Smart Agricultural Technology 2 (2022) 100049

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.elsevier.com/locate/atech

Crops yield prediction based on machine learning models: Case of West African countries

Lontsi Saadio Cedric[a], Wilfried Yves Hamilton Adoni[b,*], Rubby Aworka[a], Jérémie Thouakesseh Zoueu[c,d], Franck Kalala Mutombo[a,e], Moez Krichen[f,g], Charles Lebon Mberi Kimpolo[a]

[a] African Institute for Mathematical Sciences, Senegal Ghana Rwanda
[b] International University of Casablanca, Casablanca, Morocco
[c] University of San Pedro, San Pedro, Côte d'Ivoire
[d] National Polytechnic Institute - Felix Houphouet Boigny, Yamoussoukro, Côte d'Ivoire
[e] University of Lubumbashi, Lubumbashi, Democratic Republic of Congo
[f] Albaha University, Al Baha, Saudi Arabia
[g] REDCAD Research Unit, Sfax, Tunisia

ARTICLE INFO

Keywords:
Agriculture 4.0
Yield prediction
Smart farming
Machine learning
Logistic regression
Decision tree
k-Nearest neighbor
Climate changes

ABSTRACT

Global agricultural production, in particular, is of increasing concern to the major international organizations in charge of nutrition. The rising demand for food globally due to unprecedented population growth has led to food insecurity in some populated regions such as Africa. Another contributing factor to global food insecurity is climate change and its variability. World and African agricultural production in particular are of increasing concern to the major international organizations in charge of nutrition. The World Food Program has reported that high population growth worldwide, especially in Africa in recent years, is leading to increased food security. Moreover, farmers and agricultural decision-makers need advanced tools to help them make quick decisions that will impact the quality of agricultural yields. Climate change has been a major phenomenon in recent decades all over the world. An impact of climate change has been observed on the quality of agricultural production. The arrival of big data technology has led to new powerful analytical tools like machine learning, which have proven themselves in many areas such as medicine, finance, and biology. In this work, we propose a prediction system based on machine learning to predict the yield of six crops, namely: rice, maize, cassava, seed cotton, yams, and bananas, at the country-level in the area of West African countries throughout the year. We combined climatic data, weather data, agricultural yields, and chemical data to help decision-makers and farmers predict the annual crop yields in their country. We used a decision tree, multivariate logistic regression, and k-nearest neighbor models to build our system. We had promising results with both models when using three machine learning models. We applied a hyper-parameter tuning technique throughout cross-validation to get a better model that does not face overfitting. We found that the decision tree model performs well with a coefficient of determination(R^2) of 95.3% while the K-Nearest Neighbor model and logistic regression perform respectively with $R^2 = 93.15\%$ and $R^2 = 89.78\%$. We also study the correlation between the predicted results and the expected results. We found that the prediction results of the decision tree model and the K-Nearest Neighbor model are correlated to the expected data, which proves the efficacy of the model.

1. Introduction

The agricultural sector has a great impact on the economies of African countries and employs around two-thirds of the continent's working population. The Food and Agriculture Organization of the United Nations (FAO) considers that the transformation of the agriculture sector in Africa is at the heart of driving progress towards ending poverty, hunger, and malnutrition. Climate organizations and governments observe a big change in climatic conditions in recent decades that could influence agricultural yields. Climate change has both direct and

* Corresponding author.
E-mail addresses: cedric.l.saadio@aims-senegal.org (L.S. Cedric), wilfried.adoni@uic.ac.ma (W.Y.H. Adoni), rubby@aims.edu.gh (R. Aworka), jeremie.zoueu@inphb.ci (J.T. Zoueu), franckm@aims.ac.za (F.K. Mutombo), mkreishan@bu.edu.sa, moez.krichen@redcad.org (M. Krichen), Charles.Kimpolo@nexteinstein.org (C.L.M. Kimpolo).

https://doi.org/10.1016/j.atech.2022.100049
Received 21 December 2021; Received in revised form 27 March 2022; Accepted 29 March 2022

indirect effects on agricultural productivity, including changing rainfall patterns, droughts, flooding, and the geographical redistribution of pests and diseases (Food and Agricultural Organization of United Nations, 2020). From these facts, crop yield prediction is one of the most challenging problems in precision agriculture. The changing environmental conditions, especially global warming and climate variability, are major concerns and have an adverse impact on the future of agriculture [1]. This problem requires the use of several datasets since crop yield depends on many different factors such as climate, weather, soil, use of fertilizer, and seed variety [2]. Generally, statistical models are employed to predict the crop yield, which is time-consuming and tedious [3]. This decade, the arrival of big data has led to the use of more advanced analysis tools such as Machine Learning. A machine learning model can be descriptive or predictive, depending on the research problem and research questions [4]. While descriptive models are used to gain knowledge from the collected data and explain what has happened, predictive models are used to make predictions in the future [5]. It has been used to solve problems in many areas, such as medicine [6], biology [7], finance [8], and, most recently, agriculture.Machine Learning is an important decision support tool for crop yield prediction, including supporting decisions on what crops to grow and what to do during the growing season of the crops.

Contributions: In this work, we proposed three machine learning models for predicting six crop yields: rice, maize, cassava, seed cotton, yams and bananas in nine West African countries: Burkina Faso, Gambia, Ghana, Guinea, Mali, Mauritania, Niger, Senegal, and Togo. The prediction is made at the country's level and is done throughout the year. We found from the statistical analysis of the overall data that our chosen crops are among the most consumed and cultivated foods in West Africa. The main contributions of this paper are as follows: (1) we proposed a decision support tool to assist farmers and decision-makers in forecasting agricultural yields based on climatic conditions in their zones, in order to better combat climate change and ensure food security in the future; and (2) we proposed an advanced crop yield prediction based on machine learning models.

This work was made possible by collecting weather, agricultural, pesticide, and chemical data from the Food and Agriculture Organization for the United Nations(FAO).[1] The weather data was collected from the climate change knowledge portal (CCKP) of the World Bank Group.[2] We combined these different data sources to predict the yearly crop yield in some West African countries. We proposed three approaches based on three different machine learning models: multivariate logistic regression, k-Nearest Neighbor, and decision tree models. Multivariate logistic regression predicts crop yield with a coefficient of determination (R^2) of $83,80\%$, decision tree with R^2 of $94,65\%$, and k-Nearest Neighbor with R^2 of $95,03\%$.To the best of our knowledge, our work is among the first in that context, when compared to some related works [9] conducted in Africa. In this work, we have been confronted with the availability of climate and agricultural data for Africa. Despite this handicap, we were reassured that the data we were able to collect allowed us to design a prediction model while avoiding overfitting or underfitting cases.

Structure of the paper: The rest of this paper is organized as follows. Section 2 presents some related work in the agriculture and machine learning areas. Section 3 presents the material and methods in which we present the study area, the harvested crops, describe the data source with some analytic results, and apply some feature engineering techniques to the data. The following Section 4 presents our approach, where we present the mathematical model behind our idea, the machine learning models, and the different parameters we used for crop yield prediction. Section 5 presents the experimental results and the discussion. Finally, in Section 6, we conclude this work with some perspectives.

[1] http://www.fao.org/faostat/en/#data/

[2] https://climateknowledgeportal.worldbank.org/

2. Related work

Machine Learning is a sub-domain of artificial intelligence that gives a computer the possibility to learn from data without having to be initially programmed. Machine Learning has found more utility with the arrival of big data technology. Big data can be simply defined as a large volume of data that comes from different sources with a high velocity. Fundamentally, machine learning involves building mathematical models to help understand data [10]. Machine learning [11–13] is used to solve three main types of problems, namely supervised learning problems, unsupervised learning problems, and reinforcement problems. Mathematically, machine learning takes as input a tuple of (X, Y) where X is so-called independent features and Y represents the dependent variable or the target. We are in a supervised learning problem when the variable Y is known before the training. Supervised learning is divided into two common problems, which are regression and classification problems. The housing price prediction problem [14] is known as one of the regression problems. One talks about unsupervised learning when the variable Y is unknown before the training. Their algorithms are widely used in the case of clustering problems.

Climate change refers to long-term shifts in global or regional temperatures or weather patterns. Climate change sets many legal and regulatory challenges on how best to address global warming and reduce greenhouse gas emissions [15]. Global climate change is likely to increase the problems of food insecurity, hunger, and malnutrition for millions of people, particularly in South Asia, Sub-Saharan Africa and small islands [16]. One of the major threats to agricultural development in Africa is climate change [17]. The variability of weather conditions, temperatures, and air quality highly impacts the soil content and, thus, the quality of the agricultural yield. Hence, the current generation needs to find solutions to fight against the negative impacts of environmental consequences on crops.

Crop yield prediction retains a lot of attention for researchers around the world. You et al. [18] presented a deep learning framework for crop yield prediction using remote sensing data. That approach used a Convolutional Neural Network (CNN) with the Gaussian process component and dimensional reduction technique to forecast crop yield of mostly developing countries throughout the year. The technique was applied to a dataset of soybeans obtained by combining a sensing dataset, soil dataset, and climate dataset from the US. Their approach shows that the gaussian approach was used to improve the Root Mean Square Error (RMSE) of the model from 6,27 to 5,83 on average with the Long Short-Term Memory(LSTM) model and from 5,77 to 5,57 with the CNN model.

Another work was conducted by Paudel et al. [19] who combined agronomic principles of crop modeling with machine learning to design a machine learning baseline for large-scale crop yield prediction. Their baseline was a workflow emphasizing correctness, modularity, and reusability. Their features were created by using crop simulation outputs and weather, remote sensing, and soil data from the MARS Crop Yield Forecasting System (MCYFS) database. In their proposed workflow, three machine learning algorithms namely Gradient boosting, Support Vector Regression(SVR), and k-Nearest Neighbors was used to predict the yield of soft wheat, spring barley, sunflower, sugar beet, and potato crops at the regional level in the Netherlands, Germany, and France. Sun et al. [20] proposed a novel multilevel deep learning model coupling Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) to extract both spatial and temporal features to predict crop yield. The main aims of their work were to evaluate the performance of the proposed method for corn belt yield prediction in the US Corn Belt and to evaluate the influence of different data sets on the prediction task. They used both time-series remote sensing data, soil property data, as the inputs. Their experimentation was done in the US Corn Belt states to predict corn yield from 2013 to 2016 at the county level.

Shahhosseini et al. [21] proposed an investigative study to show the impact of coupling crop modeling and machine learning on the improve-

Table 1
Related works synthesis.

References	Authors	Year	Machine learning models	Predicted crops	Study area
[18]	You et al.	2017	• Convolution Neural Network • Gaussian	• Soybean	• United States • Developping countries
[19]	Paudel et al.	2021	• Gradient Boosting • Support Vector Regression • k-Nearest Neighbors	• Soft wheat • Spring barley • Sunflower • Sugar beet • Potatoes	• Germany • France
[20]	Sun et al.	2020	• Recurrent Neural Network • Convolution Neural Network	• Corn	• United States
[21]	Shahhosseini et al.	2021	• Deep Neural Network		
[22]	Khaki and Wang	2019	• Deep Neural Network	• Corn hybrids	
[23]	Abbas et al.	2020	• Linear regression • Elastic net • k-Nearest Neighbor • Support Vector Regression	• Potato tuber	• Atlantic Canada
[3]	Bali et al.	2021	• Recurrent Neural Network • Long Short-Term Memory	• Wheat	• India
[9]	Kaneko et al.	2019	• Deep Neural Network	• Maize	• Ethiopia • Kenya • Malawi • Nigeria • Tanzania • Zambia

ment of corn yield predictions in the US Corn Belt. Their main goals are to see if by using a hybrid approach (crop modeling + machine learning), better predictions can be obtained, and also to study which hybrid model combinations provide the most accurate predictions and to determine which crop modeling features are most effective to be integrated with machine learning for maize yield prediction. They found that adding simulation crop model variables as input features to machine learning models can decrease yield prediction RMSE from 7% to 20%, and that weather information alone is not sufficient. They specified that their proposed machine learning models need more hydrological inputs to make improved yield predictions. Khaki and Wang [22] developed a Deep Neural Network-based solution to predict yield, check yield, and yield difference of corn hybrids based on genotype and environmental (weather and soi) data.Their work was carried out as part of the 2018 Syngenta Crop Challenge. Their model was found to predict with very good accuracy, with a RMSE of 12% of the average yield and 50% of the standard deviation for the validation dataset using predicted weather data.

Another work on the prediction of potato tuber yield has been carried out by Abbas et al. [23]. From data of soil and crop properties collected through proximal sensing, they predict potato (Solanum tuberosum) tuber yield with the help of four machine learning algorithms, namely linear regression, elastic net, k-nearest neighbor and support vector regression. Six fields in Atlantic Canada, including three in Prince Edward Island (PE) and three in New Brunswick (NB), were sampled, over two growing seasons, one in 2017 and another in 2018, for soil electrical conductivity, soil moisture content, soil slope, normalized-difference vegetative index (NDVI), and soil chemistry. Their result shows that Support Vector Regression performs better than all the rest of the models with an RMSE of 5.97, 4.62, 6.60, and 6.17 t/ha for NB-2017, NB-2018, PE-2017, and PE-2018, respectively. Bali and Singla [3] used a 43-year benchmark dataset to predict wheat crop yield in the northern region of India with a deep learning-based Recurrent Neural Network (RNN) model. Their study also employed LSTM to unpack the vanishing gradient problem inherent in the RNN model. The results obtained from the RNN-LSTM model (RMSE = 147.12, MAE = 60.50), Artificial Neural Network (RMSE = 732.14, MAE = 623.13), Random Forest (RMSE = 540.88, MAE = 449.36), and Multivariate Linear Regression (RMSE = 915.64, MAE = 796.07), proved the efficacy of their proposed model (Table 1).

Kaneko et al. [9] recently proposed a crop yield study focusing on African countries.They used a deep learning architecture on satellite image data to predict maize at the district level in six countries in Africa: Ethiopia, Kenya, Malawi, Nigeria, Tanzania, and Zambia. Their model predicted with an R^2 of 0,56. We take another direction by using climate, chemical, and agricultural parameters. The impacts of climate change are most evident in crop productivity because this parameter represents the component of greatest concern to producers, as well as consumers [24].

Unfortunately, in Africa, more precisely in West Africa, little research work deals with the issue of agricultural prediction. Most of the research is based on predictions made using classical statistical models. Unfortunately, these are ineffective and unsuitable for understanding the constraints and hazards associated with the region's agricultural model. Another constraint is the lack of data. Few agricultural data sets are available and those that do exist require advanced ETL (Extract, Transform, and Load) processes to make the data usable. In this context, our work aims at developing machine learning models from scratch capable of incorporating multi-source data into their predictions, which are very close to reality. The proposed machine learning algorithms are used for predicting bananas, dry beans, cassava, rice, maize, and seed cotton at the country-level in certain west African countries throughout the year in this work.

3. Material and methods

3.1. Study area

The research area spans nine countries of West Africa and covers around 6.14 million km^2 and concerns a population estimated at approximately 381 million people. According to the UN's categorization of geographic areas, Benin, Burkina Faso, Gambia, Ghana, Guinea, Guinea-Bissau, Ivory Coast, Liberia, Mali, Mauritania, Niger, Nigeria, Senegal, Sierra Leone, and Togo, as well as Saint Helena, Ascension, and Tristan da Cunha, make up continental West Africa, according to the UN.

Due to geopolitical problems and a lack of agricultural data, the British Overseas Territory, composed of the islands of Saint Helena, Ascension, and Tristan da Cunha, is not included in this research. West Africa is rich in a diverse range of ecosystems as well as a diverse range of environmental conditions, biodiversity, and farming techniques. Agri-

Fig. 1. Bioclimatic Regions of West Africa.[3]

culture is the backbone of West Africa's economy, and the majority of the population relies on it to make a living. Generally, we have two types of crops: (1) export crops that are intended to supply large industries in Europe, America, and Asia. The majority of farms are large, and they cover, on average, more than 20 hectares. This is the case for coffee, cocoa, and hevea in the Ivory Coast and Ghana. (2) import crops that are intended for local needs in terms of food self-sufficiency. The majority of farms are modest, ranging from two to five hectares. However, while the small size of farms indicates land scarcity in densely populated areas such as Nigeria, it also reflects the limited technological tools available in rural areas. Unfortunately, poor production management and certain climatology factors harm the region's agricultural productivity rate.

West Africa is divided into five bioclimatic regions, each of which has its own climate and vegetation. The Guinean, Guineo-Congolian, Saharan, Sahelian, and Sudanian regions are depicted in Fig. 1[3] above. The Sahelian region is a large semiarid area of $350km$ that stretches from Sudan to Senegal through Tchad, Niger, and Mali. The Saharan Region is in the northern part of West Africa. This area is composed of the Sahara desert, which covers the entirety of northern West Africa (from Tchad to Mauritania). In the middle of West Africa, we find the Sudanian region. It is a large area that delimits the southern Sahel. It extends from southern Chad to southern Senegal and across much of Nigeria, Burkina, and Mali. The seasons are generally dry for 5 to 6 months, with annual rainfall ranging from 600 to 1200 mm. A little further south of the Sudanian region, the Guinean region is located between the Guinean-Congolian and Sudanian regions. The rainfall is usually high (1200 to 2220 mm). It is a region containing generally dense and closed forests. The last region is the Guineo-Congolian. It is located in the southwest and southeast of West Africa and borders the Guiean region from the north. It is the wettest area in West Africa, having double the amount of yearly rainfall as the Guinean region. We find deep woods with trees reaching heights of more than 60 m.

3.2. Harvested crops

Rapid population growth and high demand for food are driving agricultural expansion in West Africa. Agricultural production systems are diverse because of the variety of ecosystems. Table 2 depicts West Africa's agricultural expansion. For each country, we presented the top five crop yields. Firstly, we note that root crops are more common in the north, while in the south, tree crops are more common.

In the Sahara region, agricultural production is very low due to a lack of rainfall, poor soil porosity, and fertility. Millet, sorghum, and food crops (groundnuts, maize, and cowpeas) are the most common crops in the Sahelian region, especially in northern Senegal, central Chad, Mali, southern Niger, and north Burkina Faso. Yams and cassava are widely grown in the Guinean region, particularly in Ivory Coast, Ghana, Sierra Leone, and Nigeria. In the south, in the Guinean-Congolese zone, oil palm, cocoa, and cashew are harvested. We also note that the countries of the Sahara and the Sahel have a shortage of staple crops such as tomatoes, rice, etc.

3.3. Data sources

This study was based on agricultural data, annual rainfall data, climate data, weather data, and chemical data from 1990 to 2020. Agricultural data is real data from the Food and Agriculture Organization of the United Nations (FAO)[4] While the rainfall data combines information from 27 stations, including 7 stations in Senegal, 7 stations in Burkina Faso, and 13 stations in Mauritania. The climate data were obtained from the World Bank's knowledge portalCCKP).[5] We used the Apache

[3] https://eros.usgs.gov/westafrica/sites/default/files/inline-images/WA_rate_ag_expansion_v2.2.jpg

[4] http://www.fao.org/faostat/en/data/

[5] (https://climateknowledgeportal.worldbank.org/

Table 2
Percentage (%) of the top 5 crops yields produced in each country of West Africa (FAOSTAT, 2015.).

Crops	Benin	Burkina Faso	Cabo Verde	Ivory Coast	Gambia	Ghana	Guinea	Guinea Bissau	Liberia	Mali	Mauritania	Niger	Nigeria	Senegal	Sierra Leone	Tchad	Togo
Millet	1	19		1	30	3	8	3		30	3	43	6	34	2	24	3
Sorghum	3	27		1	8	4	1	4		22	42	19	11	6	2	26	13
Maize	31	12	46	4	9	15	14	3		11	5		12	5	2	7	32
Cassava	9		1	5	1	13	4	1	11				12	1	22	1	10
Cow peas		18						1		4	10	30	7	5			
Rice	2	2		5	16	3	27	22	42	11	9		6		41	4	4
Yams	7			11		6							9	37		1	4
Groundnuts	5	6		1	29	5	6	6	1	6		5	6		6	12	3
Cocoa				32		24			10				3		3		6
Oil palm fruit	1			4	1	5	9	2	3				7	1	2		1
Seed cotton	9	8		3			1	1		7			1	1		5	5
Cashwe nuts	15	1		12		1		44					1				
Sugar cane			2						4								
Pulses	1		40		3	4	2	1	1		15				7	1	1
Tomatoes	1		2			1							1				
Natural rubber				2					13				1				
Beans, dry	4			1		3					3					3	13
Sesame seed		2			2					1		1				3	
Plantains				6		5	3	3	4				1				
Coconuts	4		3					2									
Fonio							9			1							
Peas											8						

Flume NG engine[6] to ingest (Extraction-Transformation-Loading) the data to a centralized data storage. It easily allows for storing, merging, and sharing large-scale and no-structured data. Because of the lack of data, we have chosen only reliable data in particular for 6 crop yields: rice, maize, cassava, seed cotton, yams, and bananas from Burkina Faso, Gambia, Ghana, Guinea, Mali, Mauritania, Niger, Senegal, and Togo. The characteristic parameters of our prediction models are the most famous and available ones, and they have a strong influence on most types of agriculture in the region. These are:

- **Temperature ($Kelvin$):** the average temperature in the country per year. The increasing of the temperature may cause yield declines between 2.5% and 10% across several agronomic species throughout the 21st century [25].
- **Precipitation (mm):** the average quantity of precipitation in that area per year. Water is a key input to agricultural production and therefore fluctuations in water availability may impact agricultural productivity and revenue [26].
- **Pesticide ($tonnes$):** the quantity of tonne of pesticide use in one hectare of area per country per year. Pesticides are widely used in the agricultural process especially in a low-income country. The popular aim of using pesticides is for improving yield production.
- **Dioxyde of Nitrogen NO_2 (μg):** the quantity of emission nitrogen in that area per year. Nitrous oxide is emitted into the atmosphere as a result of biomass burning, and biological processes in soils [27].
- **Yield (kg/ha):** the quantity of yield per hectare per year, and per country.
- **Year**: the year that the data was collected.
- **Country/Area**: the considered country.

3.4. Features engineering

Table 3 shows a statistical summary of factors affecting crop productivity in West Africa. Temperature and rainfall are the variables dependent on climatic conditions, while pesticide and N_2O are strongly related to farmers' actions. In West Africa, temperatures vary between 7060,84K and 8259,33K, with an average of 27677,25K. The standard deviation shows that there is no disparity in temperatures. This is because the climate remains constantly warm. As for the rainfall, it varies between 92 mm and 1651 mm with a remarkable standard deviation from the average. This is explained by climatic variability. A protracted dry season and a winter (rainy season) that begins in June and finishes in September-October describe the Sahelian climate. This season is linked to the process of marine moisture transfer to the land (the West African monsoon). The average annual rainfall in the Sahelian zone is between 200 mm and 400 mm. The Sahelo-Sudanian region in the south and the Sahelo-Saharan region in the north, respectively receive an average of 400–800 mm, and 50–200 mm of rainfall each year [28]. This rainfall variability greatly impacts crop yields. There is a significant standard deviation (std=92430,1). Indeed, there is a large disparity between the expected average because of the abrupt rainfall changes that directly influence the cultivation techniques adopted by farmers.

We did a multivariate analysis by studying the multi-correlation between the parameters with Pearson formulae. Fig. 2 shows the graph of the correlation between parameters. We found that there was less dependence between these features. We remark that the yields are more correlated with nitrogen dioxide and temperature, than with rain, and pesticide parameters. This shows that rain has no large impact on the yields in those countries, which can be normal since most of them are close to the desert. We remark that there is a link between NO_2 and precipitation. This can be normal since nitrogen dioxide production is controlled by temperature, pH, the water-holding capacity of the soil,

[6] https://flume.apache.org/

Table 3
Statistical overview of the main elements influencing crop yields in West Africa.

	Yield (kg/ha)	Temperature (K)	Pesticide (t)	Rainfall (mm)	NO_2 (10^{18} μg)
mean	5811.1	7677.25	242	825	14.2
std	92430.1	271.24	271	493	9.2
min	196.4	7060.84	1	92	0.89
25%	1179.8	7487.40	39	282	8.07
50%	2214.3	7712.38	148	748	13.84
75%	71064.5	7888.84	384	1187	18.20
max	68908.9	8259.33	2061	1651	39.29

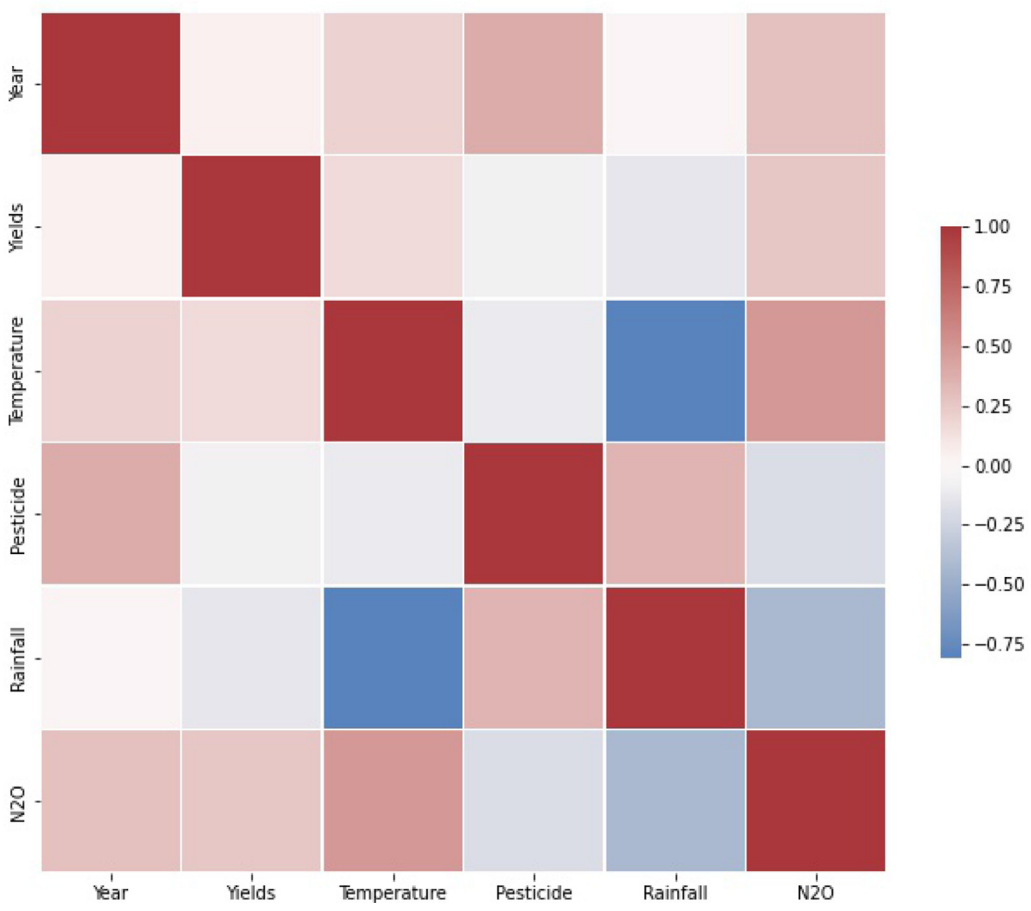


Fig. 2. Pearson Correlation between features.

irrigation practices, fertilizer rate, tillage practice, soil type, oxygen concentration, availability of carbon, vegetation, land-use practices, and use of chemicals [27].

During our data exploration and analysis process, we observed that there were data points that were noticeably different from the rest (also known as outliers) in the yield and pesticide parameters. Those values can represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data. Outliers can badly affect and mislead the training process, resulting in longer training times, less accurate models, and ultimately poorer results. We applied a log transformation technique to bring back all the outliers in the rest of the data. Log transformation is the process of calculating the logarithm values of each data vector. We took into account the country where the data were collected. We consider the country as an important parameter in our study according to the objective. Since the country is of type category, we applied the label encoding algorithm to transform the parameter country to a continuous variable. We drop the year so that our model will not take into account the year during the training. We standardized the dataset using the statistical Z-score formulae. The purpose of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

4. Proposed yield prediction models based on machine learning

We proposed a conceptual system based on machine learning models. Fig. 3 presents the pipeline of the crop yield prediction system. It consists of four steps: the ETL step, feature engineering step, model training step, evaluation step, and the final step, which consists of model deployment.

In the ETL step, we first collect the crop data from different sources. Then, we transform, clean, and process each data source. Finally, we merge and load the final dataset into centralized storage. In the feature engineering step, we applied some data analysis techniques to the final

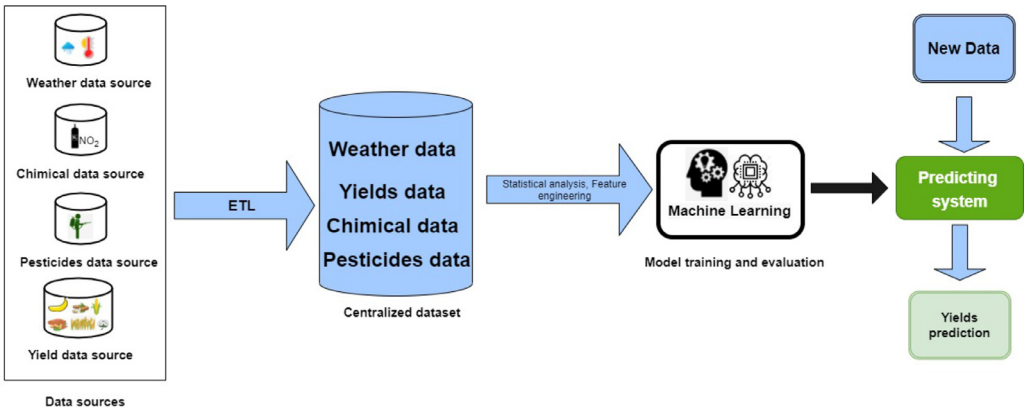


Fig. 3. Pipeline overview of crops yield prediction system.

dataset to find and understand the hidden knowledge in that dataset. The purpose of feature engineering was to prepare the crop data for machine learning models. For the model training step, we used three machine learning algorithms to train the merged dataset to get a predictive model that helps us predict the crop yield. The proposed algorithms are based on multivariate logistic regression, decision trees, and k-Nearest Neighbor models. The generalization principle of machine learning is such a way that the predicting crop model will be able to predict new crop yields with a minimal error. The machine learning model can be viewed as a complex function h in the supervised learning spectrum that takes as input a crop data matrix X and the parameters $Theta$, also known as optimization parameters, and produces an output Y. Eq. (1) gives a general mathematical representation of the process. The crop dataset can be considered as a couple of (X, Y) with $X = [x_1, x_2, \dots, x_n]$ where $x_i \in \mathbb{R}^{1\times m}$, and $Y \in \mathbb{R}^m$. X is a matrix of input data where the columns represent the crop features and the rows represent time-series data collection in each country. Y represents the predicted crop yield. m is the number of instances in the dataset, and n is the number of features taken into consideration. In our case, because of the features encoding, the number of features is $n = 21$.

$$X = \left[x_1, \dots, x_n\right] \Longrightarrow h(X, \Theta) \Longrightarrow Y \tag{1}$$

4.1. Crop multivariate logistic regression

The proposed "Crop Multivariate Logistic Regression" (CMLR) model is a supervised machine learning algorithm that can be used for regression as well as classification tasks. The proposed crop model is distinguished from multivariate linear regression in that the outcome variable (dependent variables) is dichotomous (e.g., diseased or not diseased) .

CMLR uses a logistic function to map the input variables to response/dependent variables. Eq. (2) shows the multivariate logistic regression mathematical model.

$$Y = \frac{1}{1 + e^{-[\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n]}} \tag{2}$$

The logistic function is applied to the linear combination $\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$ to calculate the response of the crop yield targets Y. Where e is the base of the natural logarithms and $\Theta = [\theta_0, \theta_1, \dots \theta_n]$ is the vector of reals unknown parameters and to stay in our context, $x_1, x_2, \dots x_n$ (eg: x_1=temperature, x_2=pesticides, $x_3 = NO_2$, x_4=precipitation), are independent variables. Y is the yield target, or the explanatory variable, or the dependent variables. Fig. 4 presents the conceptual model of our crop multivariate logistic regression used to predict the yields for all West African countries. The proposed model takes as input the climate data, the agriculture data, and the country's information to predict six crop yields (maize, yam, cotton, cassava, rice, and banana). To predict crop yield, we define a threshold. The obtained estimated probability is classified into classes based on this threshold.

After splitting the dataset into the training and testing sets, we trained the crop model across all instances x_i of the training data. We used Gradient Descent to optimize the learning parameters because it converges to a global minimum. At each epoch, we calculate the loss to assess the model's performance. The performance of the crop model is measured using the cross-entropy loss function:

$$LossLog = \sum_{i=1}^{m} = -y_i log(y_i') - (1 - y_i) log(1 - y_i') \tag{3}$$

where $y_i' = h(x_i, \theta_i)$.

4.2. Crop decision tree

The proposed "Crop Decision Tree" (CDT) model is based on the decision tree algorithm [29–31]. It is one of the most widely used supervised machine learning algorithms for regression and classification problems. The proposed crop decision tree allows establishing a system based on multiple covariates or developing advanced prediction algorithms for a yield target variable. During the training process, the crop model builds branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes [32]. We used a decision tree with binary splits, which is much more constructed using the classification and regression trees (CART) concept.

Fig. 5 presents the overall view of our crop decision tree. The proposed tree consists of three categories of nodes: root nodes, intermediate nodes, and leaf nodes. The root node is the start node of the crop graph that best divides the data. In our case, it is the precipitation. While the intermediate nodes represent the evaluated features used in the crop decision tree, they are not the final nodes. Finally, the last type of node is the leaf node. Predictions of agricultural yield are made in these types of nodes.

CDT begins with a root node and finishes with a leaf decision. The model follows a path based on the weight of the branch of that path. The weight is also known as the quantity of information that can be extracted in that branch. The branch with the highest weight is the most suitable to follow. One of our data's features is examined at each node to partition the observations during the crop training phase or to make a single data point follow a certain path while producing a yield prediction. The crop decision tree is constructed by recursively evaluating different crop input features and using the feature that best divides the data at each node.

4.3. Crop k-Nearest neighbor

The third crop yield prediction model is called "crop k-Nearest Neighbor" (Ck-NN), and is based on the k-Nearest Neighbor algorithm, which is a member of the family of supervised learning algorithms. It is used for classification as well as regression problems. Ck-NN works

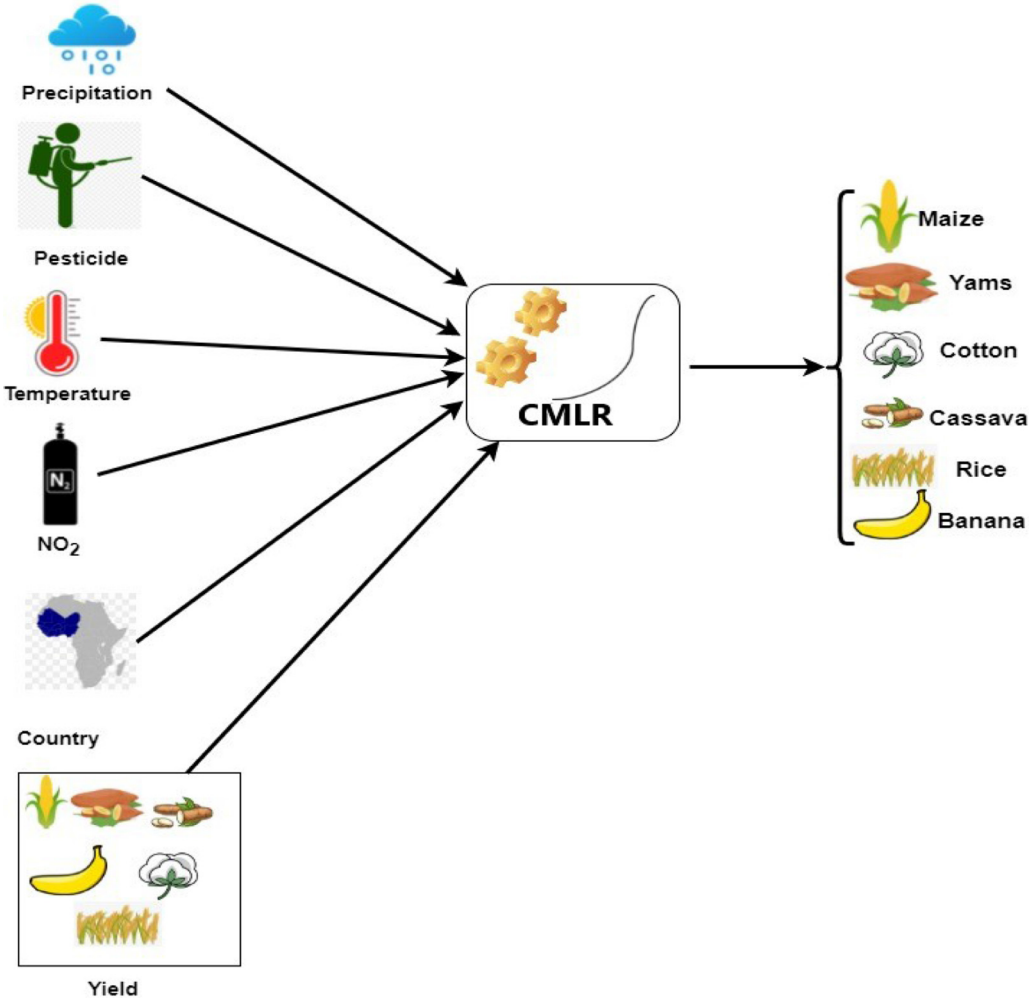


Fig. 4. Crops Multivariate Logistic Regression (CMLR) model.

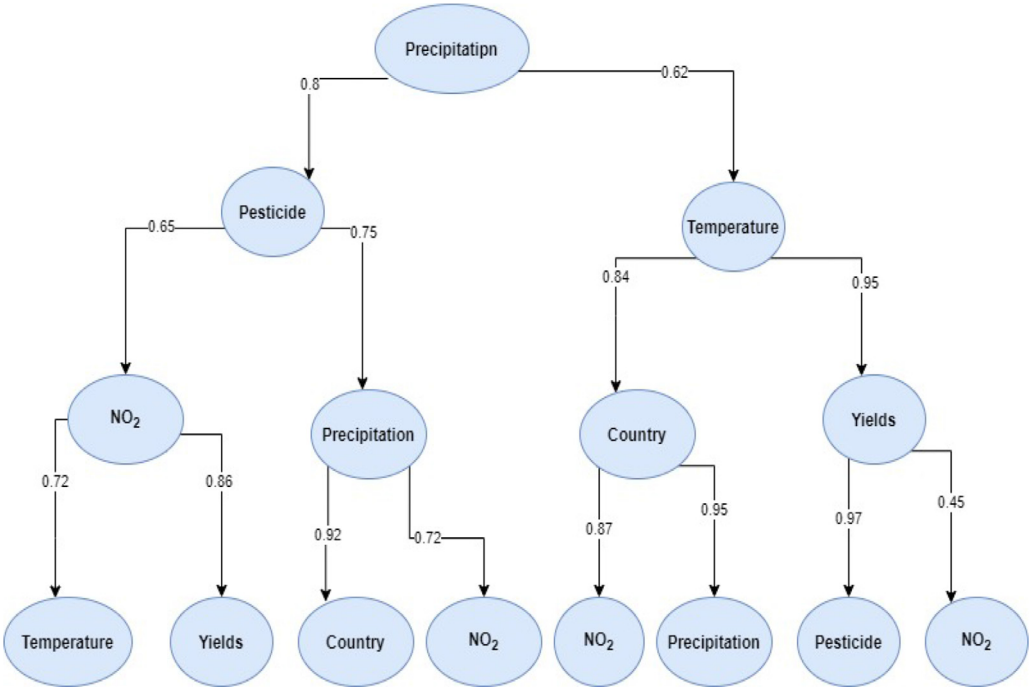


Fig. 5. Global overview of the proposed crop decision tree regression.

on the assumption that every crop data point falling near to each other is falling in the same category. It is a non-parametric algorithm, which means it does not make any assumptions about the underlying data. Ck-NN is recognized as a simple algorithm to implement and is robust to noisy training data. Algorithm 1 shows how Ck-NN predicts crop yield values. As an input, we take the crop training data and the number of neighbors. For each data point, Ck-NN finds its closest neighbor by computing the distance between that data point and the other data point. The algorithm computes the euclidean distance between each data point p_i and the fixed point p_j. The data point p_i will be finally assigned to the group of the k neighbors whose majority have the same similarities.

Input:
For given crop training data set $D = \{(x_i, y_i)\}_{i=1}^{m}$
k = number of neighbors
Searching crop similarity:
for *each* $p_i = (x_i, y_i) \in D$ **do**
  $B \leftarrow []$ *//initialization*
  for $p_j = (x_j, y_j) \in D \setminus \{p_i\}$ **do**
    $B \leftarrow \sqrt{(p_i - p_j)^2}$
  end
  Crop neighbor $p_i \leftarrow sorted(B)[:k]$
  Assign p_i *to its closest crop neighbor group*
end

5. Results and discussion

The experimentation was done on computer hardware using the Linux operating system with a Core i5 processor and 8 Go of RAM as a specification. We work in the Anaconda[7] environment with Python 3.8 as the kernel. We used the Pandas library for data manipulation and analysis. We have exploited the data preprocessing using *Pandas* and *Numpy* libraries to transform the original data into matrix data. We randomly split the data in 90% for training set and 10% for testing set with the function *train_test_split* from the library *scikit-learn*. All the functions that implement our machine learning algorithms are already available in the *scikit-learn* library, so we exploit these functions to do the work. For the hyper-parameter tuning of our models, we used *GridsearchCV* library to implement the cross validation process.

5.1. Model evaluation

We used three metrics to evaluate the systems: R^2, and MAE. The R^2, also known as the coefficient of determination measures the goodness of a prediction for a regression model. More simply, the R^2 score shows how well terms (crop data) fit the hypothesis of the prediction models. Generally, R^2 yields a score between 0 and 1. A value of 1 corresponds to a perfect crop prediction, and a value of 0 corresponds to a constant model that just predicts the mean of the training set responses [33]. There can be some cases where R^2 is negative, which means the model selected does not follow the trend of our agricultural data, therefore leading to a worse fit than the horizontal line. This case usually occurs when there are constraints on either the intercept or the slope of the linear regression line. The second metric is the Mean Absolute Error (MAE) which measures the absolute distance between the actual prediction and the expected prediction. Eq. 4 gives the mathematical expression of MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{m} |y_i - \hat{y}_i| \tag{4}$$

where m is the size of the crop yield datasets, y_i is the expected crop yield response, and $\hat{y}_i = h(x_i, \theta_i)$ is the predicted crop yield. The small the MAE of a model the better the performance of such model. The last metric is the run-time of each model which gives the temporal complexity of each model.

[7] https://www.anaconda.com/products/individual

Table 4
Performance metrics of the crop yield models with default model parameters.

	R^2(%)		MAE(kg/ha)		
Models	Train	Test	Train	Test	Runtime(sec)
CDT	99,97	94,17	0,00034	0,195	0,008
CMLR	84,42	83,80	0.288	0.315	0,121
Ck-NN	97,41	94,77	0,113	0,172	0,005

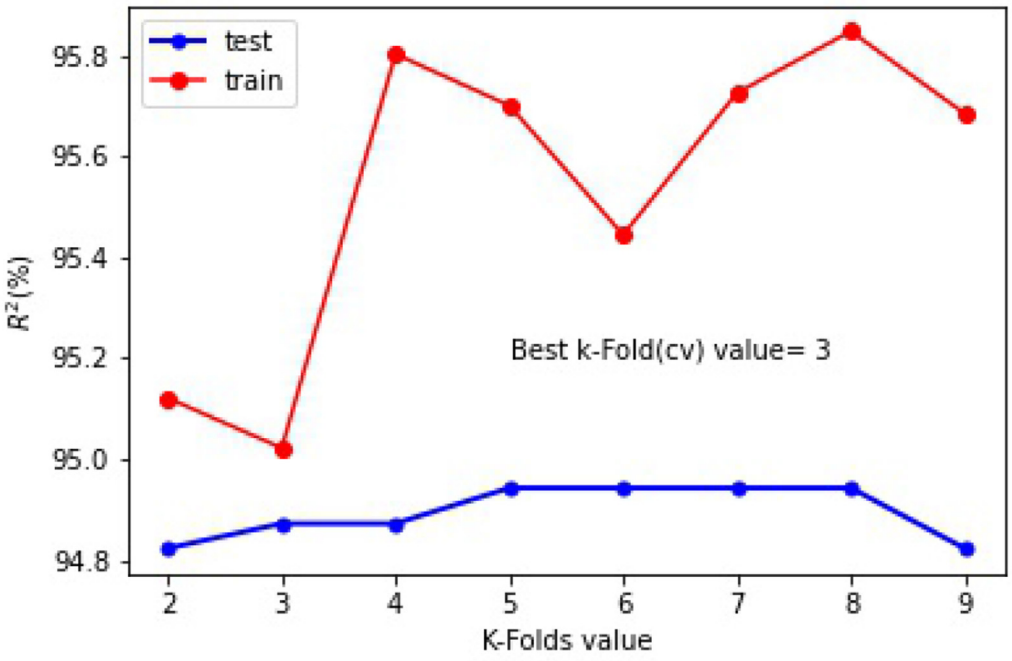
Table 4 shows the evaluation metrics of the used machine learning models with the default model parameters. The Ck-NN model outperformed the crop multivariate logistic regression and the decision tree, while the CMLR model suffered a disadvantage in execution time. The results show that the CMLR model takes two times longer than the decision tree and the Ck-NN. So, for larger data sizes, CMLR would take much longer than other models to return the yield prediction result. We evaluate the model on the train and test datasets. The CDT model with its default parameters predicts with a MAE of $0,00034 kg/ha$ on the train dataset and $0,195 kg/ha$ on the test dataset. On the train dataset, the CMLR predicts with a MAE of $0,288 kg/haand$0,315 kg/ha on the test dataset. The Ck-NN, with its default values (n_neighbors= 5), predit with a MAE of $0,113 kg/ha$ on the train data and $0,172 kg/ha$ on the test data. We can see that in the MAE criteria, there is a little gap between the training and test evaluation metrics. On the train dataset, CDT takes advantage over the rest of the other models, while on the test dataset, Ck-NN has the smallest value of MAE.

By looking at the results of the regression score on the train dataset, we can see that the result of the CDT model is the highest, with R^2 score of $99,97\%$ while the CMLR model result is the lowest. The CDT and Ck-NN models predict with nearly identical R^2 values on the test dataset.These results suggest the presence of an overfitting model as a part of the CDT model. We applied a hyper-parameter tuning technique to find the best model that fit the dataset without an overfitting. We handled this with the help of a cross validation technique.

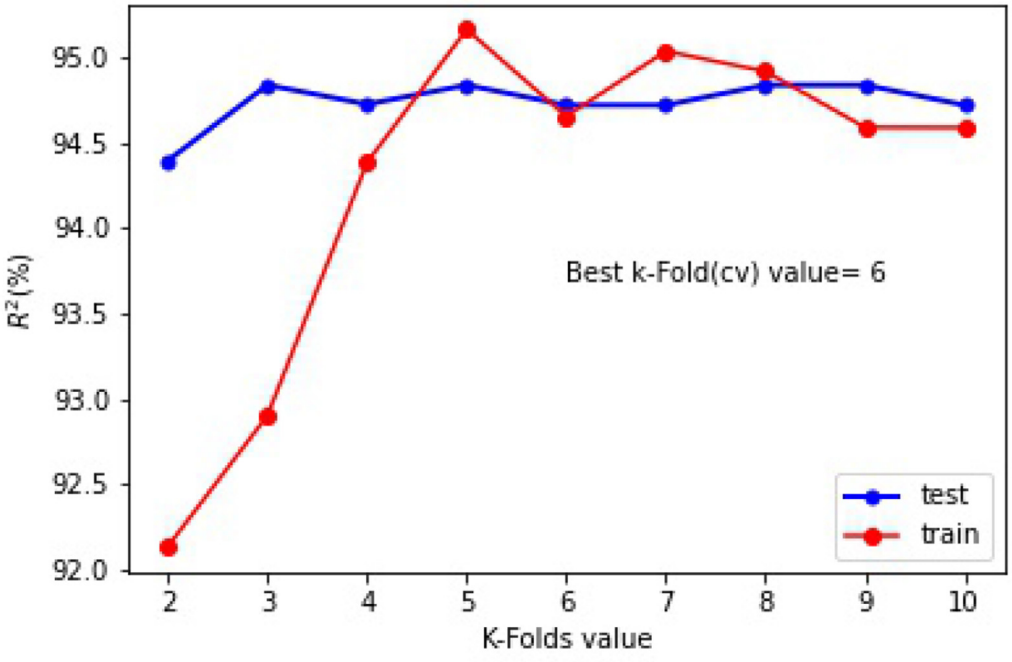
5.2. Hyper-parameters tuning

The results in Table 4 gives an impression of an overfitting model. To be sure that our model did not learn too much from the data, we looked for tuning hyper-parameters through the cross validation method in order to find an optimized model. Cross validation works by splitting the dataset into random sets (k-Folds), holding one set out as the test and training the model on the remaining groups or sets. This process is repeated for each set being held as the test set, and then the average of the models is used for the resulting model. We used the GridsearchCV library, which takes as input the machine learning model (also called an estimator), a grid of hyper-parameters, and the chosen number of groups (K-Folds or cross validation value) and provides the best estimator with its best associated hyper-parameters. We split the data using the ratio of 90/10 in such a way that 90 percent of the data was used as the training data for cross validation, and 10 percent for testing.

Fig. 6 shows the variation of the R^2 score value as a function of the k-Folds value for the Ck-NN and CDT models. The red curve represents the best score of the GridsearchCV method for each iteration of the k-Folds value, and the blue curve represents the score obtained on the test data. The use of this graph allows one to study more easily the situations of overfitting or underfitting. The greater the difference between the training and test scores, the more likely it is to have overfitting. The ideal in the evaluation process of a machine learning model is to have the training score and the test score very close.



(a) Ck-NN



(b) CDT

Fig. 6. Variation of regression score in function of K-Folds value for: (a)Ck-NN model and (b)CDT.

Table 5
Performance metrics of the crop yield models after parameters tuning.

	R^2(%)		MAE (kg/ha)		
Models	Train	Test	Train	Test	Runtime (sec)
CDT	94,72	94,65	0,162	0,088	0,008
CMLR	84,42	83,80	0.288	0.315	0,121
Ck-NN	94,83	95,03	0,133	0,160	0,005

Fig. 6 (a) shows the result of the Ck-NN model. We notice that for values of k-Folds greater than 4, there is a gap between the training and test scores. While the scores of the test data hover around 94.80% and 95,02%, those of the training data oscillate around 95.4% and 95.75%. However, this discrepancy is less felt for the K-Fold = 3, and therefore, there is less risk of overfitting for this value. This is why we have chosen it as a parameter of the GridsearchCV. The tune hyper-parameters obtained at the end of the process are: { leaf_size=20, n_neighbors=3, cv= 3}. Thus, the optimal Ck-NN model has a data point size of 20 data points per leaf, and creates similitude based on 3 crop data point neighbors.

Fig. 6 (b) shows the result of the CDT model. We notice a gap between the training and test scores for K-Folds values between 2 and 4. While for K-Folds values above 4, the training score and test values are a bit closer. We can see that for K-Folds = 6 or K-Folds = 8, we have approximately equal scores. Our choice was made on the value K-Folds = 6 and this led us to the following tuned hyper-parameters: {max_depth=12, min_samples_split=7, random_state=15, cv= 6 } with a R^2 score of 94,51% on the training data and 94,52% on the test data. Thus, the optimal CDT model has a tree that has a maximal depth of 12, a minimal sample split that can be considered at a given node is 7, and a random data shuffle value of 15.

Table 5 shows the summary results after hyper-parameters tuning and model optimization. The optimization of the parameters led us to the values in the table. We obtained a score of 94,72%; a MAE of 0,162 kg/ha on the training data and a score of 94,65%, a MAE of 0.088kg/ha on the test data with the CDT model. We also obtained a score of 94,83%; a MAE of 0,133 kg/ha on the training data and a score of 94,52%; a MAE of 0.088kg/ha on the test data with the Ck-NN model.

The three graphs in Fig. 7 the Pearson correlation between the predicted responses and the expected responses. The representation of the data shows a linear regression with a positive slope. We notice the presence of some values that deviate slightly from the large mass. This could be due to the effect of bias. But the representation of the data shows overall that there is a linearity between the parameters and thus the possibility of a reduced value of the variance.

Fig. 8 shows how the CMLR model react on each crops. On the x-axis, we have the different crops for which we want to predict the yield and on the y-axis we have the values of the test scores obtained through the model.

We can notice on this graph that banana has the highest score with 85.29% followed by cassava which had 25% less than cassava. Cotton and maize were predicted with very low rates which means that there was a bad learning of these two crops. The R^2 score of Seed cotton is negative, this shows that the regression slope of that crop is negative and there is less linearity dependence with the parameters of prediction. This may be one of the reason of the weak performance of the model.

Fig. 9 shows the R^2 of the CDT model on each crop. In contrast to the CMLR scores, all cores are positive. The model performed slightly better on all crops except corn. Cassava had the highest score followed by banana with a R^2 of $94,09\%$, and $93,53\%$ respectively. We can say that unlike yam, the model learned more from cotton, rice, banana and much more from cassava. Maize is the crop with the lowest score, which may mean that the model has not learned enough about the maize data in contrast to the other crops data.

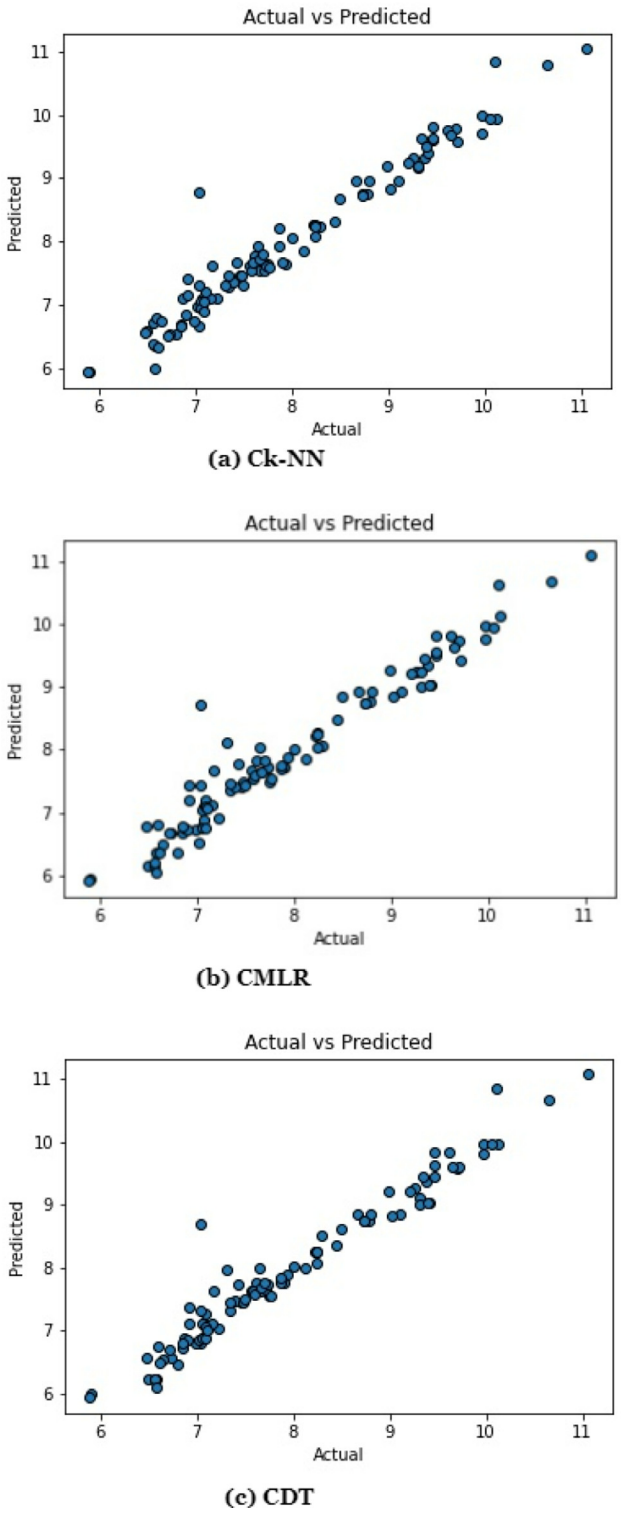


Fig. 7. Comparison between predicted and actual crop yield of each models: **(a)** Ck-NN, **(b)** CMLR and **(c)** CDT.

Fig. 10 also shows the coefficient of determination of each crops as Figs. 8 and 9 but with Ck-NN model. Cassava is the crop with the highest prediction score while maize has the lowest. The two crops have a difference in score of about 40%. Apart from banana and cassava, the other crops have somewhat closer scores, especially rice and cotton which have almost equal scores. These results show that unlike the other models, this one learned better than the other two.

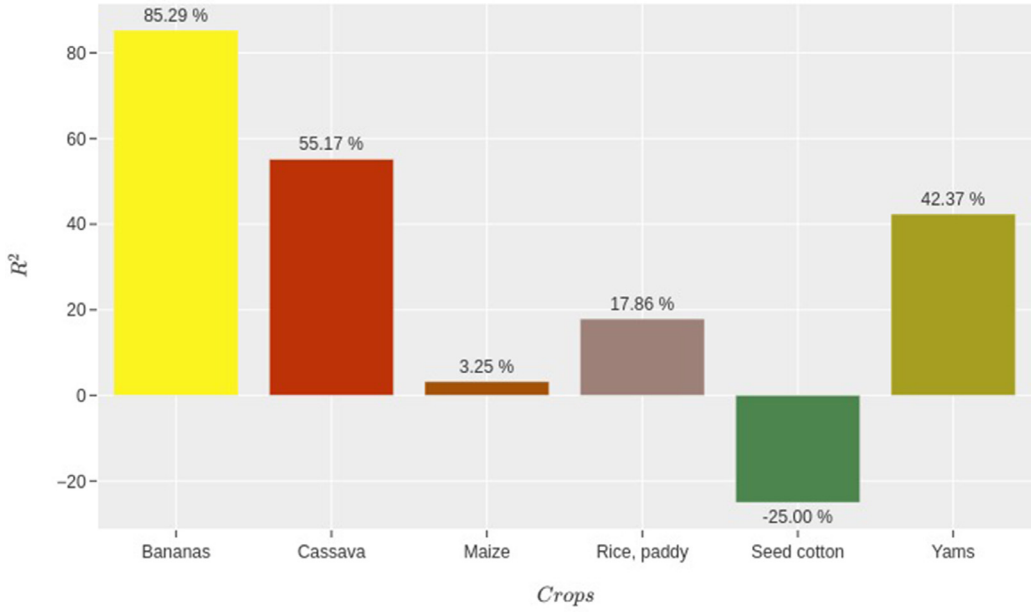


Fig. 8. R^2 scores of each crops with multivariate logistic regression model.

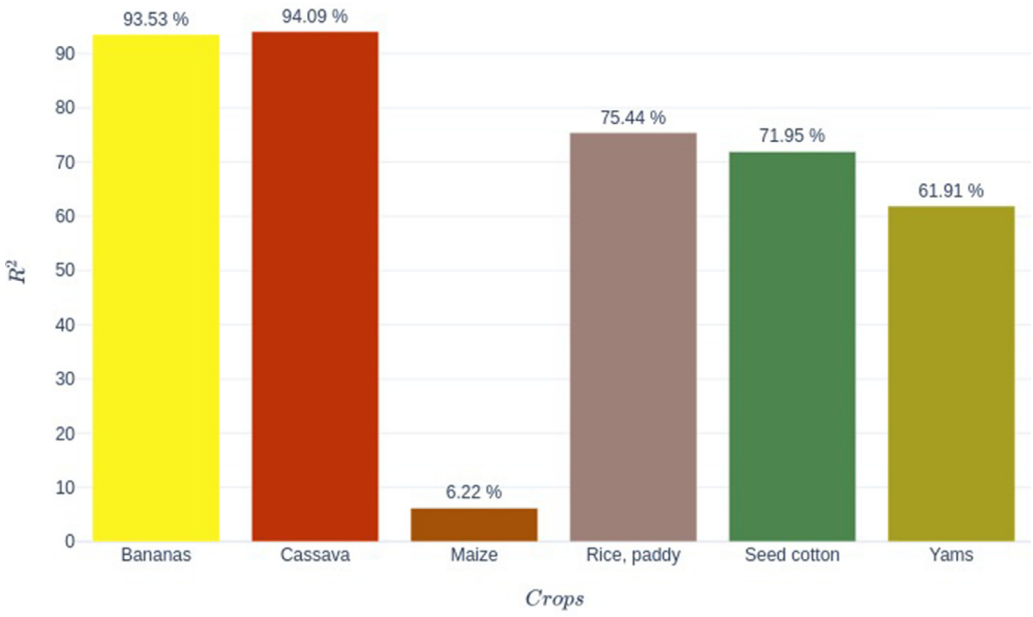


Fig. 9. R^2 scores of each crops with decision tree.

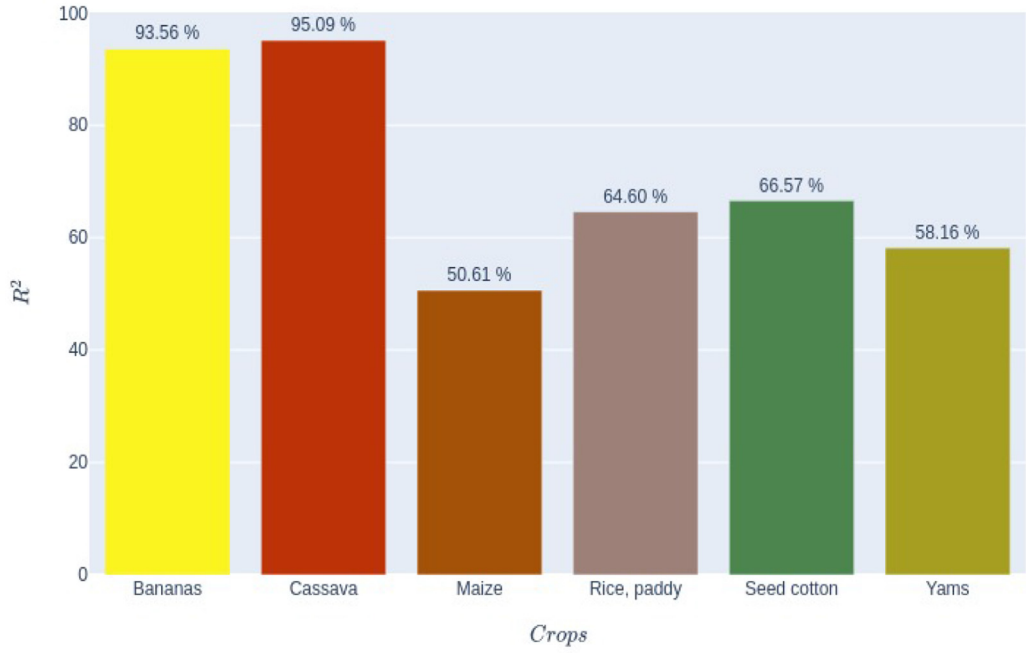


Fig. 10. R^2 scores of each crops with Ck-NN.

6. Conclusion

In this work, we proposed support decision tools for decision-level and farmers that predict at country-level six crop yields, namely bananas, yams, cassava, maize, rice, and seed cotton in some West African countries throughout the year. This work was made possible by combining yields from agricultural, chemicals, pesticides, and weather datasets collected respectively from the Food and Agriculture Organization for the United Nations[8] and the Climate Knowledge Portal World Bank.[9] With the help of ETL techniques, we merged all the different data sources into a centralized database to facilitate the process. We applied preprocessing techniques along with analytics and feature engineering techniques to the combined dataset. The purpose of the analysis was to understand the information hidden behind the data. It consisted of conducting a description of the data and studying the correlation between the variables. The engineering of the variables allowed us to prepare the data for the training of the model. We applied transformation and encoding techniques to variables. We also normalized the data to put them on the same scale. The first results of the training of the CDT and Ck-NN models were not satisfactory enough. We therefore proceeded to optimize the models through parameter tuning with the cross validation technique. The evaluation of the models showed that the Ck-NN model has the best score of all three models. Its R^2 score on test data is 95.03% with an MAE of 0.160 kg/ha while the R^2 of the CDT and CMRL models are 94.65% and 83.80% respectively. We also studied the performance of each model on crops. The results showed that the CMLR model gives a lowest performance, while the Ck-NN gives the highest performance.

The experimental results are conclusive. The proposed prediction models are generalizable in the West African region and support large-scale dataset. As perspective, we are interested to add others features such as soil data, wind data, humidity, agricultural water data, wind data, pollution data, meteorological variations data, animal species data and agricultural economic data of those countries can probably improve the model quality. To the best of our knowledge, we are among the first to work on the case of the African agriculture problem with machine learning after [9]. It will also be interesting to consider techniques related to Big Graphs [34–36] and data collected using Smartphone Sensors [37].

[8] http://www.fao.org/faostat/en/data/

[9] https://climateknowledgeportal.worldbank.org

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

All persons who have made substantial contributions to the work reported in the manuscript (e.g., technical help, writing and editing assistance, general support), but who do not meet the criteria for authorship, are named in the Acknowledgements and have given us their written permission to be named. If we have not included an Acknowledgements, then that indicates that we have not received substantial contributions from non-authors.

References

[1] R. Bhadouria, R. Singh, V.K. Singh, A. Borthakur, A. Ahamad, G. Kumar, P. Singh, Chapter 1 - agriculture in the era of climate change: consequences and effects, in: K.K. Choudhary, A. Kumar, A.K. Singh (Eds.), Climate Change and Agricultural Ecosystems, Woodhead Publishing, 2019, pp. 1–23, doi:10.1016/B978-0-12-816483-9.00001-3.

[2] X. Xu, P. Gao, X. Zhu, W. Guo, J. Ding, C. Li, M. Zhu, X. Wu, Design of an integrated climatic assessment indicator (ICAI) for wheat production: a case study in Jiangsu Province, China, Ecol. Indic. 101 (2019) 943–953, doi:10.1016/j.ecolind.2019.01.059.

[3] N. Bali, A. Singla, Deep learning based wheat crop yield prediction model in Punjab region of North India, Appl. Artif. Intell. 35 (15) (2021) 1304–1328, doi:10.1080/08839514.2021.1976091.

[4] T. van Klompenburg, A. Kassahun, C. Catal, Crop yield prediction using machine learning: a systematic literature review, Comput. Electron. Agric. 177 (2020) 105709, doi:10.1016/j.compag.2020.105709.

[5] E. Alpaydin, Introduction to Machine Learning, 2nd ed., MIT Press, 2010.

[6] P. Doupe, J. Faghmous, S. Basu, Machine learning for health services researchers, Value Health 22 (7) (2019) 808–815, doi:10.1016/j.jval.2019.02.012.

[7] D.M. Camacho, K.M. Collins, R.K. Powers, J.C. Costello, J.J. Collins, Next-generation machine learning for biological networks, Cell 173 (7) (2018) 1581–1592, doi:10.1016/j.cell.2018.05.015.

[8] S. Aziz, M.M. Dowling, H. Hammami, A. Piepenbrink, Machine Learning in Finance: A Topic Modeling Approach, SSRN, 2019, doi:10.2139/ssrn.3327277.
[9] A. Kaneko, T. Kennedy, L. Mei, C. Sintek, M. Burke, S. Ermon, D. Lobell, Deep learning for crop yield prediction in Africa, 2019.
[10] J. VanderPlas, Python Data Science Handbook, Essential Tools for Working with Data, O'reilly, 2016.
[11] Q. Abu Al-Haija, M. Krichen, W. Abu Elhaija, Machine-learning-based Darknet traffic detection system for IoT applications, Electronics 11 (4) (2022) 556.
[12] A. Mihoub, H. Snoun, M. Krichen, R.B.H. Salah, M. Kahia, Predicting COVID-19 spread level using socio-economic indicators and machine learning techniques, in: 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), IEEE, 2020, pp. 128–133.
[13] S. Srinivasan, V. Ravi, V. Sowmya, M. Krichen, D.B. Noureddine, S. Anivilla, K.P. Soman, Deep convolutional neural network based image spam classification, in: 2020 6th Conference on Data Science and Machine Learning Applications (CDMA), IEEE, 2020, pp. 112–117.
[14] Q. Truong, M. Nguyen, H. Dang, B. Mei, Housing price prediction via improved machine learning techniques, Procedia Computer Science, 2019 International Conference on Identification, Information and Knowledge in the Internet of Things, 174, 2020, pp. 433–442, doi:10.1016/j.procs.2020.06.111.
[15] J.F. McEldowney, Chapter 22 - climate change and the law, in: T.M. Letcher (Ed.), The Impacts of Climate Change, Elsevier, 2021, pp. 503–519, doi:10.1016/B978-0-12-822373-4.00018-5.
[16] A. Costa de Oliveira, N. Marini, D.R. Farias, Climate change: new breeding pressures and goals, in: N.K. Van Alfen (Ed.), Encyclopedia of Agriculture and Food Systems, Academic Press, Oxford, 2014, pp. 284–293, doi:10.1016/B978-0-444-52512-3.00005-X.
[17] T.O. Williams, M.L. Mul, O.O. Cofie, J. Kinyangi, R.B. Zougmoré, G. Wamukoya, M. Nyasimi, P. Mapfumo, C.I. Speranza, D. Amwata, et al., Climate smart agriculture in the African context(2015).
[18] J. You, X. Li, M. Low, D. Lobell, S. Ermon, Deep gaussian process for crop yield prediction based on remote sensing data, in: The Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[19] D. Paudel, H. Boogaard, A. de Wit, S. Janssen, S. Osinga, C. Pylianidis, I.N. Athanasiadis, Machine learning for large-scale crop yield forecasting, Agric. Syst. 187 (2021) 103016, doi:10.1016/j.agsy.2020.103016.
[20] J. Sun, Z. Lai, L. Di, Z. Sun, J. Tao, Y. Shen, Multilevel deep learning network for county-level corn yield estimation in the u.s. corn belt, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13 (2020) 5048–5060, doi:10.1109/JSTARS.2020.3019046.
[21] M. Shahhosseini, G. Hu, I. Huber, S. Archontoulis, Coupling machine learning and crop modeling improves crop yield prediction in the US corn belt, Sci. Rep. 11 (2021), doi:10.1038/s41598-020-80820-1.
[22] S. Khaki, L. Wang, Crop yield prediction using deep neural networks, Front. Plant Sci. 10 (2019), doi:10.3389/fpls.2019.00621.
[23] F. Abbas, H. Afzaal, A.A. Farooque, S. Tang, Crop yield prediction through proximal sensing and machine learning algorithms, Agronomy 10 (7) (2020), doi:10.3390/agronomy10071046.
[24] J.L. Hatfield, J.H. Prueger, Temperature extremes: Effect on plant growth and development, weather and climate extremes, USDA Research and Programs on Extreme Events 10 (2015) 4–10, doi:10.1016/j.wace.2015.08.001.
[25] J.L. Hatfield, K.J. Boote, B.A. Kimball, L.H. Ziska, R.C. Izaurralde, D.R. Ort, A.M. Thomson, D. Wolfe, Climate impacts on agriculture: implications for crop production (2011).
[26] M. Torres, R. Howitt, L. Rodrigues, Analyzing rainfall effects on agricultural income: why timing matters, EconomiA 20 (1) (2019) 1–14, doi:10.1016/j.econ.2019.03.006.
[27] J.R. Freney, Emission of nitrous oxide from soils used for agriculture, Nutr. Cycling Agroecosyst. 29 (1) (1997), doi:10.1023/A:1009702832489.
[28] Z. Nouaceur, La reprise des pluies et la recrudescence des inondations en Afrique de l'Ouest sahélienne, Physio-Géo 15 (2020) 89–109.
[29] N. Öcal, M. Ercan, E. Kadioglu, Predicting financial failure using decision tree algorithms: an empirical test on the manufacturing industry at Borsa Istanbul, Int. J. Econ. Finance 7 (2015) 189–206, doi:10.5539/ijef.v7n7p189.
[30] S. Divyashree, H.R. Divakar, Prediction of human health using decision tree technique, Int. J. Comput. Sci.Eng. 6 (2018) 805–808, doi:10.26438/ijcse/v6i6.805808.
[31] Z. Quan, E. Valdez, Predictive analytics of insurance claims using multivariate decision trees, Depend. Model. 6 (2018) 377–407, doi:10.1515/demo-2018-0022.
[32] Y.-y. SONG, L.U. Ying, Decision tree methods: applications for classification and prediction, Shanghai Arch. Psychiatry (2015), doi:10.11919/j.issn.1002-0829.215044.
[33] A.C. Müller, S. Guido, Introduction to Machine Learning with Python, O'Reilly, 2016.
[34] W.Y.H. Adoni, N. Tarik, M. Krichen, A. El Byed, HGraph: parallel and distributed tool for large-scale graph processing, in: 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), IEEE, 2021, pp. 115–120.
[35] H.W.Y. Adoni, T. Nahhal, M. Krichen, B. Aghezzaf, A. Elbyed, A survey of current challenges in partitioning and processing of graph-structured data in parallel and distributed systems, Distrib. Parallel Databases 38 (2) (2020) 495–530.
[36] W.Y.H. Adoni, T. Nahhal, M. Krichen, I. Assayad, et al., DHPV: a distributed algorithm for large-scale graph partitioning, J. Big Data 7 (1) (2020) 1–25.
[37] M. Krichen, Anomalies detection through smartphone sensors: a review, IEEE Sens. J. (2021).