Predicting the Rate of Aflatoxin Contamination in the White Corn Value Chain

Mahugnon Géraud Azehoun Pazou(✉), Julian Adjibi, Régis Donald Hontinfinde, Elognissè Erasme Guérin Agossadou, Vidédji Naéssé Adjahossou, Christian Djidjoho Akowanou, and Macaire B. Agbomahena

National University of Science, Technology, Engineering and Mathematics (UNSTIM), Abomey, Benin
geraud.pazou@unstim.bj
https://unstim.bj

Abstract. With the digital revolution, computer data is a resource of inestimable value. Businesses use them to understand consumer behavior, make informed decisions, and anticipate market trends. Governments rely on data to develop effective policies, monitor public health and manage resources. In this work, digital data collected in the agriculture sector allowed us to compare different machine learning models to predict aflatoxin infection levels in white corn crops. To do so, we use some qualitative and quantitative variables collected on the field, during a previous work. The compared methods are linear regression, random forests, artificial neural networks and support vector regression. The results of the analysis indicate that the random forest regression model stood out as the most effective in predicting aflatoxin infection levels. It posted an RMSE of 0.14 on the training set and 0.29 on the test set, accompanied by a coefficient of determination of 0.81, demonstrating its robustness on both data sets. This performance can be attributed to the ability of random forests to capture the complex and non-linear relationships between maize traits and aflatoxin levels. Evaluating the models on a separate test set confirmed their generalizability, that is, their ability to maintain accuracy with new data. This result constitutes a promising tool for actors in the agricultural sector, providing valuable information for risk management and strategic decision-making aimed at reducing consumer exposure to aflatoxin, thus contributing to the improvement of food safety and public health.

Keywords: machine learning · prediction models · random forest · aflatoxin contamination · white maize

Supported by UNSTIM.

K. Cheikh M. F. Kebe et al. (Eds.): InterSol 2024, LNICST 610, pp. 214–227, 2025.
https://doi.org/10.1007/978-3-031-86493-3_17

1 Introduction

Given the escalating digital revolution, the importance of leveraging data has become paramount. Just as businesses use data for consumer insight and decision-making, governments use it to develop policies and monitor vital aspects such as public health and resource allocation. Maize (Zea mays L.) is the most produced cereal in the world, with a harvest of nearly 1,100 million tonnes in 2018/2019, followed by wheat (734 million tonnes) and rice (495 million tonnes) [16]. Africa consumes 30. 0% of the global production of maize, and sub-Saharan Africa accounts for 21. 0% of the consumption. About 14 countries in Africa consume between 85.0 and 95.0% maize as their staple food and are more inclined to consume white maize, with a consumption share of around 90.0%. [1].

Maize faces various diseases, such as corn rust (Puccinia sorghi), Corn downy mildew (Peronosclerospora sorghi), Corn bunt (Ustilago maydis), Corn cyst nematodes (Heterodera zeae), corn spots of maize disease (Cercospora zeae-maydis), as well as mycotoxin infections [19]. Some mycotoxins present in foods pose a serious threat to health. Among the most toxic are aflatoxins, produced by molds that grow in soil, rotting vegetation, and cereals, especially maize. These substances can have serious acute effects after ingestion, up to acute intoxication, and can be life-threatening by damaging the liver. Aflatoxins are also genotoxic, meaning that they can damage DNA and promote the development of cancer. They can also be found in the milk of animals fed with contaminated food [17]. Faced with the major health risks posed by mycotoxins, particularly aflatoxins, monitoring and controlling their presence in foods is of paramount importance.

The contamination of maize by aflatoxins is a global issue that has sparked numerous scientific investigations. Multiple studies have been conducted to better understand the factors that contribute to this contamination and to find potential solutions. For example, Wu et al. [18] demonstrated that natural contamination by aflatoxins has serious implications for both international trade and public health. They noted that more than 100 countries have implemented regulations on aflatoxins, although this also results in economic losses for exporting countries. The authors also note that even in regulated countries, many people consume uninspected maize, contributing to the adverse effects of aflatoxins on global trade and health.

Lauren et al. [9] studied a large outbreak of aflatoxicosis in rural Kenya in 2004, resulting in 317 cases and 125 deaths. They conducted a cross-sectional survey to assess maize contamination in markets and its link to this epidemic. Their results showed that 55% of the maize samples exceeded the Kenyan regulatory limit of 20 ppb of aflatoxin, with 35% at more than 100 ppb and 7% at over 1,000 ppb. They also observed that maize coming from affected areas was more likely to be contaminated, thus entering the distribution system and contributing to the widespread contamination of corn in markets.

Kamika et al. [8] studied the ubiquitous presence of aflatoxins in maize in the Democratic Republic of Congo, a problem that is both economic and public health. Their results showed that 32% of the maize samples collected were contaminated, with levels up to 103.89 μg/kg for total aflatoxins. This contamination

was worsening throughout the supply chain, with 100% of the market samples showing levels up to 500 times higher than the maximum limit of 10 μg/kg established by the WHO. This significant increase ($p < 0.01$) between harvest and market distribution highlights the urgency of implementing strategies to control the proliferation of aflatoxins in corn.

Muga et al. [12] studied the impact of temperature, relative humidity, and moisture content on aflatoxin contamination of maize kernels. Their results showed that temperature and relative humidity had a significant effect, while moisture content did not. The contamination was higher at 30°C than at 20°C, and a relative humidity of 90% resulted in much higher aflatoxin levels than at 60%. They therefore concluded that maintaining a relative humidity below 60% makes it possible to significantly limit the contamination of corn grains by aflatoxins, thus ensuring their safety for consumption.

Kachapulula et al. [6] quantified aflatoxins in maize and peanuts in the three agroecologies of Zambia. Their results show that 17% of the market harvests exceeded the allowed limits, with higher contamination in the hottest agroecology (38%) than in the coldest and most humid (8%). They also observed that improper storage could increase contamination by more than 1000 times and that the structure of the fungal community influenced this increase. Their work highlights the need for aflatoxin management in Zambia, particularly through the use of atoxigenic biocontrol agents.

Hannah et al. [7] sought to determine the influence of post-harvest practices and storage conditions on the contamination of corn with aflatoxin in two specific countries. Their results showed that Makueni County had the highest positive sample rate, attributed to prolonged storage under poor conditions. In contrast, Baringo County showed less contamination, related to the harvest period. The authors observed that the storage type had a significant impact, explaining 11% of the variation, with burlap bags being the most contaminated. They concluded that proper drying of maize and its storage in airtight structures would reduce cases of aflatoxin contamination.

Various studies have also adopted the machine learning approach to address the challenges faced in the field of agriculture. Among these, we find that carried out by GENSERBE et al. [2] which focuses on the use of machine learning for mapping agropastoral resources in the Fitri region, northern Chad, using MSI Sentinel-2A images.

Hanadé et al. [3] proposed a modeling approach based on the assessment of drought vulnerability of agrosystems in the central Sahel, using machine learning techniques to analyze the extent of changes.

Recently, Makowski et al. [10] explored the application of supervised learning to simulate the impacts of climate change on agricultural yields, thus offering crucial perspectives for anticipating and mitigating these effects.

In this study, we attempt to automatically predict the occurrence of aflatoxin infection and the associated contamination level. To do this, we opted for an approach based on machine learning. We used data collected by Mugure et al. [7], who examined the impact of various factors such as temperature, relative

humidity and moisture content on the contamination of maize. After carrying out a statistical analysis of these data to identify the main factors influencing aflatoxin contamination, we selected, implemented and compared four prediction models.

The paper proceeds as follows: Sect.2 outlines the materials and methods employed. Section3 presents and discusses the obtained results, leading to the conclusion in Sect. 4.

2 Material and Methods

2.1 Dataset

The data used in our study comes from the work of Mugure et al. [7] carried out on storage conditions and post-harvest practices affecting aflatoxin infection in the counties of Makueni and Baringo in Kenya. These data were obtained through surveys carried out by authors on farms in these regions, covering various aspects such as agricultural practices, storage methods, environmental conditions and other relevant factors.

The field surveys were designed to collect detailed information on farmers' agricultural practices, including cultivation methods, types of seeds used, pesticides applied, and harvesting practices. Additionally, data was collected on maize storage methods after harvest, such as the type of silos or bags used, temperature and humidity conditions, and duration of storage. These information were essential to assess potential risk factors for aflatoxin infection throughout the supply chain.

Alongside the field surveys, maize samples were taken from different farms for laboratory analyses. These analyzes included aflatoxin contamination tests, which were carried out to measure contamination levels in the corn samples collected. The data obtained from these analyzes were crucial to assess the presence and level of aflatoxin in the corn samples, thus forming the target variable of our study.

Among the variables in our dataset, we included information on the type of storage (in metal bins, open storage, plastic bags) which indicates the different types of containers used to store maize. In addition, we also took into account variables such as drying of the maize, and its conservation (poorly dried, wet, or poorly preserved).

Other important variables include information on storage practices, such as "Store bags on the ground," Dry grains on a tarpaulin, and "Dry grains on bare earth roofs," which describe methods of storage. In addition, we also included variables such as $Qualitygrains_Yes$ and "Store bags on $woodenpallets_Yes$ which provide information on the quality of the grains and the storage methods used.

Using these variables, we were able to characterize in detail the storage conditions and post-harvest practices associated with the maize studied. This characterization allowed us to explore the factors that influence aflatoxin levels in maize and develop accurate prediction models to identify potential risks of aflatoxin infection (Table 1).

Table 1. Variables and Types

Qualitative Variables	Quantitative Variables
Aflatoxin Level (ppb)	Type of Storage
Day - Positivity	Poorly Dried or Wet Corn
	Improper Storage of Corn
	Drying of Corn on the Ground
	Shelling of Wet Corn
	Processing Method
	Drying
	Shelling
	Canvas Bags
	Photo Bags
	Thatch from the Attic
	Attic Sheet Steel
	Waterproof Bins
	Airtight Storage
	Method of Conservation
	Store the Bags on the Ground
	Dry the Grains on a Tarpaulin
	Dry Grains on a Bare Earthen Roof
	Grain Quality
	Store Bags on Wooden Pallets

Table 2 presents descriptive statistics for two variables: aflatoxin level (in parts per billion, ppb) and number of days.

For the aflatoxin level, we can observe an average of 35.19 ppb, with a fairly high standard deviation of 50.55 ppb, indicating significant variability in the measurements. The minimum value is 0.23 ppb and the maximum value reaches 174.99 ppb. The 25%, 50% and 75% quartiles are 4.85 ppb, 6.19 ppb and 52.03 ppb respectively.

Concerning the number of days, the average is 107.7 days, with a standard deviation of 75.22 days. The minimum value is 30 days and the maximum value is 180 days, probably corresponding to the total duration of the study.

Table 2. Statistics for Aflatoxin Level (ppb) and Days variables

	Aflatoxin Level (ppb)	Days
count	139.000000	139.000000
mean	35.185633	107.697842
std	50.545483	75.222535
min	0.230000	30.000000
25%	4.850000	30.000000
50%	6.190000	180.000000
75%	52.030000	180.000000
max	174.990000	180.000000

2.2 Data Loading and Preprocessing

During the exploratory phase of the data, we carried out an in-depth visual analysis to better understand the characteristics of the target variable, namely the aflatoxin level (Fig. 1).

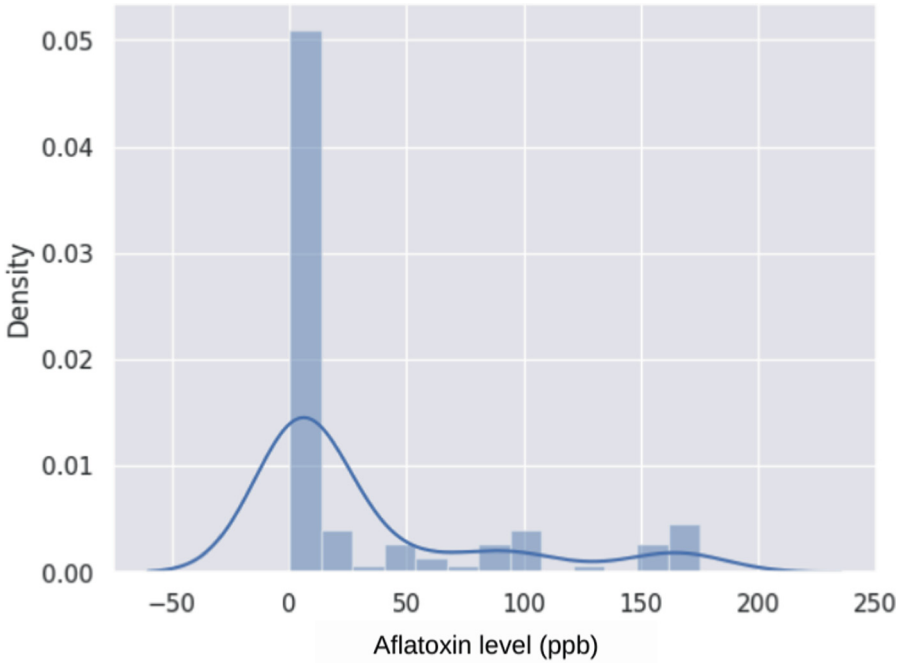

Fig. 1. Aflatoxin Level Histogram

Examination of the histogram of this variable revealed that it does not follow a normal distribution, which led us to reject traditional statistical methods for analyzing the effect of explanatory variables.

We then used box and whisker plots to visualize the impact of each of our explanatory variables on the target variable. This step allowed us to identify the most relevant variables to be included in the dataset used to train our predictive models. In addition, establishing a correlation matrix helped us to better understand the relationships between the different variables.

Once this initial exploration was completed, we proceeded to encode and standardize our variables using the One Hot Encoder library. Finally, we split our dataset into a training set (70%) and a testing set (30%), so that we can train our selected models on the training data. Data normalization was also performed to put all features on the same scale, which allowed us to avoid certain features dominating others in the model training process.

2.3 Data Exploration and Predictive Modeling

After data preprocessing, we conducted data analysis and modeling to develop prediction models for aflatoxin infection in maize. This step includes several sub-processes, including feature selection, model choice, model training and validation.

Data analysis consisted of selecting the most relevant features for predicting aflatoxin infection. This selection was based on statistical methods such as correlation analysis and machine learning techniques such as feature selection by backward elimination.

Regarding prediction models, we explored four models, chosen according to the nature of the data (qualitative explanatory variables and quantitative target variable) and the complexity of the prediction problem:

– Linear regression, a simple and easy to interpret statistical model, establishing a linear relationship between the target variable (aflatoxin level) and the predictive variables (Fig. 2).

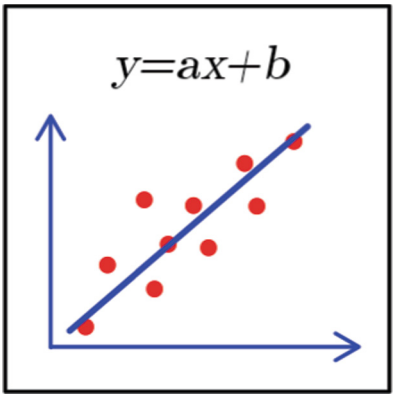

Fig. 2. Graphical representation of a linear regression [11]

– Random forests, a set of decision tree models providing a robust and efficient approach to managing complex datasets and capturing non-linear relationships (Fig. 3).

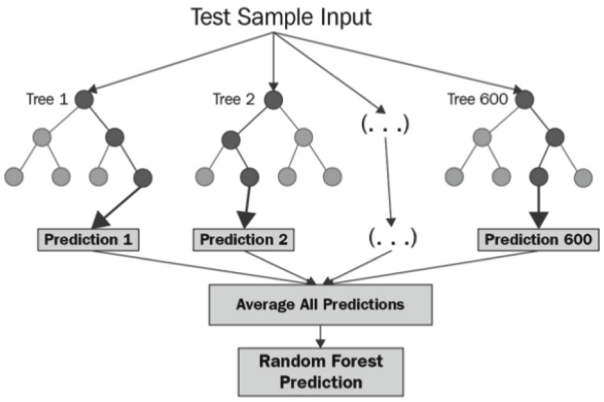

Fig. 3. Diagram above shows the structure of a Random Forest [13]

– Artificial neural networks (ANN), models inspired by the functioning of the human brain, capable of capturing complex and non-linear patterns from large data (Fig. 4).

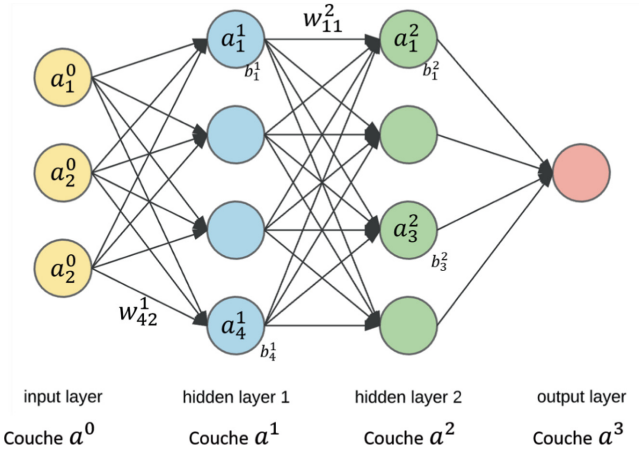

Fig. 4. Architecture of a multi-layer perceptron [14]

– Support vector regression (SVR), a machine learning method suitable for regression problems, particularly effective for dealing with non-linear data (Fig. 5).

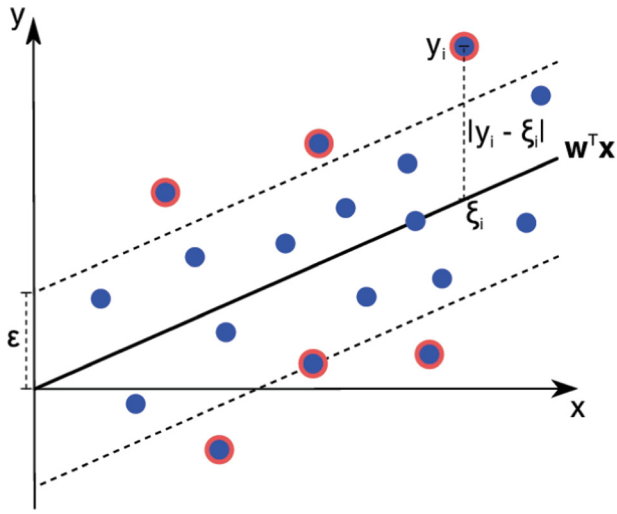

Fig. 5. Illustration of an SVR regression function [15]

We trained the models on the training dataset, using different optimization and hyperparameter tuning techniques to improve their performance.

2.4 Comparison Metric

We evaluated the performance of the models on a validation dataset to estimate their generalization capacity. We used metrics such as root mean square error (RMSE) and coefficient of determination (R^2).

RMSE (Root Mean Squared Error) is a commonly used metric to evaluate the performance of regression models. It measures the average difference between the values predicted by the model and the actual observed values. The lower the RMSE value is, the more accurate the model is in its predictions.

Mathematically, the RMSE is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{1}$$

where:

- $y_1, y_2, \ldots, y_n$ are the observed values
- $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are the predicted values
- n is the number of observations

The RMSE has the advantage of being in the same unit as the target variable, which facilitates the interpretation of the results. It further penalizes large prediction errors, making it a particularly suitable metric to evaluate the accuracy of the aflatoxin level prediction models in our study.

The coefficient of determination R^2 is a commonly used metric to evaluate the goodness of fit of a regression model. It measures the proportion of variance in the dependent variable (here, aflatoxin level) that is explained by the model.

Mathematically, the R^2 is calculated as follows:

$$R^2 = \frac{\sum_{i=1}^{T}(y_i - \bar{y})^2}{\sum_{i=1}^{T}(y_i - \hat{y})^2} \tag{2}$$

where:

- y_i are the observed values
- $\bar{y}$ is the average of the observed values
- $\hat{y}$ are the predicted values
- T is the total number of observations

The coefficient of determination R^2 varies between 0 and 1. A value of R^2 close to 1 indicates that the model explains a large part of the variance of the dependent variable, and therefore that the model has good predictive power. Conversely, a value close to 0 means that the model explains very little of the observed variance.

This metric will therefore be particularly useful to assess the overall quality of our aflatoxin level prediction models in the context of this study.

The in-depth analysis of performance metrics allowed us to objectively compare the models and select the one that offers the best predictions for our problem. This selected model will then constitute the basis of our final mathematical solution to estimate the aflatoxin level in corn.

2.5 Used Software and Tools

We implemented our models and metrics using the Python programming language, drawing on several key libraries. Pandas was used for manipulation and exploration of tabular data, while NumPy was employed for numerical calculations and processing data in the form of multidimensional arrays. Scikit-learn has played a crucial role in the application of machine learning algorithms such as linear regression, random forests, and neural networks, among others. For data visualization and graph creation to facilitate exploratory analysis, we used Matplotlib and Seaborn. The integration of these different Python libraries allowed us to efficiently carry out the different stages of developing our predictive models, ranging from data preprocessing to evaluating model performance.

3 Results and Discussion

3.1 Presentation of the Prediction Models Used

The analysis of the results made it possible to conclude that the random forest regression model (Random Forest Regressor) offers the best performance for predicting aflatoxin levels in maize. Indeed, this model obtained the lowest RMSE

(Root Mean Squared Error) value on the test set, indicating that it produces the most accurate predictions (see Table 3). Figure 6 presents the prediction accuracy of the different models.

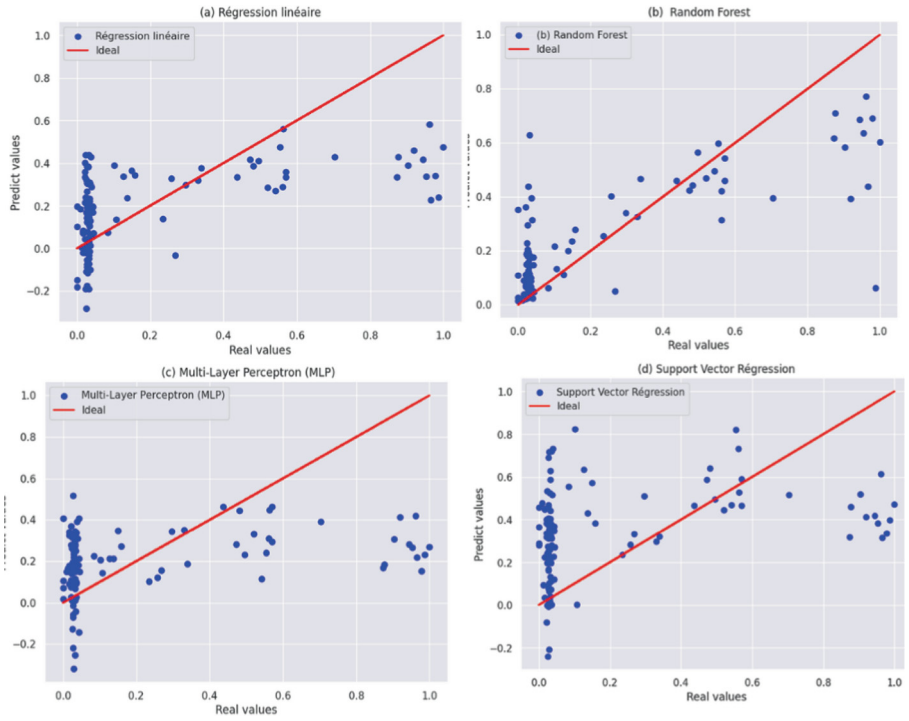

Fig. 6. Models Prediction accuracy: (a) Support Vector Regression (b) RandomForest Regression (c) Multi Layer Perceptron (d) Linear Regression.

Table 3. Metrics values for the different models

Models	Training		Test	
	R2	RMSE	R2	RMSE
Linear Regression	0.007	0.24	0.026	0.27
Random Forest Regressor	**0.81**	**0.14**	**0.81**	**0.29**
Support Vector Regression (SVR)	0.01	0.31	0.04	0.26
MLP Neural Network	4.63e–05	1.44	0.015	1.63

Several aspects deserve to be highlighted and interpreted in the results of this study:

The superiority of random forests can be attributed to their ability to capture the complex and nonlinear relationships between maize traits and aflatoxin levels. Random forests are robust to noisy data and can handle a large number of explanatory variables, making them a particularly suitable choice for this prediction problem.

Evaluating the performance of the models on a separate test set confirmed their generalization capacity, that is, their ability to maintain their accuracy when confronted with new data. This step is crucial to guarantee the robustness and reliability of the models developed in real conditions.

Finally, it should be emphasized that the validity of the results obtained is reinforced by their consistency with previous knowledge, in particular previous studies by MUGURE et al. [7] on storage conditions and practices post-harvest affecting aflatoxin infection in maize. This convergence between our results and existing research constitutes an additional confirmation factor of the relevance of our methodological approach and the reliability of our conclusions.

3.2 Pratical Implications for Farmers

Based on this study, farmers can reduce aflatoxin exposure by ensuring thorough drying of maize before storage, using airtight containers, maintaining clean and well-ventilated storage areas with low humidity, storing bags on wooden pallets instead of the ground, regularly inspecting stored maize, minimizing storage duration, and avoiding the shelling of wet maize. These practices, informed by our findings on the factors influencing aflatoxin levels, can significantly mitigate contamination risks, protecting both farmer livelihoods and consumer health.

4 Conclusion

In this study, we sought to use machine learning methods to predict aflatoxin levels in maize, which represents a major challenge for food security in the region. We explored several regression models, including linear regression, random forests, and neural networks, to evaluate their ability to accurately predict aflatoxin levels. Our results showed that random forest regression outperformed other models in terms of predictive performance, highlighting the importance of variables such as storage type and post-harvest practices in determining aflatoxin levels. This methodological approach constitutes a promising tool for actors in the agricultural sector, providing valuable information for risk management and strategic decision-making aimed at reducing consumer exposure to aflatoxins, and thus contributing to improving food safety and public health in the region.

References

1. Analyse de la taille et de la part du marché de la production de maïs en Afrique - Tendances et prévisions de croissance (2024–2029). https://www.mordorintelligence.com/fr/industry-reports/african-maize-market

2. Genserbe, B.M., Assoma, V.T., Kouame, K., N'guessan, B.V.H.: Machine learning appliquée aux images MSI Sentinel-2A pour la cartographie des ressources agropastorales dans le Fitri au nord du Tchad. Sciences Appliquées et de l'Ingénieur **5**(1), 57–64 (2023). issn: 2630-1164. http://publication.lecames.org/index.php/ing/article/view/29869
3. Hanadé Houmma, I., et al.: Drought vulnerability of central sahel agrosystems: a modelling-approach based on magnitudes of changes and machine learning techniques. Int. J. Remote Sens. **44**(14), 4262–4300 (2023). https://doi.org/10.1080/01431161.2023.2234094
4. Gouvernement du Bénin. Stratégie nationale pour l'e-Agriculture au Bénin 2020–2024. French (2020). https://faolex.fao.org/docs/pdf/ben210399.pdf
5. International Fund for Agricultural Development. International Fund for Agricultural Development. https://www.ifad.org/fr/web/operations/w/pays/benin
6. Kachapulula, P.W., et al.: Aflatoxin contamination of groundnut and maize in Zambia: observed and potential concentrations. J. Appl. Microbiol. **122**(6), 1471–1482 (2017). issn: 1364-5072. https://doi.org/10.1111/jam.13448. https://academic.oup.com/jambio/article-pdf/122/6/1471/47332725/jambio1471.pdf
7. Kamano, H.M., et al.: Storage conditions and postharvest practices lead to aflatoxin contamination in maize in two counties (Makueni and Baringo) in Kenya. Open Agric. **7**(1), 910–919 (2022). https://doi.org/10.1515/opag-2021-0054
8. Kamika, I., Ngbolua, H.N., Tekere, M.: Occurrence of aflatoxin contamination in maize throughout the supply chain in the Democratic Republic of Congo. Food Control **69**, 292–296 (2016). issn: 0956-7135. https://doi.org/10.1016/j.foodcont.2016.05.014. https://www.sciencedirect.com/science/article/pii/S0956713516302481
9. Lewis, L., et al.: Aflatoxin contamination of commercial maize products during an outbreak of acute aflatoxicosis in Eastern and Central Kenya. Environ. Health Perspect. **113**(12), 1763– 1767 (2005). https://doi.org/10.1289/ehp.7998. https://ehp.niehs.nih.gov/doi/pdf/10.1289/ehp.7998
10. Makowski, D., Chen, M.: Apprentissage supervisé pour simuler l'effet du changement climatique sur les rendements agricoles. In: INRAE, AgroParisTech, Université Paris-Saclay, France (2024)
11. Module Régression Linéaire. Consulté le 11 avril 2024. https://www.privateteacher.ch/Module-Regression-Lineaire
12. Muga, F.C., Marenya, M.O., Workneh, T.S.: Effect of temperature, relative humidity and moisture on aflatoxin contamination of stored maize kernels. Bulgarian J. Agric. Sci. **25**(2), 271–277 (2019)
13. Random Forest Regression. Consulté le 11 avril 2024. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84
14. Réseau de neurones : on va essayer de démystifier un peu tout ça. Consulté le 11 avril 2024. https://www.aspexit.com/reseau-de-neurones-onva-essayer-de-demystifier-un-peu-tout-ca-1/
15. Support Vector Regression (SVR): Illustration of an SVR regression function. Consulté le 11 avril 2024. https://www.researchgate.net/figure/Support-vector-regression-SVR-Illustration-of-an-SVR-regressionfunction-represented_fig12_248396465
16. Toutes les données sur la production céréalière. https://www.mccormick.it/fr/toutes-les-donnees-sur-la-production-cerealiere/
17. World Health Organization. Mycotoxines. https://www.who.int/news-room/fact-sheets/detail/mycotoxins

18. Wu, F.: Global impacts of aflatoxin in maize: trade and human health. World Mycotoxin J. **8**(2), 137–142 (2015). https://doi.org/10.3920/WMJ2014.1737. https://doi.org/10.3920/WMJ2014.1737
19. Yallou, C.G.: Le maïs au Bénin: atouts et perspectives. In: Direction de la recherche agronomique. https://www.fao.org/3/X5158F/x5158f0g.htm