

Using pattern mining to determine fine climatic parameters for maize yield in Benin

Souand Peace Gloria Tahi¹, Vinasetan Ratheil Houndji², Castro Gbêmêmali Hounmenou¹,
Romain Glèlè Kakai¹

¹Laboratoire de Biomathématiques et d'Estimations Forestières, Faculty of Agronomic Sciences, University of Abomey-Calavi, Cotonou, Benin

²Institut de Formation et de Recherche en Informatique, University of Abomey-Calavi, Cotonou, Benin

Article Info

Article history:

Received Dec 3, 2023

Revised Mar 29, 2024

Accepted Apr 17, 2024

Keywords:

Association rules

Climatic pattern

Machine learning

Yield prediction

Zea mays

ABSTRACT

This study investigates the relationships between Benin's climate and maize production to develop an association rule algorithm for accurate yield prediction. The datasets utilized extend 26 years (1995 to 2020) and include climate and maize yield data from five districts with synoptic weather stations in two agroclimatic zones (Sudanian and Sudano-Guinean). Climate variables were combined with yield using "year" and "districts" to find the association rules. Several techniques were used to determine the correlation between weather parameters and maize yields: support vector machines, K nearest neighbor, artificial neural networks, decision trees, and recurrent neural networks. The most performed method was the decision tree ($R^2=0.998$, mean squared error (MSE)=0.021, and mean absolute error (MAE)=0.0008). This model is difficult to understand, though the frequent pattern growth technique was then applied to the dataset to facilitate the discovery of the rules. The Sudano-Guinean zone exhibits high maize yields for medium minimum and maximum temperature values, rainfall, evapotranspiration, and humidity. In the Sudanian zone, medium minimum and maximum temperatures and maximum humidity levels are associated with high maize yields. The discovered association rules showed that optimizing maize output might be done dependably and effectively.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Souand Peace Gloria Tahi

Laboratoire de Biomathématiques et d'Estimations Forestières, Faculty of Agronomic Sciences

University of Abomey-Calavi

04 BP 1525 Cotonou, Bénin

Email: souandtahi@gmail.com

1. INTRODUCTION

Agriculture is the primary source of economic growth in most developing countries. Achieving food security requires, among others, an increase in crop production. However, several factors limit crop yields: climate, soil composition, fertilizer use, resource availability, and political and socioeconomic variables [1]. Nevertheless, climate change remains the primary cause restricting agricultural yield [2]. It led to significant disruptions in certain regions [3], thereby affecting agricultural productivity. Nowadays, it is no longer possible for a farmer to produce without knowing the yield forecast, which suggests adopting smart agriculture. Statistical and computational techniques are increasingly used by farmers, agronomists, and policymakers for decision-making [4]. For this reason, predicting climate change impacts on crop yields, economy, and the environment from high-precision models becomes necessary [5].

Crop yield trends usually follow the non-linear process. Machine learning (ML) may then be considered as an appropriate method for modeling crop yields [6], [7]. Indeed, the program uses artificial intelligence to make the correlations between the input and output variables non-linear [8]. Nowadays, supervised ML techniques have gained significant recognition in crop yield prediction, leveraging sophisticated models such as deep neural networks [9]–[11]. These advanced methods excel at capturing intricate patterns within agricultural data. Nevertheless, these methods are intricate, and their complexity lies in their intensive training requirements and implementation, often necessitating large datasets and substantial computational resources. Despite their effectiveness, their lack of interpretability can pose a challenge. Interpreting results remains a significant challenge with ML techniques. One such example is decision trees (DT), which offer some interpretability by producing simple rules in the form of a tree. But, the more complex the tree, the harder it is to interpret. However, there is untapped potential for other ML techniques, such as association rules, which are underused in crop yield prediction. This paper focuses on the association rules algorithms, used to find unseen or desired patterns in large datasets. Widely used in other fields such as transactional datasets [12], association rules offer the possibility of discovering correlations between various agronomic, meteorological, and environmental parameters, to find the relations among variables. This study bridges the gap between the present-day status of the art in crop yield prediction and the opportunities offered by association rules.

There are several algorithms to search association rules, such as Apriori (AA), eclat, frequent pattern growth (FPG), and rapid association rule mining (RARM). Some previous works applied pattern-finding techniques in agriculture. For example, Kaur and Attwal [13] used association rules techniques to evaluate the effect of temperature and precipitation on rice yield in India. They concluded that rice yield depends on temperature. In Kenya, Silas and Nderu [14] used the Apriori algorithm to predict tea production and concluded that high production is likely to extend into January and December. In Pakistan, Supro [15], using association rules and artificial neural network (ANN), revealed that rice yield is high when the temperature is high, and humidity is medium. Moreover, the relationship between temperature, rainfall, soil nitrogen, and rice production was ascertained by Rao *et al.* [1] using Apriori, Eclat, and AprioriTid. They found that the Apriori algorithm outperformed the other two algorithms to find the relationship between weather parameters and maize yields. Salankar *et al.* [16] used Apriori, FPG, and clustering to propose a solution for crop suggestion at a crop set. They observed that the proposed approach effectively suggested the best crop to grow based on environmental factors.

This paper investigates the association rules between climate parameters and yield to predict maize yield in Benin (West Africa). Maize is indeed an important cereal crop in people's diets worldwide, particularly in West Africa [17]. Annual maize consumption in Benin is 55 kg/capita [18], ranking the country at the top of countries with high maize consumption. However, following the effects of climate change, a gradual decline in yields and production volume has been observed in recent years [18]. It is therefore urgent to determine the rules between climate parameters and yield. For this purpose, i) maize yield is predicted from different supervised ML methods, ii) the dataset is augmented from the best-identified method, and iii) the generated data is used to establish the association rules from the FPG algorithm.

The rest of the paper is organized as follows. Section 2 presents the data collection methodology, generation, and analysis. Then section 3 deals with the main results obtained and discussion. Finally, section 4 focuses on the conclusion.

2. METHOD

2.1. Data collection and preprocessing

2.1.1. Dataset considered

Secondary data from two agroclimatic zones in Benin (Sudanian and Sudano-Guinean) are used to assess climatic factors and maize yield relationships in Benin. These data were obtained over 26 years from 1995 to 2020 in five synoptic weather stations located in the Sudano-Guinean zone (Bohicon and Save districts) and Sudanian zone (Parakou, Natitingou, and Kandi districts). Data cover variables such as evapotranspiration (mm), sunstroke (h), rainfall (mm), minimum and maximum temperatures (°C), minimum and maximum humidity (%), and overall maize yield (kg/ha). The climate data (temperature, humidity, rainfall, sunstroke, evapotranspiration) are obtained at the "Agence pour la Sécurité de la Navigation Aérienne en Afrique (ASECNA)" in Benin. The annual maize yield data linked to the districts are obtained at the "Direction des Statistiques Agricoles (DSA)". Furthermore, daily data on climate variables and annual yields are also obtained from DSA. In Benin, maize is cultivated from mid-April to mid-October (7 months). Hence, only data covering the production period are collected, 1,503 daily climate records. From the daily data, averaged monthly climate data is aggregated to yield data for each year and the district. Finally, 875 observations are utilized for the analysis. Figure 1 displays the distribution of the climate parameters by districts. Evapotranspiration was high in the Kandi district shown in Figure 1(a), while rainfall was higher in

the Natitingou and Kandi districts shown in Figure 1(b). Sunstroke shown in Figure 1(c) and temperature shown in Figures 1(d) and 1(e) over the last 26 years were also high in the Kandi district. Bohicon and Savè had the highest maximum humidity shown in Figure 1(f), while Natitingou, Savè, and Bohicon recorded the highest minimum humidity levels shown in Figure 1(g). Regarding maize yield, Kandi had the highest value, followed by Bohicon and Natitingou shown in Figure 1(h).

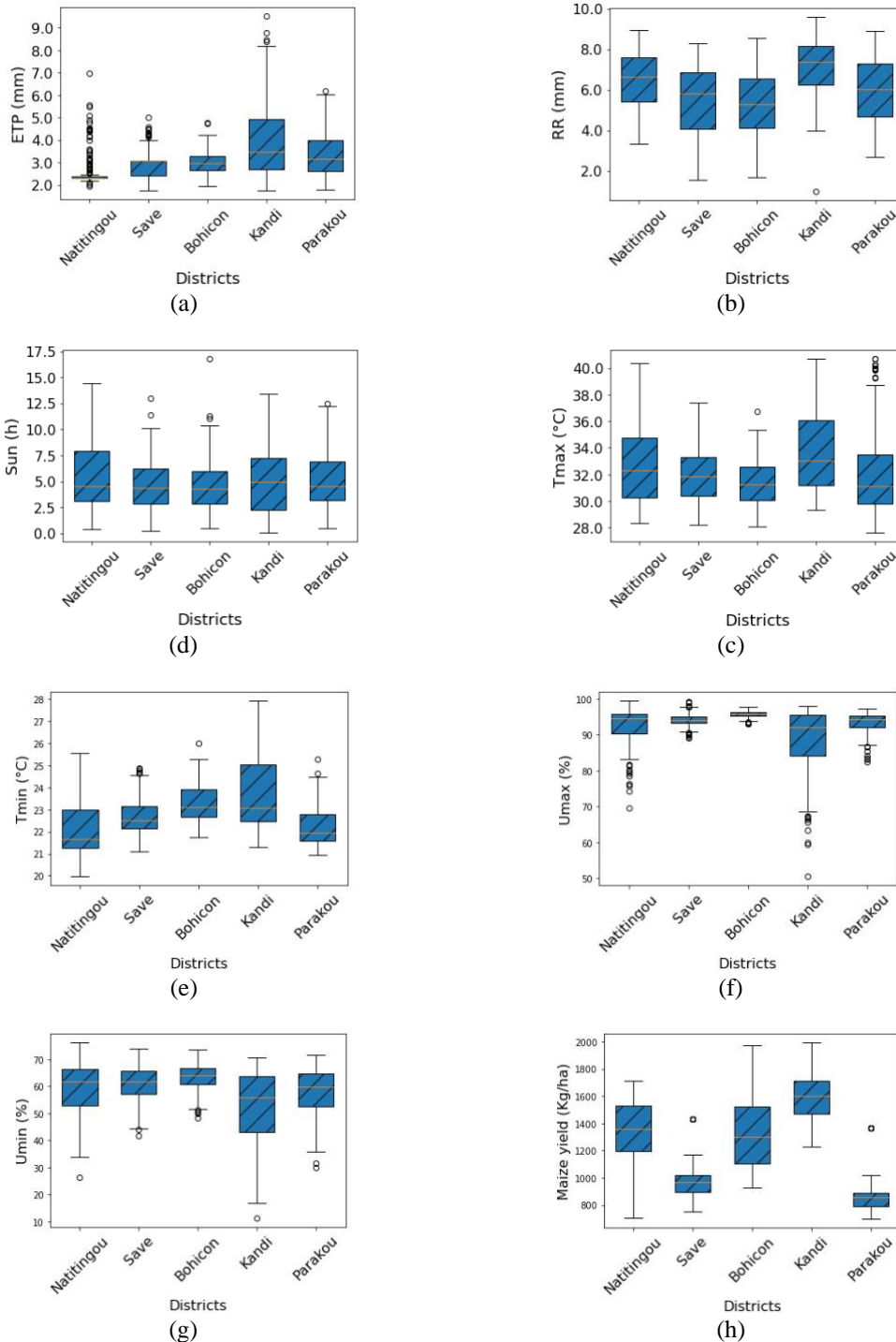


Figure 1. The distribution of climate and yield parameters for the 1995-2020 period linked to each district. (a) evapotranspiration, (b) rainfall, (c) sunstroke, (d) maximum temperature, (e) minimum temperature, (f) maximum humidity, (g) minimum humidity, and (h) Maize yield

a) Maize yield dataset

The maize yield trend by hectare for 26 years is similar for Bohicon, Kandi, and Natitingou. This is a downward trend for Kandi, with the peak noted in 1997 (2,000 kg/ha), while Bohicon presents a contrasting trend with 3 peaks in 1997 (1,480 kg/ha), 2007 (1,630 kg/ha), and 2015 (1,900 kg/ha) ha), respectively shown in Figure 2. Maize yield in Natitingou also shows a bimodal trend, with peaks in 2007 (1,700 kg/ha) and 2012 (1,800 kg/ha). Parakou and save recorded the lowest maize yield shown in Figure 2.

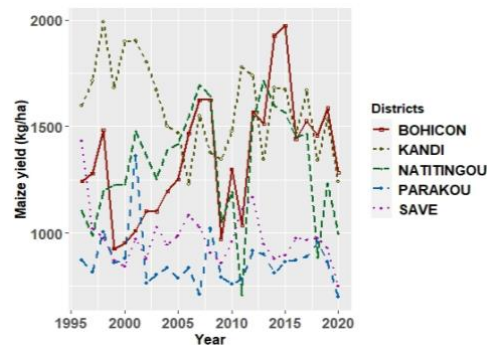


Figure 2. Temporal variation in maize yield in Kandi, Parakou, Save, Natitingou, and Bohicon

b) Trend of climatic parameters in the production season from 1995 to 2020 in the study areas

During the last 26 years, Kandi has registered a high potential for evapotranspiration compared to the other districts shown in Figure 3(a). A marked increase was recorded in 1995 and 2000, whereas a pronounced decline was registered in 2001. It was constant from 1995 to 2011 (ETP=2.286 mm) before reaching a peak of 5.1 mm in 2020. Overall, Kandi showed the highest evapotranspiration values, whereas Bohicon recorded the lowest trends. Regarding the rainfall, similar trends were noticed for the districts considered. Parakou and Natitingou recorded relatively high rainfall values, whereas save and Bohicon show the lowest trends shown in Figure 3(b). Sunstroke shows the same trends as evapotranspiration shown in Figure 3(c). From 1995 to 2000, Save, Bohicon, Kandi, and Parakou recorded the lowest maximum temperature values` under 25 °C, and similar and fluctuating trends between 32 °C and 35 °C from 2000 to 2020 shown in Figure 3(d). Regarding the minimum temperatures, Save, Natitingou, and Parakou show similar trends over the agricultural seasons for the last 26 years, whereas Bohicon and Kandi recorded similar fluctuations shown in Figure 3(e). However, the minimum temperatures remained higher in Kandi compared to other districts. The trend was relatively high for the Bohicon district regarding the maximum and minimum humidity. During 2013, maximum humidity in Savè peaked at 96.64% shown in Figure 3(f), while minimum humidity peaked in 2011 and 2014 shown in Figure 3(g). Similarly, the maximum and minimum humidity were relatively low in the Kandi district. In Natitingou, maximum humidity has been progressively decreasing since 2008.

2.1.2. Data generated

New weather and yield data were generated for each of the five districts in agro-climatic zones to provide reliable association rules for high maize yield. This simulation used the supervised learning model with $R^2=0.99$ and $MAE=0.0008$. Thus, the input variables of generated data (climate parameters) follow distributions and parameters similar to those in the actual data. Ten thousand (10,000) observations have been generated for each district in both agro-climatic zones. Data variables were encoded in ordinal (low, medium, and high) based on the fixed threshold values. Observations with high yield were selected from each generated database to build the rules. The threshold values for all variables are given by the agroecological zone in Table 1.

2.1.3. Data preprocessing

The rules in the two agro-climatic zones were obtained by applying supervised and unsupervised ML models. The supervised models optimize maize yield, find suitable relationships between climate parameters and yield, and identify the best model. The best model identified was not easily interpretable. Thus, this model with random parts was used to complete the available data. Therefore, the unsupervised model was used to determine the rules that are easily interpretable and understandable.

Before performing the supervised modeling, a correlation analysis between independent and dependent variables was performed. No climatic variables were correlated to maize yield. An additional correlation analysis was performed among the independent variables. The threshold of correlation was fixed to 80%. Using this threshold, the minimum humidity was eliminated from the predictors. Missing daily data was subjected to simple imputation before aggregation, and the outliers were subsequently removed following the interquartile range approach. Data were then standardized and partitioned using 5-fold cross-validation. The models were trained on 700 observations and validated on the remaining 175 data.

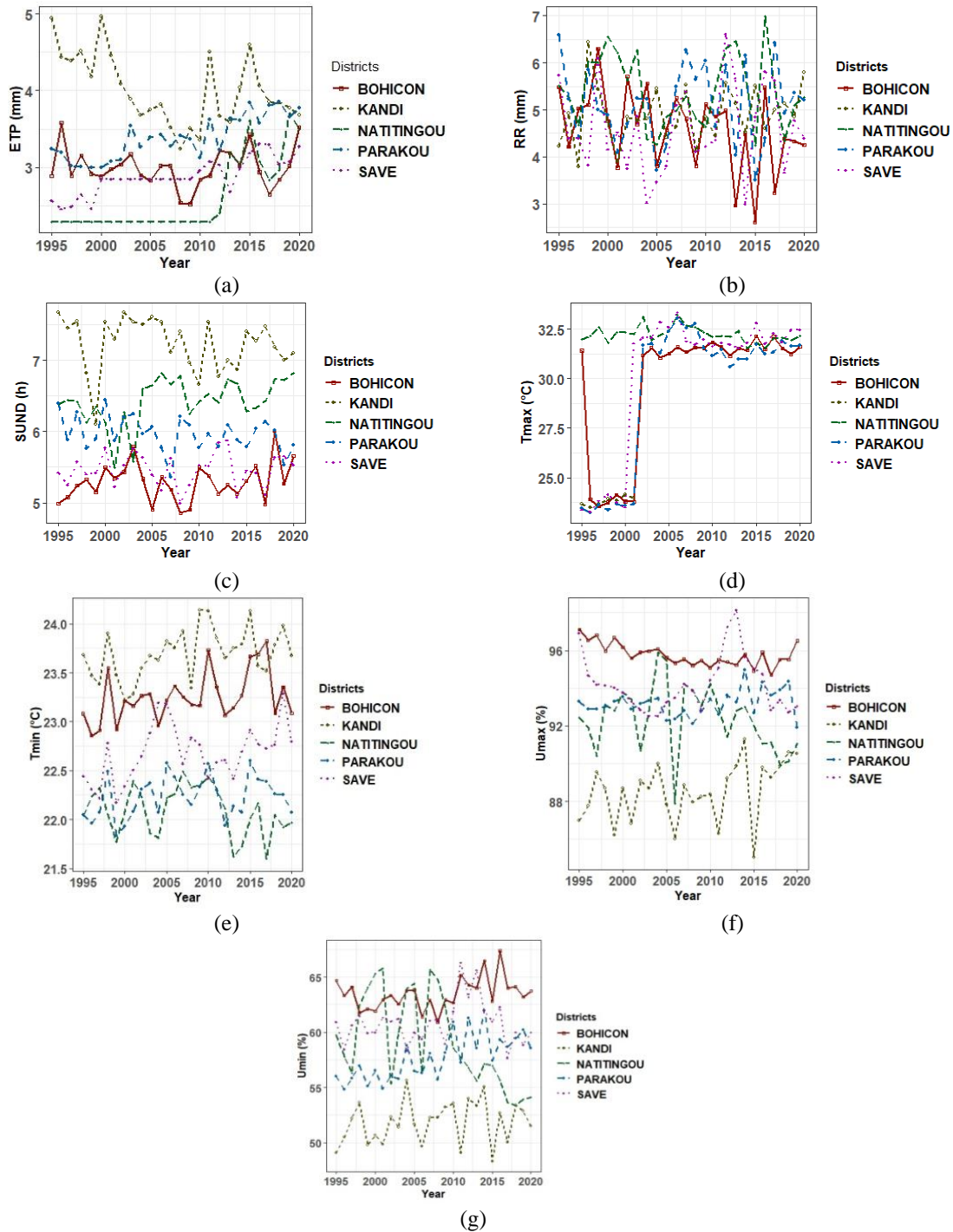


Figure 3. Variations in weather parameters over time in Kandi, Parakou, Save, Natitingou, and Bohicon. (a) evapotranspiration, (b) rainfall, (c) sunstroke, (d) maximum temperature, (e) minimum temperature, (f) maximum humidity, and (g) minimum humidity

Table 1. Threshold values for discretization

Attributes	Low	Medium	High
Sudanian zone			
ETP	<1.618	1.618-4.801	≥4.801
RR	<2.163	2.163-8.206	≥8.206
Sun	<5.221	5.221-7.883	≥7.883
Tmax	<29.371	29.371-35.971	≥35.971
Tmin	<21.337	21.337-24.015	≥24.015
Umax	<87.324	87.324-102.28	≥102.28
Yield	<1076.614	1076.614-1459.829	≥1459.829
Sudano-Guinean zone			
ETP	<2.513	2.513-3.628	≥3.628
RR	<2.314	2.314-6.767	≥6.767
Sun	<3.891	3.891-6.952	≥6.952
Tmax	<29.836	29.836-33.599	≥33.599
Tmin	<22.088	22.088-23.88	≥23.88
Umax	<93.497	93.497-96.302	≥96.302
Yield	<1016.311	1016.311- 1279.765	≥1279.765

2.2. Prediction of maize yields with supervised machine learning techniques

ML algorithms are widely renowned for their ability to model non-linear correlations between input and output variables. To understand the existing non-linear relationship between climate parameters and maize yield, five (5) supervised ML techniques were explored on the dataset in Python 3.8.11 software to find relationships between aggregated yield and weather features. These methods were k-nearest neighbor (KNN), support vector machine (SVM), DT, ANN with multilayer perceptron, and recurrent neural network (RNN) with long short-term memory. Input variables were standardized before modeling. The data was divided into training and testing using five-fold cross-validation. The various methods were assessed using the test data after being trained on the training set. The “Scikit-Learn” library was used for the KNN, SVM regression, and DT models, while “Keras” was used for the ANN and RNN models.

The fitting of the models to aggregate outputs was simple. This is because each output corresponds to a set of input vectors. The problem was resolved by generating an artificial output for each input vector. These artificial outputs were defined as the mean of the output value throughout the aggregate collections to which they belong. The following subsections provide an overview of the models used for the prediction.

2.2.1. Overview of the models used

As mentioned above, five algorithms were considered: KNN, SVM, DT, ANN, and RNN. These algorithms are a non-parametric ML method used to understand the non-linear link between the variables that are input and output. In KNN regression, prediction is the mean of the k nearest neighbor values [19]. SVM goal is to minimize the prediction bias by finding an excellent hyper-plane to reduce differences between predicted and observed values [20]. A DT is a supervised non-parametric learning approach that forecasts the value of a target variable by training on simple decision rules based on data characteristics. The ANN is based on conceptual neurons [21], [22]. It consists of the input, hidden, and output layers of neurons. In this work, ANN with multilayer perceptron architecture is used. RNN is also used; it is an ANN class where connections between the nodes form a graph along a time sequence [23]. RNN can process a long sequence of input variables using their memory. Its architecture consists of input layers, hidden layers, and output layers. Each layer is independent of the other. No layer stores the preceding outputs.

2.2.2. Comparison criteria

To evaluate the different supervised models used, the cross-validation (CV) technique was used. Due to its ease of use, applicability, and effectiveness in preventing the over-fitting problem, it is an often employed algorithm selection technique [24], [25]. The best model is widely agreed to be the one with the minimum estimation error. Using the coefficient of determination (R^2), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE), the models' performance was evaluated. R^2 indicates the linear relationship between the observed maize yield and the predictions. MAE measures the percentage of the mean deviation of the predicted maize yield from observations, while RMSE measures the deviation of the predicted maize yield from observations. MSE measures the square of the deviation from observations. The detailed formulas are defined as follows:

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y}_i)(f_i - \bar{f}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2 (f_i - \bar{f}_i)^2} \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - f_i)^2} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - f_i| \quad (3)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - f_i)^2 \quad (4)$$

y_i is the observed maize yield, \bar{y}_i is the corresponding mean value, n is the number of data points for the ML model. f_i is the predicted maize yield, \bar{f}_i is the corresponding mean value, and p is the total number of explanatory variables in the model. The closer R^2 is to 1 or the closer MSE, MAE, and RMSE are to 0, the better the model's prediction performance. All four metrics were used to choose the best model.

2.3. Prediction of maize yields with unsupervised machine learning methods

Association rules are defined by ML to establish the relationship between variables. The FPG algorithm is used to establish the rules from the “*mlxtend*” package. FPG is an improved version of the well-known Apriori algorithm with accurate extraction efficiency [26]. To do this, it sets a minimum threshold, which is used to remove irrelevant information in the data and form frequent itemsets. These later are used to assess the similarity degree of different variables and establish the association rules.

Input and output variables were encoded in low, medium, and high based on the fixed threshold values. Minimum support was fixed as 0.25. As for the threshold confidence, it was set to 0.75. The rules are then ordered and screened based on confidence and lift. After filtering, only the best rules from each agro-climatic zone were retained.

2.3.1. Criteria for choosing the best association rules

Different metrics were used to obtain the best rules, such as support, confidence, and lift. The support displays the proportion of transactions, including items X and Y. The confidence expresses how many transactions have X and Y. At the same time, the lift assesses their link. n represents the total number of transactions, $n(X \cup Y)$ is the number of transactions with X or Y, and $n(X)$ is the number of transactions with X. A rule is useful when the support and confidence are higher than the minimum support and minimum confidence threshold, and lift is greater or equal to 1.

$$Support(X \rightarrow Y) = P(X, Y) = \frac{n(X \cup Y)}{n} \quad (5)$$

$$Confidence(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \quad (6)$$

$$lift(X \rightarrow Y) = \frac{P(X, Y)}{P(X)P(Y)} \quad (7)$$

3. RESULTS AND DISCUSSION

3.1. Maize yield modeling

Table 2 presents the performance of the models using comparison metrics. Based on the results, all models had an MSE value less than 0.03 kg/ha. The DT technique recorded the lowest MAE and RMSE values (0.0008 kg/ha and 0.012 kg/ha, respectively). This conclusion was confirmed by Veenadhari *et al.* [27] in establishing the relationship between climate parameters and soybean yield. Ashwinirani and Vidyavathi [28] applied this method to improve the methodology for modeling sugarcane yield. According to Fan *et al.* [29], DT has the ability not only to capture non-linear relationships but also to capture extreme cases. Among the different supervised methods applied, ANN and RNN were found to be the least predictive. ANN and RNN are more sensitive to sample size, which is relatively low in our study. KNN also has better accuracy than the other models, except the DT. This result is corroborated by Shakoor *et al.* [30], who revealed the superiority of DT compared to KNN regarding prediction error percentage. Furthermore, the research by Jahan and Shahariar [31] used a DT algorithm to predict fertilizer treatment in maize cultivation and observed good accuracy. Reyes *et al.* [32] showed that KNN was a promising method for predicting winter grain yields.

Figure 4 illustrates a linear relationship in all graphs, suggesting little difference between the actual and predicted maize yield. The predicted variable of KNN and SVM models is not fully correlated with the actual variable shown in Figures 4(a) and 4(b). The DT shows a perfect correlation between actual and

predicted variables shown in Figure 4(c). Moreover, the predicted variable of both ANN and RNN models is not entirely correlated with the actual variable, as shown in Figures 4(d) and 4(e). The correlation coefficient (R^2) for each model indicates the model's threshold of predicting maize yield from weather parameters. However, a higher R^2 means that the strength of the relationship between the yield and climate variables is convincing. All five models used to predict maize yield have R^2 values higher than 0.60. DT recorded the highest R^2 which is greater than 0.95. Consequently, all of the discussions above confirmed that the DT model performs well compared to other models.

Table 2. RMSE, MAE, R^2 , and MSE of KNN, SVM, ANN, RNN, and DT

Models	RMSE	MAE	R^2	MSE
KNN	0.132	0.098	0.755	0.017
SVM	0.145	0.110	0.704	0.021
ANN	0.16	0.125	0.629	0.028
RNN	0.156	0.116	0.664	0.023
DT	0.012	0.0008	0.998	0.021

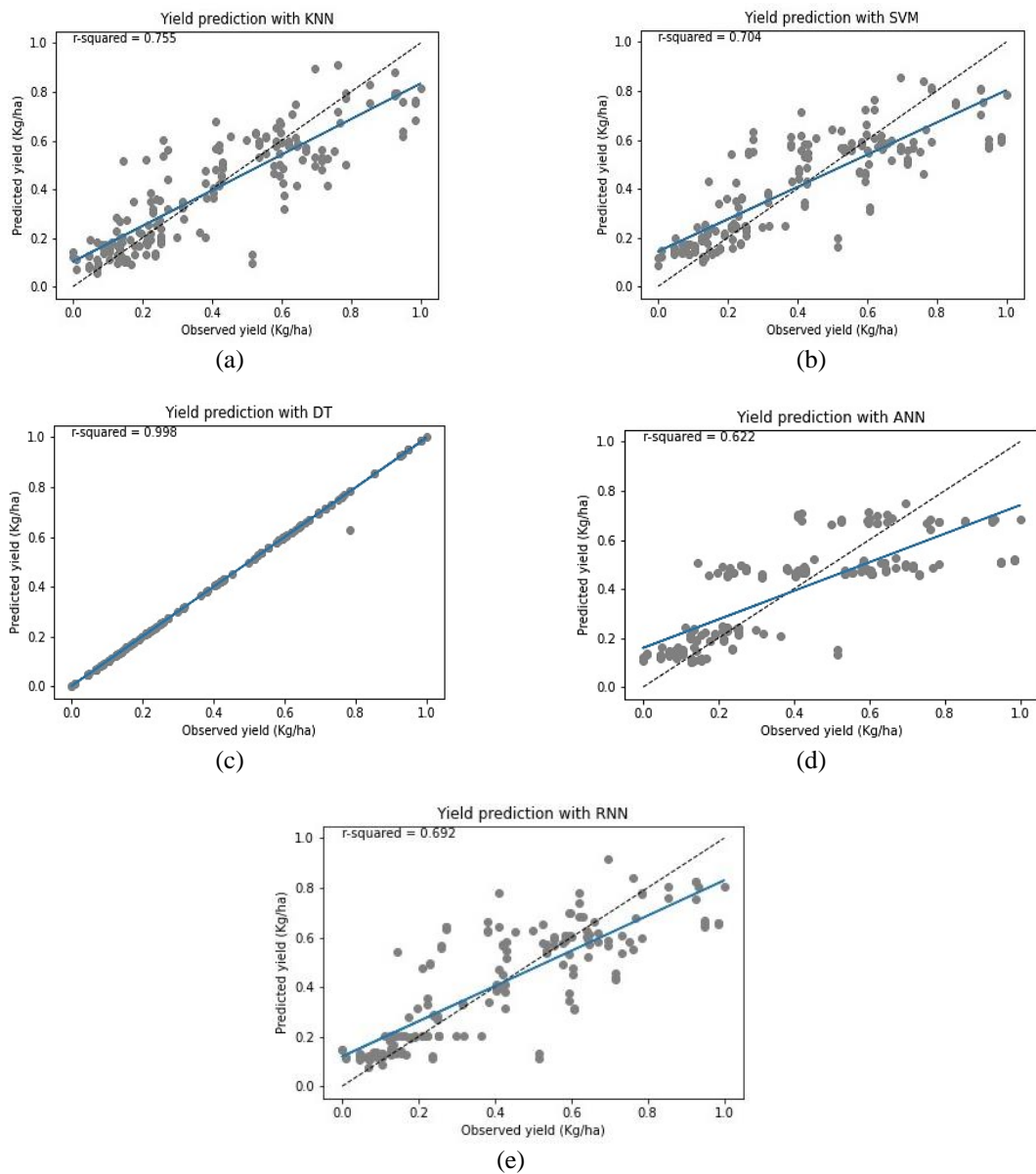


Figure 4. Scatter plots of actual output versus predicted output using different ML techniques: (a) KNN algorithm, (b) SVM, (c) DT, (d) ANN, and (e) RNN. Dashed line: regression line, blue lines: error bars

3.2. Association rules

New data were generated using the DT model to have credible rules to identify early maize yield from weather characteristics, 10,000 for each district. The FPG algorithm is applied to the data. Since FPG uses items, the numerical attribute values are converted into ordinal attributes (low, medium, and high scores) based on fixed threshold values. Minimum support and minimum confidence were, respectively, 0.25 and 0.75. These values have been chosen not too small or too large to allow the FPG algorithm to obtain all the available elements from the database without under or overestimation. The unfiltered list of the rules was obtained after modeling. It was then sorted by lift and confidence. Only the best rules were chosen from this new rule list in each agro-climatic zone. According to the rule selection criteria described above, the strong association rules were selected in each agro-climatic zone. The list of the trained rules had at least three attributes for antecedents of each of the rules formed. The best rules were chosen from each district in both agro-climatic. The association rules listed in Tables 3 and 4 suggest the most suitable conditions for obtaining high maize yield. In Tables 3 and 4, the support values indicate the involvement of a given set of items in the antecedent of association rules. The confidence value expresses support for occurrences of transactions, while lift shows an independent relationship between antecedent and consequent.

Table 3. Association rules for weather and maize yield in Sudano-Guinean zone obtained using simulated data for Bohicon, Kandi, Save, Parakou, and Natitingou districts

Agro climatic zones	Antecedents	Consequents	Support	Confidence	Lift	
Sudano-guinean		Bohicon				
		Bohicon				
		Tmin middle, RR middle, ETP middle	Yield high	0.444	1	1
		Tmin middle, RR middle, Sun middle	Yield high	0.349	1	1
		ETP middle, RR middle, Umax middle, Sun middle	Yield high	0.259	1	1
		RR middle, Umax middle, Sun middle	Yield high	0.32	1	1
		ETP middle, Umax middle, Sun middle, Tmax middle	Yield high	0.258	1	1
		Save				
		Tmin middle, Umax middle, Tmax middle	Yield high	0.411	1	1
		Tmin middle, ETP high, Tmax middle	Yield high	0.357	1	1
		Tmin middle, RR middle, Umax middle	Yield high	0.392	1	1
		Tmin middle, RR middle, Umax middle, Tmax middle	Yield high	0.269	1	1
		RR middle, Umax middle, Tmax middle	Yield high	0.339	1	1

Table 4. Association rules for weather and maize yield in Sudanian zone obtained using simulated data for Bohicon, Kandi, Save, Parakou, and Natitingou districts

Agro-climatic zones	Antecedents	Consequents	Support	Confidence	Lift	
Sudanian zone		Parakou				
		Parakou				
		Tmin middle, Umax middle, Tmax middle	Yield high	0.513	1	1
		ETP middle, Sun middle, Tmax middle	Yield high	0.391	1	1
		Tmin middle, ETP middle, Sun middle	Yield high	0.385	1	1
		Tmin middle, ETP middle, Sun middle, Tmax middle	Yield high	0.319	1	1
		Tmin middle, ETP middle, Umax middle, Sun middle	Yield high	0.311	1	1
		Kandi				
		ETP middle, Umax middle, Tmax middle	Yield high	0.616	1	1
		RR middle, Umax middle, Sun middle, Tmax middle	Yield high	0.324	1	1
		Tmin middle, RR middle, Tmax middle	Yield high	0.417	1	1
		Tmin middle, RR middle, Umax middle, Tmax middle	Yield high	0.374	1	1
		Tmin middle, RR middle, ETP middle, Tmax middle	Yield high	0.361	1	1
		Natitingou				
		Umax low, ETP middle, Tmin middle	Yield high	0.605	1	1
		Umax low, ETP middle, Tmin middle, Sun middle, RR middle	Yield high	0.294	1	1
		RR middle, ETP middle, Sun middle	Yield high	0.475	1	1
		Tmin middle, RR middle, Sun middle	Yield high	0.376	1	1
		Tmin middle, RR middle, ETP middle, Sun middle	Yield high	0.375	1	1

The list of rules defines associations between the medium value of the minimum and the maximum temperature, humidity, sunstroke, rainfall, and evapotranspiration. For example, the best maize trends were observed in the Sudano-Guinean zone for medium values of minimum temperature, rainfall, and evapotranspiration ($ETP=[2.513-3.628 \text{ mm } [; RR=[2.314-6.767 \text{ mm } [; Tmin=[22.088-23.88 \text{ } ^\circ\text{C }]$). Maize yield was also high when the minimum temperature, maximum temperature, and maximum humidity were medium ($Tmin=[22.088-23.88 \text{ } ^\circ\text{C } [; Tmax=[29.836-33.599^\circ\text{C } [; Umax=[93.497-96.302\%]$). This means that maize yield is high in this agro-climatic zone when these climate attributes are medium over the production period. In the Sudanian zone, the yield was high when evapotranspiration, maximum

humidity, and maximum temperature were medium ($ETP=[1.618-4.801mm$ [; $U_{max}=[87.324-102.28$ % [; $T_{max}=[29.371-35.971$ °C [) or when evapotranspiration and minimum temperature were medium and maximum humidity was low ($ETP=[1.618-4.801mm$ [; $U_{max}< 87.324\%$; $T_{min}=[21.337-24.015$ °C[). However, when the values of minimum temperature, maximum temperature, and maximum humidity were medium, the maize yield was also high in the Sudanian zone ($T_{max}=[29.371-35.971$ °C [; $U_{max}=[87.324-102.28\%$ [; $T_{min}=[21.337-24.015$ °C [). Apart from a low humidity obtained in the Soudanian zone, none of the obtained rules had a low attribute. A similar study Kaur and Attwal [13] on the effect of temperature and rainfall on rice yield in Punjab city corroborates the results. They concluded that high rice yield is observed when the temperature is medium during the vegetative phase or when rainfall is low during the grain filling and maturity phase. Supro's [15] work on rice in India supports the results. He showed that rice yield is high under high temperatures and medium humidity. Rao *et al.* [1] estimating paddy yield predictors using temperature, rainfall, soil pH, and nitrogen, observed that high yields were associated with high temperature, rainfall, and medium pH and nitrogen. No rules found by these authors are developed with low attributes. Cirad's [33] study highlights climate change's significant impact on maize yields in sub-Saharan Africa. It emphasizes that a 4 °C temperature increase could decrease maize yields by approximately 14%. Furthermore, the study suggests that adaptation of fertilization is necessary for maize cultivation, especially in the case of substantial temperature increases. Indeed, research by Gunathilake *et al.* [34] revealed that increases in minimum temperature are generally more significant than increases in maximum temperature. The study by Koudahe *et al.* [35] on the impact of climate variability on crop yields in Southern Togo concluded that temperature increase significantly affects the yields of maize and beans. Precipitation also remains a fundamental phenomenon of the climate system, considerably impacting all aspects of life, including agriculture [36]. However, maize was also shown to be less reactive to rising atmospheric CO₂ and less affected by higher temperatures or lower rainfall without nitrogen fertilization [37]. But it becomes more vulnerable to nitrogen deficiency, with more significant negative effects on yield. These conclusions underscore the crucial importance of adapting crop fertilization, particularly for maize, in the face of the challenges of climate change in sub-Saharan Africa. Therefore, it is imperative to implement sustainable water management and climate change adaptation strategies that can help improve crop performance in sub-Saharan Africa. The results of this study highlight the essential conditions for achieving optimal maize yields. To our knowledge, no prediction of maize yield has been made in Benin using association rules. Despite the contribution of this study, some limitations should be considered in future works. This study only considers climatic factors as input variables and thus disregards other factors such as fertilizers, pest management, agricultural practices, and varieties. These factors can be combined with the climatic variables to provide better rules.

4. CONCLUSION

This paper has identified fine climatic parameters for high maize yield using pattern mining in Benin (West Africa). Firstly, an excellent supervised model (decision tree) that predicts maize yield based on weather parameters is proposed. Since this model is not easily interpretable, the FPG algorithm is used to establish the relation between weather and maize yield. The results showed that most of the rules for high maize yield are associated with the medium values of minimum and maximum temperature, humidity, sunstroke, rainfall, and evapotranspiration. The best trends are observed in the Sudano-Guinean and Sudanian zones for medium values of minimum temperature, maximum temperature, humidity or rainfall, and evapotranspiration. These identified rules are a reliable and promising new approach to improving maize yield in Benin. From this study, farmers can implement preventive measures or adapt their cultivation practices to mitigate the risks of climate change.

ACKNOWLEDGEMENTS

This work was supported by German Academic Exchange Service (DAAD) with grant number 91786177 and Scholarship Program in Artificial Intelligence for Development (AI4D) Africa. Funded by the International Development Research Centre (IDRC) and the Swedish International Cooperation Agency (SIDA), and managed by the African Centre for Technology Studies and Development (ACTS) with grant number 10016012.

REFERENCES




- [1] P. R. Rao, S. P. Gowda, and R. J. Prathibha, "Paddy yield predictor using temperature, rainfall, soil pH, and nitrogen," *Lecture Notes in Electrical Engineering*, vol. 545, pp. 245–253, 2019, doi: 10.1007/978-981-13-5802-9_23.
- [2] H. H. Kanwal, I. Ahmad, A. Ahmad, and Y. Li, "Yield forecasting and assessment of interannual wheat yield variability using

- machine learning approach in semi-arid environment,” *Pakistan Journal of Agricultural Sciences*, vol. 58, no. 2, pp. 461–470, 2021, doi: 10.21162/PAKJAS/21.661.
- [3] C. Karunanayake, M. B. Gunathilake, and U. Rathnayake, “Inflow forecast of Iranamadu reservoir, Sri Lanka, under projected climate scenarios using artificial neural networks,” *Applied Computational Intelligence and Soft Computing*, vol. 2020, pp. 1–11, Nov. 2020, doi: 10.1155/2020/8821627.
- [4] M. Shahhosseini, G. Hu, S. Khaki, and S. V. Archontoulis, “Corn yield prediction with ensemble CNN-DNN,” *Frontiers in Plant Science*, vol. 12, 2021, doi: 10.3389/fpls.2021.709008.
- [5] Y. J. N. Kumar, V. Spandana, V. S. Vaishnavi, K. Neha, and V. G. R. R. Devi, “Supervised machine learning approach for crop yield prediction in agriculture sector,” *Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020*, pp. 736–741, 2020, doi: 10.1109/ICCES48766.2020.09137868.
- [6] Y. Guo *et al.*, “Integrated phenology and climate in rice yields prediction using machine learning methods,” *Ecological Indicators*, vol. 120, 2021, doi: 10.1016/j.ecolind.2020.106935.
- [7] S. Iniyani, V. A. Varma, and C. T. Naidu, “Crop yield prediction using machine learning techniques,” *Advances in Engineering Software*, vol. 175, Jan. 2023, doi: 10.1016/j.advengsoft.2022.103326.
- [8] K. Manley, C. Nyelele, and B. N. Egoh, “A review of machine learning and big data applications in addressing ecosystem service research gaps,” *Ecosystem Services*, vol. 57, 2022, doi: 10.1016/j.ecoser.2022.101478.
- [9] Y. Sucharitha, P. C. S. Reddy, and T. N. Chitti, “Deep learning based framework for crop yield prediction,” *AIP Conference Proceedings*, vol. 2548, no. 1, 2023, doi: 10.1063/5.0118526.
- [10] J. Wang, H. Si, Z. Gao, and L. Shi, “Winter wheat yield prediction using an LSTM model from MODIS LAI products,” *Agriculture*, vol. 12, no. 10, 2022, doi: 10.3390/agriculture12101707.
- [11] S. Khaki, L. Wang, and S. V. Archontoulis, “A CNN-RNN framework for crop yield prediction,” *Frontiers in Plant Science*, vol. 10, 2020, doi: 10.3389/fpls.2019.01750.
- [12] M. H. Santos, “Application of association rule method using apriori algorithm to find sales patterns case study of Indomaret Tanjung Anom,” *Brilliance: Research of Artificial Intelligence*, vol. 1, no. 2, pp. 54–66, 2021, doi: 10.47709/brilliance.v1i2.1228.
- [13] K. Kaur and K. S. Attwal, “Effect of temperature and rainfall on paddy yield using data mining,” *Proceedings of the 7th International Conference Confluence 2017 on Cloud Computing, Data Science and Engineering*, pp. 506–511, 2017, doi: 10.1109/CONFLUENCE.2017.7943204.
- [14] N. M. Silas and L. Nderu, “Prediction of tea production in Kenya using clustering and association rule mining techniques,” *American Journal of Computer Science and Information Technology*, vol. 5, no. 2, 2017, doi: 10.21767/2349-3917.100006.
- [15] I. A. Supro, “Rice yield prediction and optimization using association rules and neural network methods to enhance agribusiness,” *Indian Journal of Science and Technology*, vol. 13, no. 13, pp. 1367–1379, 2020, doi: 10.17485/ijst/v13i13.79.
- [16] S. Salankar, A. Salankar, A. Sune, P. Suryavansh, and H. Kumar, “Crop suggestion using data mining approaches,” *2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT, 2021*, doi: 10.1109/ICCCNT51525.2021.9579999.
- [17] L. Nkurunziza *et al.*, “The potential benefits and trade-offs of using sub-surface water retention technology on coarse-textured soils: Impacts of water and nutrient saving on maize production and soil carbon sequestration,” *Frontiers in Sustainable Food Systems*, vol. 3, 2019, doi: 10.3389/fsufs.2019.00071.
- [18] N. R. A. Adjovi *et al.*, “Variation climatique et production vivrière au Sud-Bénin: Cas de la commune de Bohicon,” *Afrique Science*, vol. 15, no. 2, pp. 32–43, 2019.
- [19] E. Fix and J. L. Hodges, “Discriminatory analysis. nonparametric discrimination: Consistency properties,” *International Statistical Review / Revue Internationale de Statistique*, vol. 57, no. 3, Dec. 1989, doi: 10.1007/BF02459570.
- [20] H. Drucker, D. Wu, and V. N. Vapnik, “Support vector machines for spam categorization,” *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1048–1054, 1999, doi: 10.1109/72.788645.
- [21] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, doi: 10.1037/h0042519.
- [22] W. S. McCulloch and W. H. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, Apr. 1943, doi: 10.7551/mitpress/12274.003.0011.
- [23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, doi: 10.1038/323533a0.
- [24] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” *Statistics Surveys*, vol. 4, pp. 40–79, 2010, doi: 10.1214/09-SS054.
- [25] J. Han *et al.*, “Prediction of winter wheat yield based on multi-source data and machine learning in China,” *Remote Sensing*, vol. 12, no. 2, 2020, doi: 10.3390/rs12020236.
- [26] Y. Zeng, S. Yin, J. Liu, and M. Zhang, “Research of improved FP-growth algorithm in association rules mining,” *Scientific Programming*, vol. 2015, 2015, doi: 10.1155/2015/910281.
- [27] S. Veenadhari, D. B. Mishra, and D. C. Singh, “Soybean productivity modelling using decision tree algorithms,” *International Journal of Computer Applications*, vol. 27, no. 7, pp. 11–15, 2011, doi: 10.5120/3314-4549.
- [28] Ashwinirani and B. M. Vidyavathi, “Ameliorated methodology for the design of sugarcane yield prediction using decision tree,” *Compusoft*, vol. 4, no. 7, pp. 1882–1889, 2015.
- [29] J. Fan, A. Jintrawet, and C. Sangchyoswat, “The relationships between extreme precipitation and rice and maize yields using machine learning in Sichuan Province, China,” *Current Applied Science and Technology*, vol. 20, no. 3, pp. 453–469, 2020, doi: 10.14456/cast.2020.30.
- [30] M. T. Shakoore, K. Rahman, S. N. Rayta, and A. Chakrabarty, “Agricultural production output prediction using supervised machine learning techniques,” *2017 1st International Conference on Next Generation Computing Applications, NextComp 2017*, pp. 182–187, 2017, doi: 10.1109/NEXTCOMP.2017.8016196.
- [31] N. Jahan and R. Shahariar, “Predicting fertilizer treatment of maize using decision tree algorithm,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 3, pp. 1427–1434, 2020, doi: 10.11591/ijeecs.v20.i3.pp1427-1434.
- [32] J. R. L. Reyes *et al.*, “Early prediction of winter cereals yield: A preliminary study based on machine learning technique as a strategic tool for fertilization and field management,” *GIoTS 2020 - Global Internet of Things Summit, Proceedings, 2020*, doi: 10.1109/GIOTS49054.2020.9119598.
- [33] Cirad, “La pauvreté des sols et la faible fertilisation en Afrique masquent les effets du changement climatique sur les cultures,” *Cirad*, 2020. [Online]. Available: <https://www.cirad.fr/espace-presse/communiqués-de-presse/2020/adapter-la-fertilisation-des-cultures-en-afrique-au-contexte-du-changement-climatique>
- [34] M. B. Gunathilake, Y. V. Amaratunga, A. Perera, I. M. Chathuranika, A. S. Gunathilake, and U. Rathnayake, “Evaluation of




- future climate and potential impact on streamflow in the upper Nan river basin of Northern Thailand,” *Advances in Meteorology*, vol. 2020, pp. 1–15, Oct. 2020, doi: 10.1155/2020/8881118.
- [35] K. Koudahe, D. Koffi, J. A. Kayode, S. O. Awokola, and A. A. Adebola, “Impact of climate variability on crop yields in Southern Togo,” *Environment Pollution and Climate Change*, vol. 2, no. 1, 2018, doi: 10.4172/2573-458x.1000148.
- [36] M. A. Priatna and E. C. Djamal, “Precipitation prediction using recurrent neural networks and long short-term memory,” *Telkommika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 5, pp. 2525–2532, 2020, doi: 10.12928/TELKOMNIKA.V18I5.14887.
- [37] G. N. Falconnier *et al.*, “Modelling climate change impacts on maize yields under low nitrogen input conditions in sub-Saharan Africa,” *Global Change Biology*, vol. 26, no. 10, pp. 5942–5964, Oct. 2020, doi: 10.1111/gcb.15261.

BIOGRAPHIES OF AUTHORS






Souand Peace Gloria Tahī    holds a Bachelor’s degree in Agronomy (2015) and a Master’s degree in Biostatistics in 2018 at the University of Abomey-Calavi, Benin. She started a Ph.D. thesis on maize yield optimization with machine learning methods in October 2020 under the supervision of Professor Glèlè Kakaï. She can be contacted at email: souandtahi@gmail.com.






Dr. Eng. Vinasetan Ratheil Houndji    is a Senior Lecturer in Artificial Intelligence at the University of Abomey-Calavi (UAC). He obtained his Ph.D. in Computer Science at the Catholic University of Leuven (UCL) in Belgium and the University of Abomey-Calavi (UAC) in Benin in 2017. His research interests are mainly machine learning, constraint programming, combinatorial optimization, and the ethical use of artificial intelligence. He is currently the Head of the Department of Software Engineering at the Institute of Training and Research in Computer Science (IFRI, UAC) and Chair at Ratheil Foundation for Responsible and Efficient Artificial Intelligence (FRIARE). He can be contacted at email: vrateilhoundji@gmail.com.



Dr. Eng. Castro Gbêmémali Hounmenou    is fascinated by data science and research, he has a Ph.D. in Statistics-Probability, specialty Machine Learning, and Data Mining after a master’s degree in statistics option Biostatistics and twenty certificates in data science. He worked in several positions since 2011, moving from consultant in data management and analysis to data science research. Author or co-author of twenty-six articles published in at least peer-reviewed journals. He is a member of several learned societies, including Artificial Intelligence for Development (AI4D), Africa and Applied Malaria Modeling Network (AMMnet) and Data Science Community since 2021. He can be contacted at email: castrohounmenou@gmail.com.



Prof. Dr. Romain Glèlè Kakaï    is a Full Professor of Biostatistics and Forest Estimations, and the Director of the Laboratory of Biomathematics and Forestry Estimations (LABEF) at the University of Abomey-Calavi (UAC, Benin). He is the Coordinator of the Master’s degree program in Biostatistics and the doctoral program in Biometry at the same University. He obtained his degree in Agronomy Engineering in 2000, his Master’s degree, and Ph.D. in Biostatistics at the University of Liege (Belgium) in 2001 and 2005. He became Full Professor (CAMES) in 2015. He is a member of several professional and scholar societies, including Artificial Intelligence for Development (AI4D) Africa, World Economic Forum Scientific Community (2012), TWAS-Young Affiliate (2011-2015), and World Academy of Young Scientists (2012-2018). He was also the Chair of the Africa-Germany Network for Excellence in Science (AGNES, 2019-2023). He can be contacted at email: glele.romain@gmail.com.