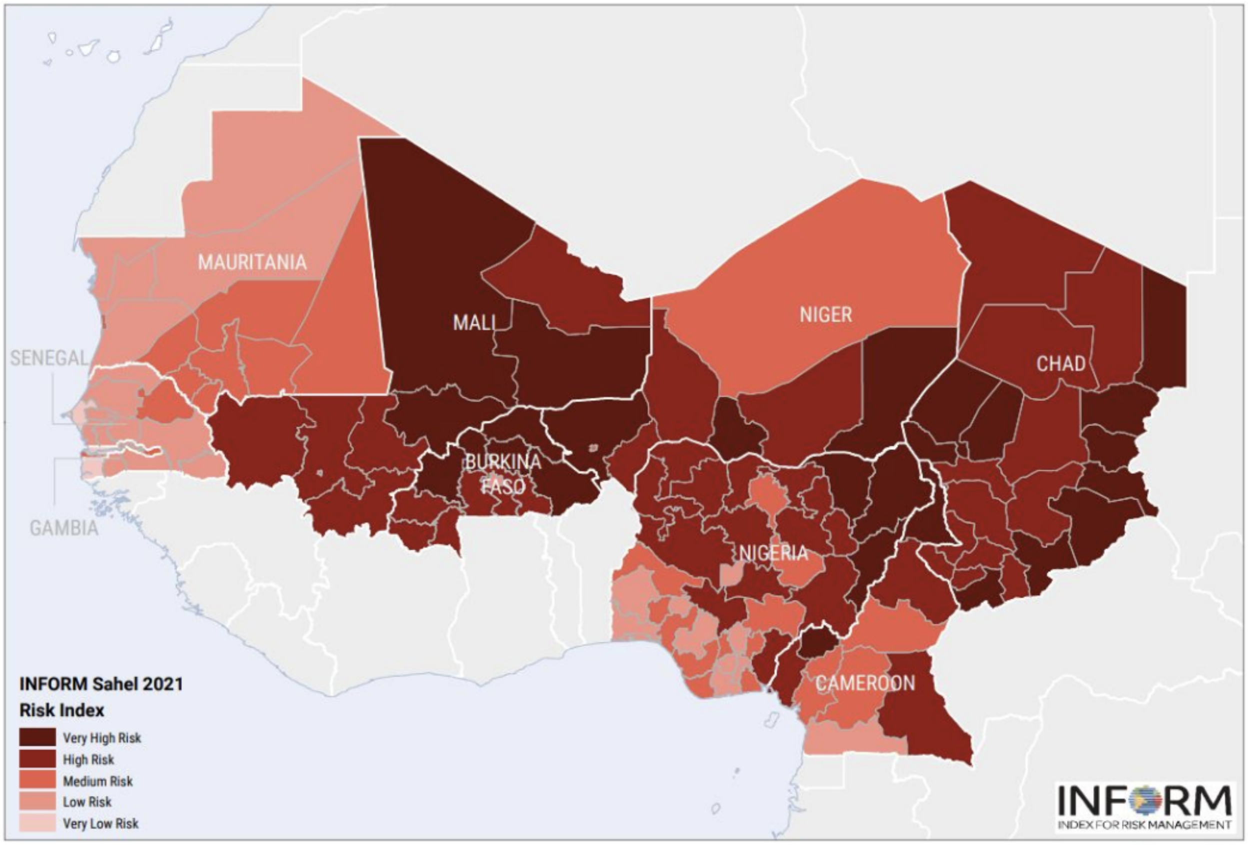
Prediction of the Hotspots of Child Malnutrition

Submission number: 25

April 20, 2023



September 2021

Contents

1 Introduction

Wasting is an acute form of malnutrition defined as a low weight-to-height ratio (WHO, 2023). It usually indicates recent and severe weight loss, and a weakening immune system, which increases the risks of infections and autoimmune diseases (Katona and Katona-Apte, 2008). The prevalence of wasting increases the risk of severe malnutrition (stunting), long-term negative consequences for development, and increased mortality (Karlsson et al., 2022). It has also been found that childhood wasting decreases the ability to learn and makes them less productive later in life (Isanaka et al., 2021). A recent study estimated 45.4M under 5 years were wasted, and 13.6M were severely wasted (UNICEF, 2021).

There is a strong need for prevention of all types of malnutrition. Targeting wasting specifically lowers the probability of developing more severe malnutrition, and can positively influence children to reach their physical and cognitive potential (McDonald et al., 2013). While some programs target the nutrition and well-being of mothers during pregnancy, other programs focus on improving already existing cases of wasting in children, improving coverage and care, or reducing the costs for families and communities. Many prevention programs have been effective in reducing childhood wasting cases, so the challenges lay in rapidly identifying hotspots where the burden of wasting is high.

In a given time period, the burden of wasting is determined by its prevalence at the beginning of the period and the incidence of new cases during that time. Prevalence estimates are often readily available to program managers, while incidence estimates are much harder to obtain. It is, therefore, crucial to identify regions where the burden (defined as prevalent + incident cases) is high, in order for program managers to immediately target with remediation efforts.

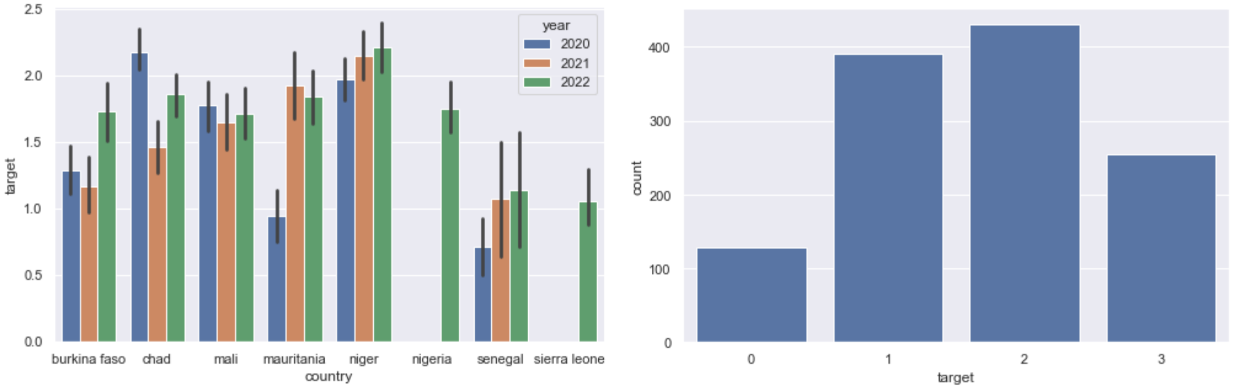
This paper presents predictions of the burden of wasting using three different machine learning methods. Due to the high number of missing values in the data set, we use methods that can handle this. The burden is presented as a categorical variable with four levels from zero to three, where zero represents a low burden while three represents a very high burden. Therefore, we present both exact and fuzzy accuracy metrics, which are more useful for policy-making decisions.

We start with the data description in Section 2. Section 3 provides an in-depth explanation of the prediction algorithms, while Section 4 presents the results and Section 5 presents policy recommendations and a Conclusion.

2 Data description

In order for a top-performing machine learning algorithm to learn patterns and make accurate predictions, it should be provided with data. Top-performing athletes mind their food intake before a big game. Food poisoning, for instance, would be catastrophic and leads to poor performance. In this case, the data can be considered the food of the model, so it is essential that it is of high quality. Mostly, datasets are contaminated and contain, for instance, duplicates or missing values. Also, it should be noted that the data assembled in, for instance, surveys or medical datasets may be subjective as it is collected by people, who may have different beliefs, opinions or interpretations. The initial dataset contains 1204 entries and 69 columns. Some duplicate columns such as 'lack_of_coping_capacity' and 'Infrastructure' were deleted. Most of the variables can be subdivided into political (e.g. governance, institutional), environmental (e.g. hazard and risk of flooding), health (e.g. diarrhoea and malaria fever) and socio-economic (e.g. inequality and conflict intensity) variables. We converted the categorical values to dummies or numerical values, depending on the nature of the variable. We constructed the following variables: 'level_1' is the region within a particular city, 'level_2' denotes the second level of administration (i.e. a particular region in a country), 'level_3' represents a city or village in a particular country.

In Figure 1a we exhibit the average targets with 95% confidence intervals for the different countries. We can clearly see that the average target level is lower for, for instance, Senegal and Sierra Leone. We observe the highest average target value in Niger. Also, we learn that Sierra Leone and Nigeria only collected data in 2022. Figure 1b exhibits the counts for each of the risk levels present in the target variable. We observe that the target value is imbalanced since we observe much more values of 1 and 2 compared to 0 and 3.



(a) Barplot denoting the mean of the target variable per country per year.

(b) Countplot presenting the number of counted values per malnutrition risk level.

Figure 1: Visualization of the target variable distribution.

Also, we evaluated the number of missing values in each column. Out of the 69 variables, 26 had more than 10% missing values. For seven, the percentage was even over 50%. Most statistical models cannot handle missing data, so we tried to impute the missing values by using a k-Nearest Neighbour imputing mechanism on the country level. In words, entries are compared to other instances in a particular country. Those instances are the so-called neighbours of the entry with missing values. Subsequently, the Euclidean distance between the entry and its neighbours is computed and used to find the k-Nearest Neighbours. Finally, the average of the values of the neighbours is used to impute the missing values. For the remaining missing values we used the median of the full column, or, in the case of categories, we added a category called 'Unknown'.

3 Model

Since the greatest challenge of our task was to deal with the amount of missing data present in the dataset we tried two different approaches. We first tried to impute the data as discussed in Section 2, but the amount of missing data was such that the approximation introduced in the data by the imputation was harming the prediction more than helping it. For this reason, we then decided to use a model that can handle missing values, in our case we used the histogram-based gradient boosting method. Our procedure was the following, since our goal was a one-step-ahead forecast the model is tuned and cross-validated on the subsample of data from 2020 using the target variable for 2021:

$$y_{T+1} = f(X_T), \quad \text{T=2020}$$

We then validate our results using the subsample of data from 2021 as an out-of-sample test sample, using the target variable from 2022 to validate our prediction. This means that we compute our out-of-sample accuracies using:

$$\hat{y}_{T+1} = f(X_{T+1}), \quad \text{T=2021}$$

Once we validated our approach and selected the best performing model we can join the dataset to train the whole engine on a bigger dataset and use this final model on the data from 2022 to produce some predictions for 2023 and discuss the implications of these predictions.

We will report two types of accuracy: the exact accuracy and a fuzzy accuracy measure. Our fuzzy accuracy measure will define as correct the predictions that fall in an interval ±1 from the true value. This measure gives good feedback about how useful is actually our point prediction. In practice, if we want to allow for policy intervention the most costly type of error we can make is to advise not to intervene when it is actually necessary or to suggest an intervention when it is not necessary. This means that the distinction between an observation marked as category 2 and an observation marked as category 3 is much less relevant than being able to distinguish between a category 1 and a category 3.
Then we have that our accuracy metric can be defined as:

$$p_{i,x} = \mathbb{1}_{[\hat{y}_{i,T+1} = y_{i,T+1}]},$$

for the exact accuracy and:

$$p_{i,x} = \mathbb{1}_{[(\hat{y}_{i,T+1} - y_{i,T+1})^2 \leq 1]},$$

for the fuzzy accuracy measure.

The second approach we used is to try to break down this complex problem into smaller tasks to allow the model to make the most use of the problematic dataset. We simplified the problem asking the engine to predict whether the situation in a given county/province is severe or not, in other words answering the policy question should we intervene or not? We then have a proxy target variable defined as:

$$y^*_{T+1} = \begin{cases} 1 & \text{if } y_{T+1} \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

By doing so we obtained a simpler problem for the engine in the form of a binary prediction that represents whether the target variable is smaller than a certain threshold.

The final step we took is to use the Mixture of Experts (MoE) framework. We divided the task in simpler tasks and assigned to each of them one model, an expert. Each expert produces its predictions and these outputs are fed to a gating model that combines them into a final prediction. Here below we present the structure of our MoE model.

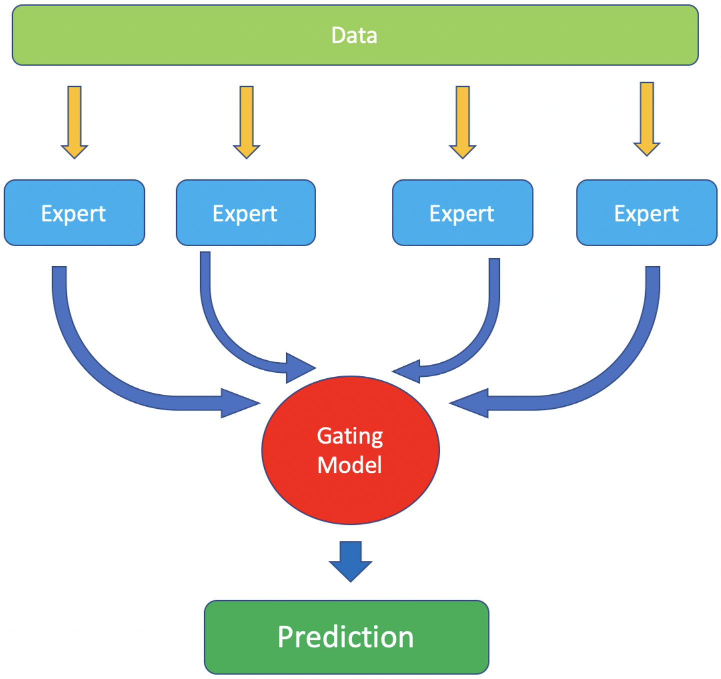


Figure 2: Structure of our Mixture of Experts

In this case we used as experts models that have the task to compute a simple binary prediction about the seriousness of the malnutrition problem in a given area. So we will have a binary prediction for the category extremely severe/ not extremely severe, one for severe/not severe and so on. In other words the expert i will be assigned to the proxy target variable $y^*_{i,T+1}$, defined as follows:

$$y^*_{T+1} = \begin{cases} 1 & \text{if } y_{T+1} \leq k_i \\ 0 & \text{otherwise} \end{cases}$$

where $k_i \in \{0, 1, 2\}$.

Finally we will use these models to perform some predictions about 2023 using the data available for 2022 and we will comment on these predictions.

4 Results

We will now discuss the results of the three approaches that we presented in the model section we have respectively:

- Histogram-based gradient boosting model;
- Binary prediction model;
- Mixture-of-Experts model.

We start by presenting the outcome of the Histogram-based gradient boosting model prediction in Table 1, even if the point prediction lacks some precision the fuzzy accuracy highlights the reliability of this measure for policy intervention. Figure 3 shows a confusion matrix of the predicted states of the Histogram Gradient Boosting model.

	Exact Accuracy	Fuzzy Accuracy
Model 1	51%	92%

Table 1: Accuracy for the Histogram Gradient Boosting model

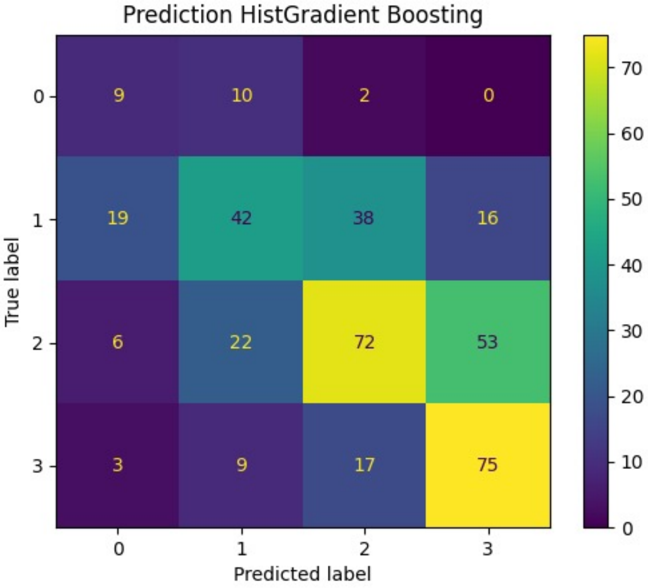


Figure 3: confusion matrix of the prediction of the target variable over the year 2021

Although most machine learning models are considered to be so-called ‘black-box’ algorithms, it is in most cases possible to extract the feature importance to identify the most important factors that drive the predictions. To quantify which variables affect our predictions most, we use Permutation feature importance. The permutation score of a feature is the amount the score of a model decreases after randomly shuffling the data of that feature. The larger the permutation score, the more important that feature was when making predictions. Figures 4 and ’5 show all features with a non-zero permutation score for the binary and the non-binary models. It is interesting to see that only a small fraction of the features, 6 and 14 respectively, had the most predictive power. The most influential factors for predicting whether the target is severe/not severe are food insecurity, the risk of the previous year (lags), the number of internally displaced persons (IDPs), severe acute malnutrition prevalence (SAM prevalence), natural risks and the city or village someone lives in. For the non-binary prediction problem (see Figure 5), the most important features are less pronounced. However, indicators of food security, and health again give the algorithm the most predictive power. Furthermore, features describing the geographical location are also among the most important variables. The latter indicates that policy might need to be region specific instead of for an entire country.

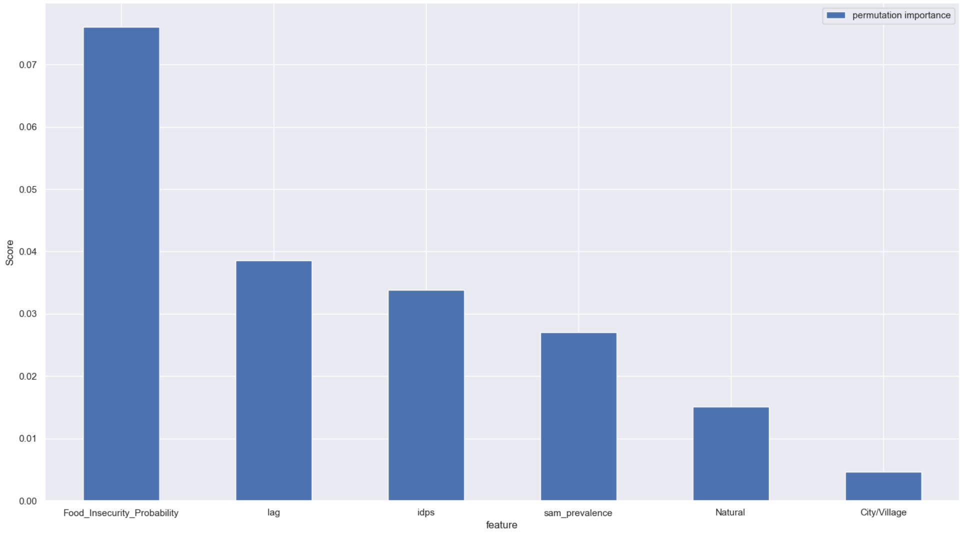


Figure 4: Permutation feature importance plot for the Binary Histogram-Gradient Boosting model

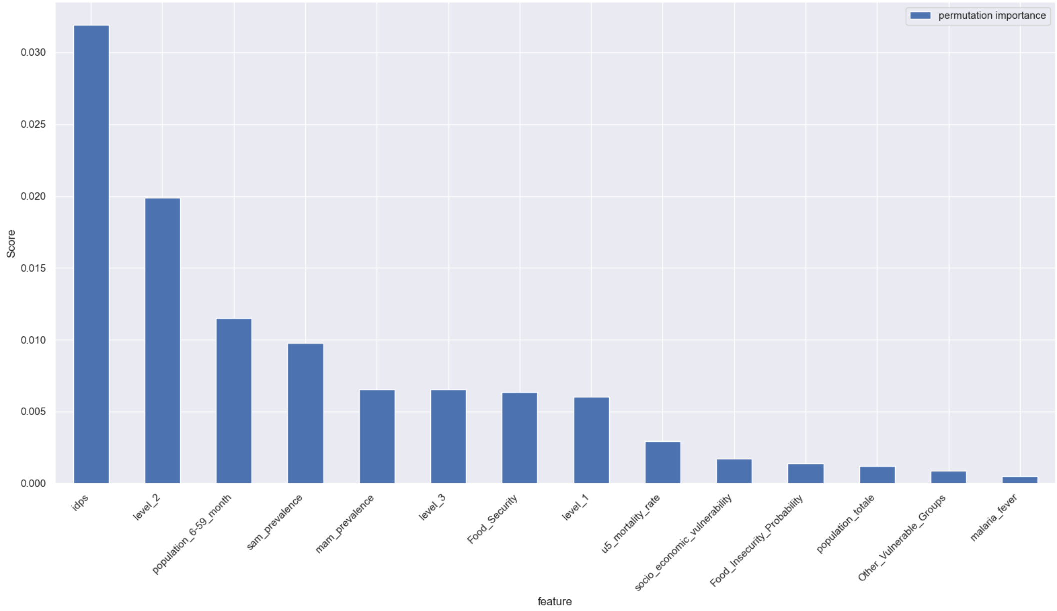


Figure 5: Permutation feature importance plot for the non-binary Histogram-Gradient Boosting model

The second model we present is the binary prediction model. In this case, the point prediction of our model improves significantly in accuracy thanks to the simplification of the task we give to the engine. The accuracy on our test sample is depicted in Table 2.

	Exact Accuracy	Fuzzy Accuracy
Binary Model	81%	·

Table 2: Accuracy for the Histogram Gradient Boosting binary model

The confusion matrix for the predicted states is presented in Figure 6.

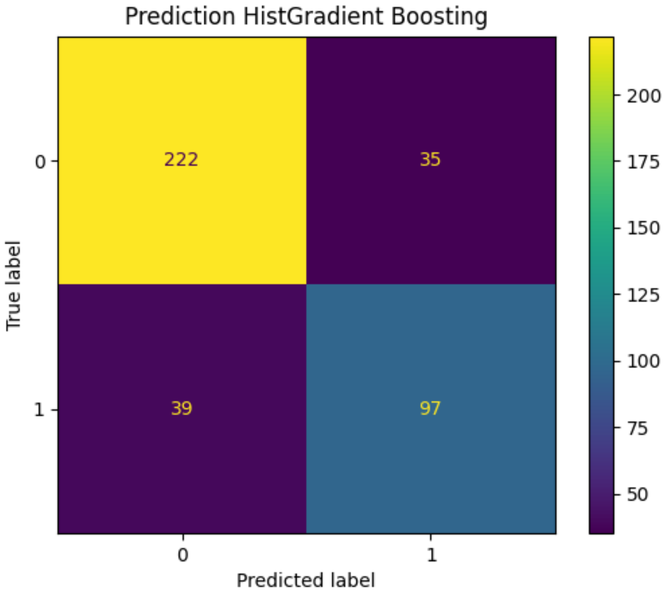


Figure 6: Confusion matrix of the binary prediction of the severe/not severe target variable over the year 2021

The final model we present is the MoE model. This framework allows our model to almost completely close the gap of the fuzzy accuracy to 100%. This makes sure that our predictions of the target variable, even if they are still not completely accurate, represent reliable predictions of the situation in the province/county of interest and can be used as the starting point for policy intervention. Here we show the accuracies of the MoE framework compared to the first model we presented:

	Exact Accuracy	Fuzzy Accuracy
Model 1	51%	92%
MoE Model	54%	97%

Table 3: Accuracy for the Histogram Gradient Boosting MoE model

From the table we can see the improvement achieved using the MoE framework. The confusion matrix for the predicted states of the MoE model is provided in Figure 7.

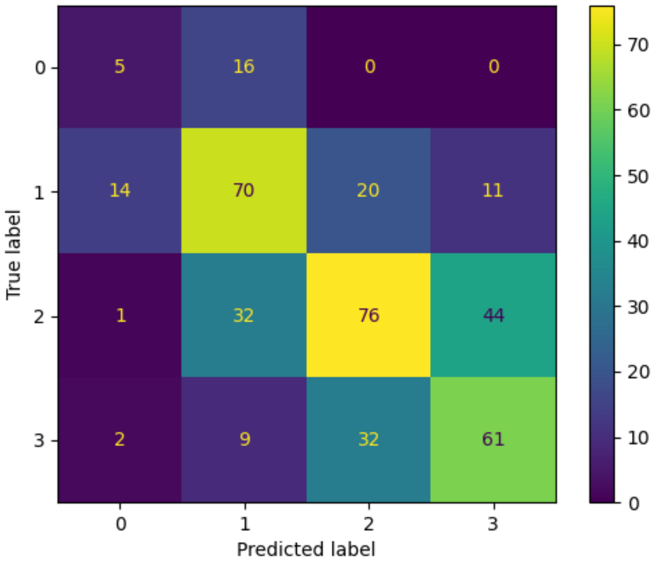


Figure 7: Confusion matrix of the MoE model for the target variable over the year 2021

5 Conclusion and policy recommendations

We will now conclude our findings and provide policy recommendations.

5.1 Concluding remarks

This paper proposes machine learning algorithms to predict the hotspots of child malnutrition in the Sahel region as accurately as possible. Because the amount of data was rather limited and contained many missing values, a (histogram) gradient boosting technique is exploited to find as many useful patterns within the data as possible. Imputing data using advanced machine learning techniques, such as kNN imputing, did not result in more accurate results. One of the advantages of the gradient boosting method is its flexibility, also when it comes to handling missing values.

The results show that the proposed method is very reliable in terms of policy interventions. When dividing the targets into not severe and severe, we find that the model predicts severity with high accuracy. Even more impressive is the fuzzy accuracy of the MoE histogram gradient boosting algorithm, which is close to 97%. Therefore, it is very reliable to construct policy recommendations from the models introduced in this paper. The fuzzy accuracy measure is more appropriate in this setting, as, first of all, the measurements of the target variable are subjective. Second, in practice, the most costly type of error we can make is to advise not to intervene when it is actually necessary. Therefore, a distinction between predictions 2 or 3 is surmountable, whereas the difference between 1 and 3 may result in not intervening at all, which may result in a high level of malnutrition.

In our analysis, we also elaborate on the feature importance. We can clearly see that variables such as food insecurity, the area and IDPs play a big role in predicting the hotspots of child malnutrition.

5.2 Policy recommendations

In the previous sections we discussed machine learning algorithms and evaluation criteria to quantify the problem. We will now translate our quantitative findings to the actual and more important problem: provide policy recommendations to minimize the risk of malnutrition among children. With that in mind, as discussed in 3, we recommend using either our binary prediction model or the histogram-based gradient boosting model, using the fuzzy accuracy measure, for two reasons. First of all, as explained by dr. Amrit, there is a high risk of measurement error in the target values. Therefore, a target value of 2 might mean a 3, or vice versa, for various survey takers. Second, the problem of interest is not merely predicting a subjective number, but rather intervening as accurately as possible, i.e. when there is a high risk of malnutrition.

Besides finding the most severe cases, it is essential to identify the factors determining the risk of malnutrition. From the permutation importance plot (see Figure 4 and 5), we know that the most important factors for predicting the target can be categorized into three categories: Food security, health, and geographical location. Those factors may sound trivial considering the problem statement, but it shows that our model learned from the data, which underscores the reliability of the proposed algorithm. This means that to prevent malnutrition, it is crucial to look at those factors first, as the model tells us that an increase in one of those factors most likely results in a higher risk of malnutrition. Furthermore, we might be able to increase the predictive power by better data collection in these categories. If we have fewer missing values, we might get better insights into the future. While data collection might be hard in some areas, focusing on these categories might alleviate some of the burdens. Also, our results indicate that policy should be tailored to specific areas since the severity of malnutrition differs between regions and cities.

Figure 8 presents the evolution of the distribution of malnutrition risk over the years. The first three years (2020-2022) are the observed years, whereas 2023 is our prediction. We see a clear increase in the number of (very) high malnutrition-risk areas over the years. Our model predicts that there will be even more areas in danger in 2023. Considering the current state of affairs, these results are, albeit shocking, not surprising. As explained by dr. Amrit, the Covid-19 pandemic has had a huge impact on food insecurity in the Sahel region. Also, the terrible war in Ukraine jeopardizes the food security in African countries. As predicted by our model and common sense, this results in higher malnutrition risks. Policy interventions should thus focus on food security, especially in times of worldwide crises such as Covid-19 and wars.

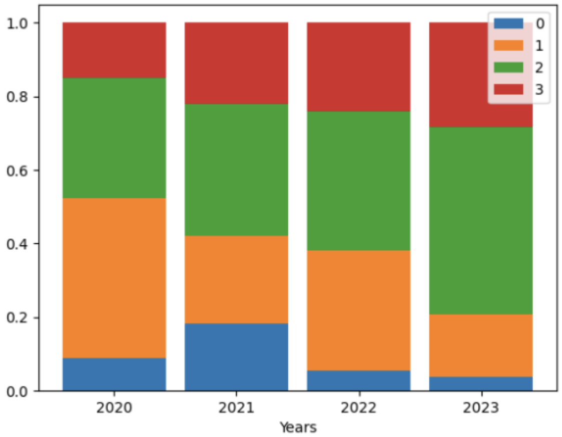


Figure 8: Evolution of the distribution of malnutrition risk over the years

References

Isanaka, S., Andersen, C. T., Cousens, S., Myatt, M., Briend, A., Krasevec, J., Hayashi, C., Mayberry, A., Mwirigi, L., Guerrero, S., and et al. (2021). Improving estimates of the burden of severe wasting: Analysis of secondary prevalence and incidence data from 352 sites. *BMJ Global Health*, 6(3).

Karlsson, O., Kim, R., Guerrero, S., Hasman, A., and Subramanian, S. (2022). Child wasting before and after age two years: A cross-sectional study of 94 countries. *eClinicalMedicine*, 46:101353.

Katona, P. and Katona-Apte, J. (2008). The interaction between nutrition and infection. *Clinical Infectious Diseases*, 46(10):1582–1588.

McDonald, C. M., Manji, K. P., Kupka, R., Bellinger, D. C., Spiegelman, D., Kisenge, R., Msamanga, G., Fawzi, W. W., and Duggan, C. P. (2013). Stunting and wasting are associated with poorer psychomotor and mental development in hiv-exposed tanzanian infants. *The Journal of Nutrition*, 143(2):204–214.

UNICEF (2021). *Levels and trends in child malnutrition.*

WHO (2023). Malnutrition. *World Health Organization.*