Value		0.9268
Neg Pred Value		0.7727
Prevalence		0.6825
Detection Rate		0.6032
Detection Prevalence		0.6508
Balanced Accuracy		0.8669
'Positive' Class		below the average

Source: Findings (2025)

6. Discussion

The paper intended to find out the extent to which the sample split approach contributes to the body of knowledge by revealing the likelihood of the smallholder farmers who are at risk of producing low maize productivity. In line with the descriptive statistics, some indispensable findings have been revealed along with studied small farmers. Firstly, it has been revealed that smallholder farmers experience a low average past harvest from the maize crop. The low average maize productivity of the past harvest may contribute to low motivation to cultivate more. Secondly, it has been found on average that, it is too expensive to buy fertilizers compared to the large-scale farmers classified above the average. This may be attributed to the fact that the study participants experience limited financial resources to afford to buy in bulk compared to the giant farmers. Thirdly, on average most of these farmers use the local seed instead of improved seeds and this is among the reasons that suggest the low yield. Dominantly, they use local seed may be because of the limited financial and assets resources. Fourth, the findings showed that they are not able to keep a large number of goats on average and probably fail to supplement the natural manure for restoring land fertility.

Based on the results in Table 2, the present paper has revealed crucial results in line with the statistical significance. The low p-value of 5.218831e-58 for past harvest proposes that this variable has a meaningful impact on maize productivity. This shows that interventions pointing to this variable could significantly raise productivity. Significantly, the findings propose that the cost of fertilizers impacts productivity in terms of affording to buy farm inputs. Thus, smallholder producers below average are likely to suffer because of high input costs. This, in turn, may affect their capability to invest in essential resources for agricultural

sustainability. Given the household size and livestock results, the findings propose that the household size and the number of goats suggest that families with livestock may experience diverse production dynamics. Thus, the findings suggest that livestock management is key to improving productivity. The local seed variable affects productivity significantly among smallholder producers. The findings suggest that increasing the level of productivity, access to improved seed varieties, or better seed management practices are key.

Concerning Table 3 of the presented results mean across the below and above the average, the comparison across the identified key dominant variable has been taken into account. In terms of mean group comparison, the results have revealed very interesting results. This means that each independent variable (past harvests, cost of fertilizer, household size, quantity of local seed, quantity of maize sales, and number of goats) depicts distinct variances between the two groups. *Past Harvests:* the findings depict that the past harvest for below-average producers is 5.36 kgs, whereas for above-average producers, it is relatively higher at 17.12 kgs. In this perspective, the results propose that the past harvest performance is a strong gauge of the present yield likely. *Cost of Fertilizer:* small producers who are classified as below-average incur an average cost of 43,805.01 Tanzanian shillings (Tshs), whereas above-average producers spend 149,062.87 Tshs. This discrepancy shows that the more efficient farmers may have a greater chance to invest in fertilizers, and in turn, contribute to high yields. *Household size*: The findings show that the mean household size for below-average small farmers is nearly 5.78, whereas above-average producers have a somewhat larger household size of 6.48. The findings may suggest that larger households can contribute more labor for farming. *Quantity of Local Seed*: the findings indicate that the below-average small farmers use an average of 4.24 kgs of local seed, while above-average smallholders use 3.53 kgs. The results suggest that above-average producers may use higher-quality seeds or more efficient practices, even with less quantity. *Quantity of maize sales:* there is a stark contrast in maize sales: below-average producers sell an average of 1.62 kgs, while above-average producers sell 12.23 kgs. This substantial difference highlights the economic implications of production efficiency and market access. Number of goats: the findings indicate that the average number of goats is 1.12 for below-average producers and 2.30 for above-average producers respectively. These results propose that livestock ownership might be associated with better productivity. This in turn increases the additional in-flow for boosting farming activities.

The QDA results show that both production-related factors and household characteristics significantly distinguish between below- and above-average maize yield groups. Past harvest and fertilizer expenditure were the strongest discriminators (Wilks' Lambda = 0.54 and 0.46, $p < 0.001$), emphasizing the role of input use. Maize sales and household size also contributed, reflecting market orientation and potential labor supply. Local seed use had a moderate effect ($p < 0.05$), while goat ownership was marginally significant. Higher fertilizer spending among better-performing farmers may reflect not just investment capacity, but also improved access to credit, markets, or extension services, underscoring systemic access disparities.

While the results show that above-average farmers spend significantly more on fertilizer, this difference may not only reflect greater investment willingness or agronomic knowledge, but also underlying differences in access to credit, extension services, or input markets. Farmers with better financial access or geographic proximity to input suppliers may find it easier to acquire and apply fertilizers. These structural advantages may amplify productivity independently of intrinsic farming practices, suggesting that input access

constraints, not just individual decisions, shape yield outcomes. On the other hand, larger households may offer more labor for farm activities, they also consume more resources, potentially straining food and income. Therefore, the impact of household size on productivity depends on the balance between labor contribution and consumption needs.

Based on the training sample, the study has revealed the highest positive predictive value results/posterior mean distribution to be high compared to the net predictive value in the training sample. This represents the average value of the parameter of the updated distribution given the observed measurement from the selected input variables. In the quadratic classification estimated results, the posterior mean for each class represents the estimated probability of that class given the input data where most of the sampled smallholder farmers were found to produce maize yield below the average. This is a class having the highest estimated probability. The highest posterior mean or highest estimated probability results imply that smallholder farmers are likely to produce below the average and hence fail to meet expectations. The posterior mean intends to detect the group that the model guesses to be the most likely or greatest plausible result for the selected predictors. Given that assessing the model fit using the test set, the findings indicated that both sensitivity and specificity are also used. Since their values are close to one for the cases being correctly classified below and above the average, respectively, there is a low rate of misclassification (Kaivanto, 2008).

Furthermore, the apparent error rate was further compared to the out-of-sample (test set) to assess the accuracy of the selected model. The results have discovered the smallest expected error (12.7%). The lowest error rate signifies that the sample spilt results on out–of–sample is effective in increasing the accuracy results of the classifier quadratic model (Shmuel, 2010). On the other hand, the current findings, based on the sample partition based on the test, support the claim that the out-of-sample (test) performance is reliable as benchmarked to the practical results depicted by the training set (Montesinos López *et al.,* 2022).

This is interesting because it addresses a core concern in predictive modeling: whether a model trained on one dataset (training set) can generalize well to new, unseen data (test set). When the sample partition shows that the QDA model performs similarly on the test data as it does on the training data, it indicates the model has good out-of-sample reliability and is not overfitting. This is interesting because it demonstrates that the model's predictive power is not limited to the training sample; it generalizes well to new observations. The alignment between test and training results supports the model's robustness and practical utility, strengthening confidence in using the QDA model to inform real-world policy or decision-making for improving maize yields.

In terms of generalized results, such as low rates of yield, the current study findings are comparable to the past studies (Akudugu *et al.,* 2012; Wiggins, S. & Keats, 2013; Woniala, J, Nyombi, 2014) as reported that smallholder farmers are faced with low crop yield in developing countries. However, the previous studies did not discuss the question of how likely the study sample of small farmers is to fail to meet the highest expectation of maize productivity. Therefore, the present study has added value to the body of knowledge including the analysis with likelihood analysis results on the status of the maize productivity. On the performance of the classifier based on the training, the current study results have been judged by the apparent error rates (13.3%). The current results do not conquer with previous findings as revealed by

(Olarinde *et al.,* 2010). Importantly, the present study has been well generalized as a resulting model of the sample split approach.

Generally, the categorization accuracy directs the model's effectiveness in predicting productivity categories. This high accuracy proposes that the proposed quadratic discriminant model can be reliably used to identify farmers who may need support or targeting interventions aimed at improving the performance of these farmers. In this present paper, the results revealed a high likelihood that producers identified as below average truly are in that category. This enhances confidence in targeting support programs effectively

7. Conclusion

The present study was designed to find out the extent to which the sample splitting approach minimizes the loss function of the quadratic supervised classifier by considering the case of maize yield in Tanzania. Furthermore, it has been revealed that the discriminatory variables (past harvest in kilograms, quantification of fertilizer in kilogram, quantity of maize sale, quantity of the local seed, household size, and several goat rearing) that were retained aided in classifying the smallholder farmers to the respective classes of being below. Through these retained subsets of variables, it has been discovered that quadratic discriminant machine learning has improved its performance.

Given the methodological (proposed discriminant technique) justification, the findings indicated a test set to outperform compared to the training set. On the other hand, the present study has revealed an interesting result indicating small farmers are likely to fail to meet their expectations because the positive predictive value or posterior mean is high with the test sample where loss function was found to decrease. This is an indication that Also, more of the study participants produced below their expectations, which was not documented in the previous studies. Again, the results discovered that the selected inputs add value by updating the belief having been evidenced from the previous studies.

Along with the discovered results, however, the present study has come up with an answer to what factors can contribute to better yield for small farmers. In this regard, the past yield, improved seed, quantity of fertilizers used, household size of the source of manpower, volume of sales of the harvest, and number of goats as a source of manure and income diversification via selling the goats.

The present study is important to farmers and policy practitioners because of the following key issues: Initially, identifying Performance Gaps: by classifying farmers according to their yield performance, the paper has uncovered the elements that lead to either low or high yields. This insight can enable farmers to comprehend their position in comparison to their counterparts and identify particular areas of their agricultural practices that may require enhancement. On the other hand, the above-average producers are investing more in inputs, suggesting a “rich get richer” effect, where wealthier farmers continue to advance while poorer ones fall behind. To address this, policies should promote Inclusive strategies that are essential to ensure all farmers benefit from growth.

Secondly, customized agricultural practices: the findings obtained from quadratic modeling can assist in formulating specific recommendations for various groups of farmers. For instance, those with yields below the average may need targeted interventions, including optimized crop rotation methods, improved soil

management strategies, or advanced irrigation techniques. Conversely, farmers achieving above-average yields can be analyzed to uncover best practices that could be implemented by their peers. Thirdly, data-informed decision-making: The findings from the research offer farmers tangible, data-informed insights that can guide their choices regarding crop selection, resource distribution, and investment in innovative technologies or practices. This approach can enhance risk management and facilitate more strategic planning within their operations. With these results, the present study recommends the policy intervention cutting across the underlined focus.

Economic recommendations: To improve smallholder maize productivity in Tanzania, interventions should address the key factors of past harvest, fertilizer cost, household size, local seed use, maize sales, and goat ownership. Expanding input subsidies through programs like NAIVS can ease cost barriers, while strengthening community seed systems can boost access to quality local varieties. Promoting crop-livestock integration and improving post-harvest practices can enhance sustainability and food security. Tailored extension services and improved market access through infrastructure or digital tools are also crucial. These efforts should align with national strategies such as ASDP II for effective implementation. However, including digital-based innovations could further strengthen the classification modeling with minimal risk, improving precision and implementation efficiency.

Also, it noted that the variation revealed in terms of the identified key dominant discriminating variables suggests that producers above average are likely to invest more in better farm inputs. This, in turn which in turn leads to better harvests. In other words, better farm inputs influence households to enhance their overall productivity and sales. Therefore, it recommends addressing these identified gaps to improve the efficiency of the producers below average. Based on the sample division technique, on the other hand, the following are important to be in place: first is the targeted Interventions; on identified high sensitivity and predictive accuracy, the study recommends allowing agricultural agencies to focus resources on producers classified as below average. This can improve their likelihood of increasing productivity as well as the economic practicability. Secondly, the need for resource allocation is greatly important, this is because the findings have revealed the category of farmers who are struggling. Therefore, the allocation of resources in terms of subsidies, training programs, and access to better seeds and fertilizers is key.

Declaration statements

Competing interest: There is no conflict of interest

Funding: Author has no
Acknowledgment: A special thanks should go to the small farmers in the studied area for their willingness to respond to survey questions supported by Mzumbe University between October 2015 to mid of February 2016.

Data availability: Data may be available upon request.

References

Adimassu, Z., & Kessler, A. (2016). Factors affecting farmers ' coping and adaptation strategies to perceived trends of declining rainfall and crop productivity in the central Rift valley of Ethiopia. *Environmental Systems Research*. https://doi.org/10.1186/s40068-016-0065-2

Ahmed, B., Haji, J., & Geta, E. (2013). Analysis of farm households' technical efficiency in the production of smallholder farmers: the case of Girawa district, Ethiopia. American-Eurasian *Journal of Agricultural & Environmental Sciences*, 13(12), 1615–1621. https://doi.org/10.5829/idosi.aejaes.2013.13.12.12310

Akudugu, M. A., Guo, E., & Dadzie, S. K. (2012). Adoption of Modern Agricultural Production Technologies by Farm Households in Ghana: What Factors Influence their Decisions? *Journal of Biology, Agriculture and Healthcare*, 2(3), 1–13. https://www.researchgate.net/profile/Samuel_Dadzie/publication/235751741_Adoption_of_modern_agricultural_production_technologies_by_farm_households_in_Ghana_What_factors_influence_their_decisions/links/00463533b1249ebdf1000000.pdf

Alhassan, A., Salifu, H., & Adebanji, A. O. (2016). Discriminant analysis of farmers adoption of improved maize varieties in Wa Municipality, Upper West Region of Ghana. *SpringerPlus*, 5(1). https://doi.org/10.1186/s40064-016-3196-z

Andersson Djurfeldt, A., Djurfeldt, G., Hillbom, E., Isinika, A. C., Joshua, M. D. K., Kaleng'a, W. C., Kalindi, A., Msuya, E., Mulwafu, W., & Wamulume, M. (2019). Is there such a thing as sustainable agricultural intensification in smallholder-based farming in sub-Saharan Africa? Understanding yield differences in relation to gender in Malawi, Tanzania and Zambia. *Development Studies Research*, 6(1), 62–75. https://doi.org/10.1080/21665095.2019.1593048

Barrett, C. B., Bellemare, M. F., & Hou, J. Y. (2010). Reconsidering conventional explanations of the inverse productivity–size relationship. *World Development*, 38(1), 88–97.

Becker, G. S. (1964). Human Capital: *A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.

Chirwa, E., et al. (2025). Cost-benefit analysis of climate-smart agricultural practices in Southern Africa: Gender considerations and adoption challenges. *Frontiers in Sustainable Food Systems*.https://doi.org/10.3389/fsufs.2025.1499982

Coase, R. H. (1960). The Problem of Social Cost. *Journal of Law and Economics*, 3, 1-44.

Cobb, C. W., & Douglas, P. H. (1928). A Theory of Production. *The American Economic Review*, 18(1), 139-165.

El Ouardighi, A., El Akadi, A., & Aboutajdine, D. (2007). Feature selection on supervised classification using Wilk's Lambda statistic. ISCIII'07: 3rd International Symposium on Computational *Intelligence and Intelligent Informatics; Proceedings*, 51–55. https://doi.org/10.1109/ISCIII.2007.367361

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

Epule, T. E., Chehbouni, A., & Dhiba, D. (2022). Recent Patterns in Maize Yield and Harvest Area across Africa. *Agronomy*, 12(2). https://doi.org/10.3390/agronomy12020374

Erenstein, O., Jaleta, M., Sonder, K., Mottaleb, K., & Prasanna, B. M. (2022). Global maize production, consumption and trade: trends and R&D implications. *Food Security*, 14(5), 1295–1319. https://doi.org/10.1007/s12571-022-01288-7

Fall, F., Ky, Y., & Birba, O. (2015). Analyzing the Mobile-Banking Adoption Process among Low-Income Populations : A Sequential Logit Model To cite this version : HAL Id : halshs-01225149 Volume 35 , Issue 4 Analyzing the Mobile-Banking Adoption Process among Low-Income Populations : A Sequen. 35(4), 2085–2013.

FAO. (2010). Global Strategy to Improve Agricultural and Rural Statistics Report of the Friends of the Chair on Agricultural. 3(February).

FAO. (2023). Agricultural production statistics Agricultural production statistics 2000-2022. Faostat Analytical Brief 79. https://www.fao.org/faostat/en/#data/QCL

Friendly, M., & Sigal, M. (2020). Visualizing Tests for Equality of Covariance Matrices. *American Statistician*, 74(2), 144–155. https://doi.org/10.1080/00031305.2018.1497537

GRIN Publishing. https://www.grin.com/document/1081232

Härdle, W., & Simar, L. (2003). Applied Multivariate Statistical Analysis. Applied Multivariate Statistical Analysis, October. https://doi.org/10.1007/978-3-662-05802-2

Haug, R., & Hella, J. (2013). The art of balancing food security: Securing availability and affordability of food in Tanzania. *Food Security*, 5(3), 415–426. https://doi.org/10.1007/s12571-013-0266-8

Jiamwattanapong, K., & Ingadapa, N. (2021). On Testing Homogeneity of Covariance Matrices with Box's M and the Approximate Tests for Multivariate Data. *Advances in Image and Video Processing*, 9(5). https://doi.org/10.14738/aivp.95.11115

Joseph, V. R. (2022). Optimal ratio for data splitting. Statistical Analysis and Data Mining, 15(4), 531–538. https://doi.org/10.1002/sam.11583

Joseph, V. R., & Vakayil, A. (2022). SPlit: An Optimal Method for Data Splitting. Technometrics, 64(2), 166–176. https://doi.org/10.1080/00401706.2021.1921037

Kaivanto, K. (2008). Maximization of the sum of sensitivity and specificity as a diagnostic cutpoint criterion. *Journal of Clinical Epidemiology*, 61(5), 517–518. https://doi.org/10.1016/j.jclinepi.2007.10.011

Kakar, M., Kiani, A., & Baig, A. (2016). Determinants of Agricultural Productivity: Empirical Evidence from Pakistan's Economy. *Global Economics Review*, I(I), 1–12. https://doi.org/10.31703/ger.2016(i-i).01

Korjus, K., Hebart, M. N., & Vicente, R. (2016). An efficient data partitioning to improve classification performance while keeping parameters interpretable. *PLoS ONE*, 11(8), 1–16. https://doi.org/10.1371/journal.pone.0161788

Maindonald, J. H. (2012). *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery* by Graham Williams. In International Statistical Review (Vol. 80, Issue 1). https://doi.org/10.1111/j.1751-5823.2012.00179_23.x

Martin, T. M., Harten, P., Young, D. M., Muratov, E. N., Golbraikh, A., Zhu, H., & Tropsha, A. (2012). Does the rational selection of training and test sets improve the outcome of QSAR modeling? *Journal of Chemical Information and Modeling*, 52(10), 2570–2578. https://doi.org/10.1021/ci300338w

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez G., J., Buckler, E., & Doebley, J. (2002). A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences*, 99(9), 6080–6084. https://doi.org/10.1073/pnas.052125199

Me-Nsope, N., & Larkins, M. (2016). This document is discoverable and free to researchers across the globe due to the work of AgEcon Search. Help ensure our sustainability. In *Journal of Gender, Agriculture and Food Security* (Vol. 1, Issue 3).

Mgonja, M. A., et al. (2017). The impact of agricultural input subsidies on maize productivity in Tanzania: Evidence from smallholder farmers. *African Journal of Agricultural Research*, 12(9), 745-754. https://doi.org/10.5897/AJAR2016.11674

Mkonda, M. Y., & He, X. (2018). Agricultural history nexus food security and policy framework in Tanzania. *Agriculture & Food Security*, 1–11. https://doi.org/10.1186/s40066-018-0228-7

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Multivariate Statistical Machine Learning Methods for Genomic Prediction. In Multivariate Statistical Machine Learning Methods for Genomic Prediction. https://doi.org/10.1007/978-3-030-89010-0

Morais, C. L. M., & Lima, K. M. G. (2018). Principal component analysis with linear and quadratic discriminant analysis for identification of cancer samples based on mass spectrometry. *Journal of the Brazilian Chemical Society*, 29(3), 472–481. https://doi.org/10.21577/0103-5053.20170159

Mupangwa, W., Chipindu, L., Nyagumbo, I., Mkuhlani, S., & Sisito, G. (2020). Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in Eastern and Southern Africa. *SN Applied Sciences*, 2(5), 1–14. https://doi.org/10.1007/s42452-020-2711-6

Nuno, D. B., & Baker, M. M. (2021). The Determinants of Agricultural Crop Productivity among Smallholder Households in Haramaya Distinct, Eastern Ethiopia. *Grassroots Journal of Natural Resources*, 04(04), 146–153. https://doi.org/10.33002/nr2581.6853.040410

Obasi, P. (2013). Factors Affecting Agricultural Productivity among Arable Crop Farmers in Imo State, Nigeria. *American Journal of Experimental Agriculture*, 3(2), 443–454. https://doi.org/10.9734/ajea/2013/2030

Olarinde, L. O., Manyong, V. M., & Akintola, J. O. (2010). Factors influencing risk aversion among maize farmers in the Northern Guinea Savanna of Nigeria: Implications for sustainable crop development programmes. *Journal of Food, Agriculture and Environment*, 8(1), 128–134.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Pingali, P., & Sunder, N. (2017). Transitioning toward nutrition-sensitive food systems in developing countries. *Annual Review of Resource Economics*, 9, 439–459

Piperno, D. R., & Flannery, K. V. (2001). The earliest archaeological maize (Zea mays L.) from highland Mexico: New accelerator mass spectrometry dates and their implications. *Proceedings of the National Academy of Sciences*, 98(4), 2101–2103. https://doi.org/10.1073/pnas.98.4.2101

Ranum, P., Peña-Rosas, J. P., & Garcia-Casal, M. N. (2014). Global maize production,

Raykov, T., & Marcoulides, G. A. (2008). An Introduction to Applied Multivariate Analysis. In An Introduction to Applied Multivariate Analysis. https://doi.org/10.4324/9780203809532

Ribeiro-Duthie, A. C., Gale, F., & Murphy-Gregory, H. (2021). Fair trade and staple foods: A systematic review. *Journal of Cleaner Production*, 279, 123586. https://doi.org/10.1016/j.jclepro.2020.123586

Saitoti, S., & Ngalawa, H. (2020). Challenges in implementing fertilizer subsidy programs in Tanzania: A review. *Journal of Development and Agricultural Economics*, 12(4), 183-190. https://doi.org/10.5897/JDAE2020.1181

Shi, W., & Tao, F. (2014). Vulnerability of African maize yield to climate change and variability during 1961-2010. Food Security, 6(4), 471–481. https://doi.org/10.1007/s12571-014-0370-4

Shita, A., Kumar, N., & Singh, S. (2018). Determinants of Agricultural Productivity in Ethiopia: ARDL Approach. The Indian Economic Journal, 66(3–4), 365–374. https://doi.org/10.1177/0019466220941418

Shmueli, G. (2010). To explain or to predict? Statistical Science, 25(3), 289–310. https://doi.org/10.1214/10-STS330

Solomon, H. (2019). The effect of formal education on the productivity of maize farmers in Ethiopia: Evidence from a Cobb–Douglas production function.

Srivastava, S., Gupta, M. R., & Frigyik, B. A. (2007). Bayesian quadratic discriminant analysis. Journal of Machine Learning Research, 8, 1277–1305.

Székely, G. J., & Rizzo, M. L. (2013). Journal of Statistical Planning and Inference Energy statistics : A class of statistics based on distances. Journal of Statistical Planning and Inference, 143(8), 1249–1272. http://dx.doi.org/10.1016/j.jspi.2013.03.018

Tesfaye, K., Gbegbelegbe, S., Cairns, J. E., Shiferaw, B., Prasanna, B. M., Sonder, K., Boote, K., Makumbi, D., & Robertson, R. (2015). Maize systems under climate change in sub-Saharan Africa: Potential impacts on production and food security. International Journal of Climate Change Strategies and Management, 7(3), 247–271. https://doi.org/10.1108/IJCCSM-01-2014-0005

The Tanzania National Sample Census of. (2011). 4.

Urassa, J. K. (2010). Rural household livelihoods , crop production and well- being after a period of trade reforms : a case study of Rukwa , Tanzania Thesis submitted to the University of Sussex in partial fulfilment of the requirements.

Urgessa, T. (2015). The Determinants of Agricultural Productivity and Rural Household Income in Ethiopia. Ethiopian Journal of Economics, 24(2), 63–91.

Utilization, and consumption. Annals of the New York Academy of Sciences, 1312(1), 105–112. https://doi.org/10.1111/nyas.12396

Wiggins, S. & Keats, S. (2013). Smallholder agriculture's contribution to better nutrition. Report for the Hunger Alliance. March.

Woniala, J, Nyombi, K. (2014). Soil Fertility Management by Smallholder Farmers and the Impact on Soil Chemical Properties in Sironko District, Uganda. Research Journal of Agriculture and Forestry Sciences, 2(1), 5–10. http://www.isca.in/AGRI_FORESTRY/Archive/v2/i1/2.ISCA-RJAFS-2013-070.pdf

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. Journal of Analysis and Testing, 2(3), 249–262. https://doi.org/10.1007/s41664-018-0068-2

Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2011). Statistical Methods in Diagnostic Medicine (2nd ed.). Wiley.