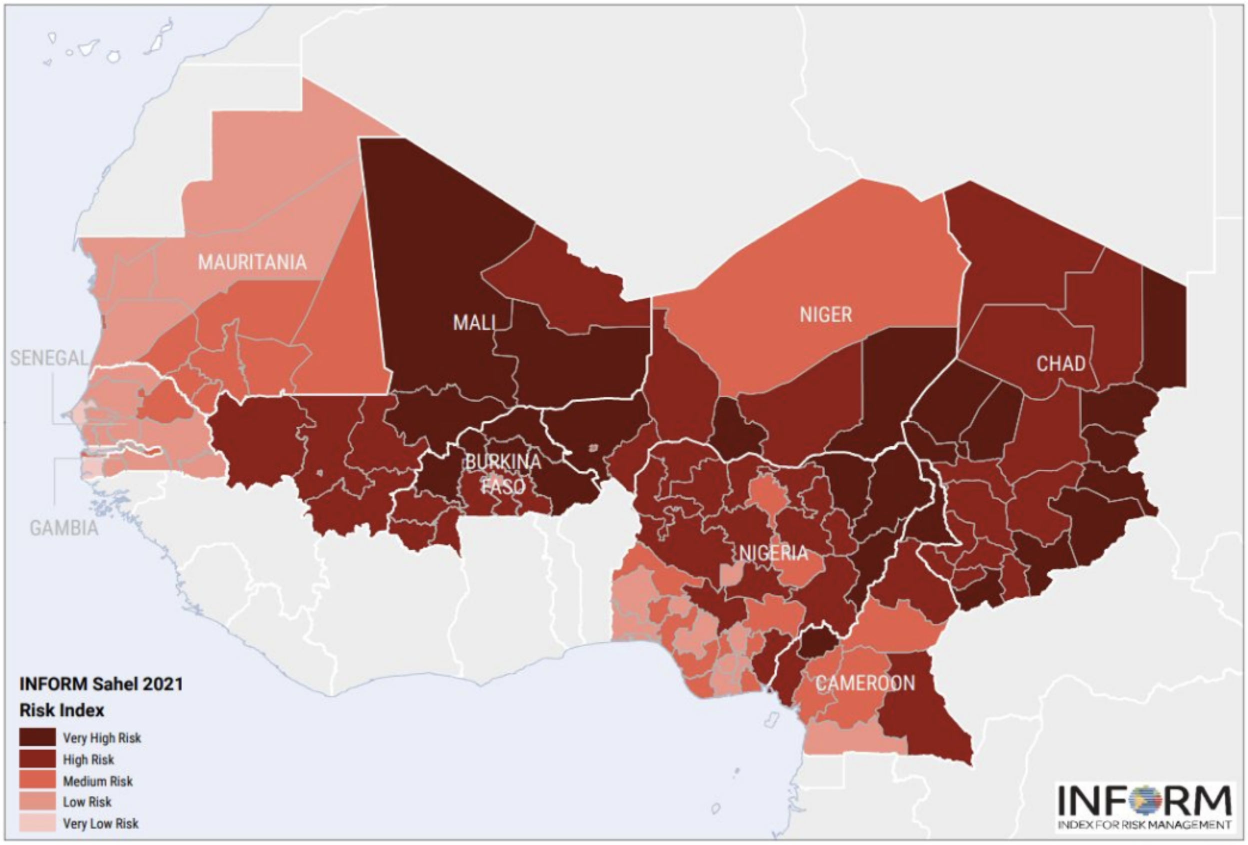
Prediction of the Hotspots of Food Insecurity in Chad

Gabriele Mingoli, Marcello de Wit, Benjamin Weggelaar and Barend Spanjers

Vrije Universiteit Amsterdam, Submission number: 25

April 25, 2023



September 2021

Contents

1 Introduction

Food insecurity is a persistent and complex issue that affects millions of people around the world. Food insecurity is defined as the lack of access to affordable and nutritious food that is essential for a healthy life. Food insecurity is caused by a variety of factors such as poverty, conflicts or climate. The consequences can be extreme for a given population since it can lead to chronic illnesses and malnutrition. Severe cases of malnutrition, such as wasting and stunting in children, can have long-term negative consequences for development, as well as increased mortality and morbidity (Karlsson et al., 2022). Addressing food insecurity through programs aimed at food affordability, access and distribution have been found to be effective (Fortin et al., 2015; Mokgomo et al., 2022). In order for program managers to intervene in certain regions, there is a need for identifying hotspots and asses the severity of the situation. Providing forecasts of hotspot regions for food insecurity can generate more time for program managers to organize and coordinate efforts to mitigate the impact of food insecurity.

Located in central Africa, the landlocked Sahelian nation of Chad has one of the highest rates of hunger in the world, with 42% of its people living below the poverty line (WFP, 2023). The country's hunger and poverty have been made worse by environmental degradation, rapid desertification, and conflict over depleting natural resources. Among those most impacted by the deterioration of the global climate are the people of Chad. The detrimental effects of COVID-19 on socioeconomic activities accentuate the poverty of a population that is already extremely vulnerable (Dasgupta and Robinson, 2021).

The aim of this paper is to support policymakers by offering food insecurity predictions at the administrative region level for the country of Chad. Food insecurity is classified according to the five phases of Cadre Harmonisé, where phase 1 represents minimal food insecurity and phase 5 represents famine. This paper continues with Section 2, which offers a description of the data. Section 3 explains in detail the methodology used for predicting food insecurity, while Section 4 offers the results. We finalize the paper with Section 5, which provides 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. Highly ranked chess players mind their food intake before and during a big game. It should be nutritious and easily digestible. Also, 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.

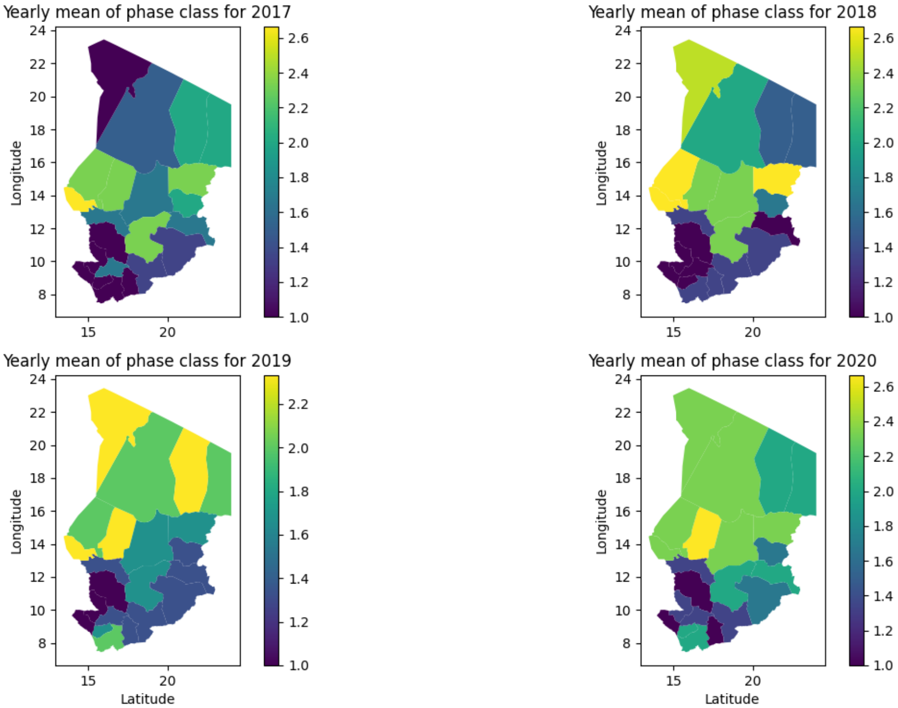
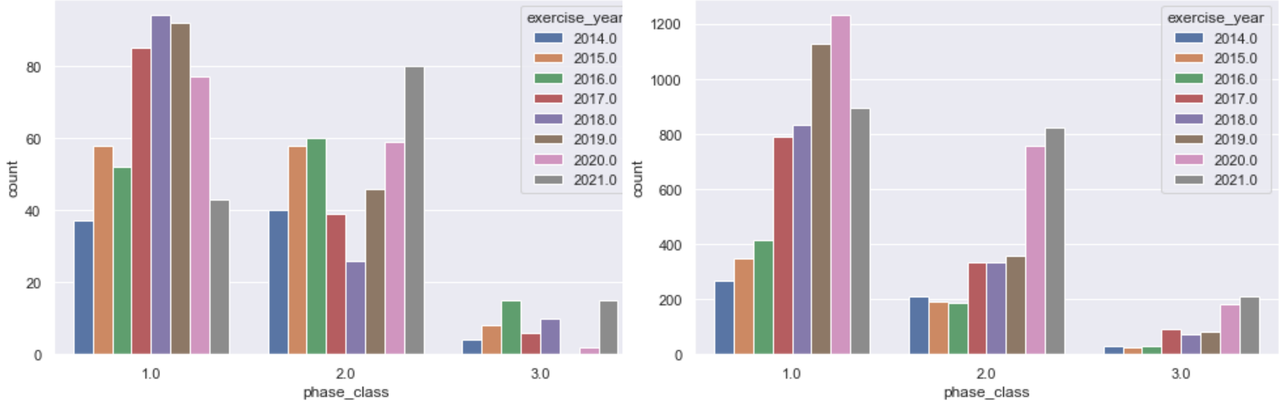

Yearly mean of phase class for 2017
Yearly mean of phase class for 2018
Yearly mean of phase class for 2019
Yearly mean of phase class for 2020
Longitude
Latitude
24
22
20
18
16
14
12
10
8
15
20
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0

Figure 1: Geographical Representation of the target variable

In figure 1 we constructed a time series maps of the target variable. Figure 1 shows the evolution of phase classes over time. We can clearly see that the phase increases over time. Furthermore, the classes differ per region. The south is more food secure compared to the north.

In order to provide accurate predictions, we make use of three separate datasets. The first dataset contains 331 rows and 31 columns, whereas the second dataset contains 19856 rows and 21 columns. The columns of the first dataset are subdivided into 7 categories: location, time, climate, disasters, conflict, the number of people in each phase, and cereal prices. The second dataset has more instances but less information about those instances and can be subdivided into location, time and phases. We aggregated the two datasets and merged food price data for Chad[1]. The data ranges from the years 2015-2021, where each year is subdivided into three categories: Jan-May, Jun-Aug, and Sep-Dec. Since the amount of features in the aggregated datasets is limited, we make use of a third dataset. This dataset contains information on food prices over the last few years. We only use the years that are in common with the aggregated dataset.

Figure 2(a) shows the counts per phase class for Chad over the years. It can be seen that, in most of the years, most of the phase classes were marked 1. In 2021, however, there is a sudden spike in phase class 2. This might be the result of the worldwide Covid-19 epidemic, as we can see that more mass is shifted from class 1 to 2 from 2019 on. When comparing Figure 2(a) with 2(b), we see that the distribution is more or less the same, where we see larger values for phase classes 2 and 3 in 2019 and 2020 compared to earlier years.



((a)) Countplot phase class variable for Chad

((b)) Countplot phase class full dataset

Figure 2: Visualization of the counts of the target variable for Chad (a) and the full dataset (b).

[1] We downloaded the data from https://data.humdata.org/dataset/wfp-food-prices-for-chad.

3 Model

The biggest challenge in dealing with this task was to deal with the unequal distribution of the data coming from different datasets, this resulted in most of the observations having only part of the features and in the end a great number of missing values. We rely on Histogram-Based Gradient Boosting methods to deal with this fragmented dataset. We use this method in the context of the Mixture of Experts (MoE) framework. The problem the engine faces can be split into subproblems such that we can assign a submodel (an expert) to each of these subproblems. There will be then a gating model to deal with the information produced by each of these experts. Figure 3 provides the structure of the framework:

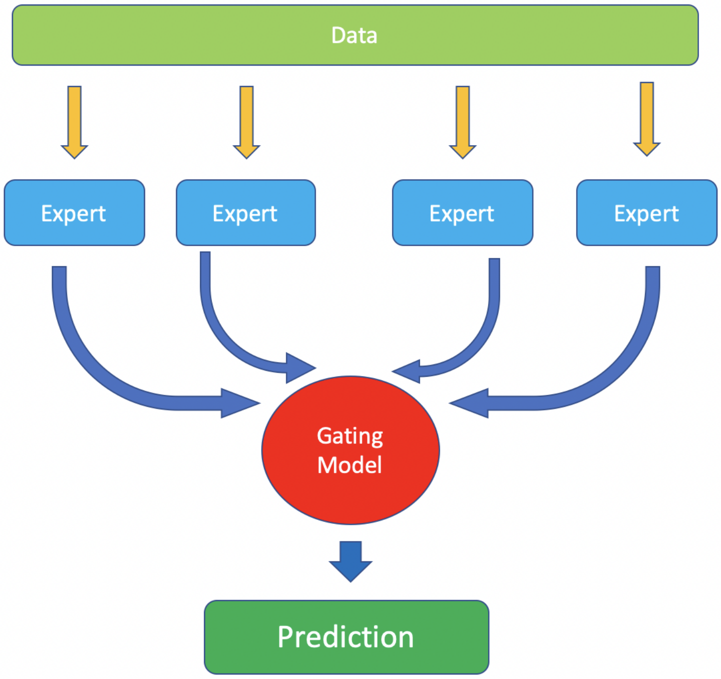


Figure 3: Structure of our MoE framework.

Given the nature of the problem seems natural to break down the prediction of the phase the next year into individual forecasts of the number of people in a given risk level. This allows us to define and train three different experts (categories 4 and 5 have too few observations so we merged them into a "risky" category defined by the union of 3,4 and 5) to predict what will be the number of people in each risk class the next year. Figure 4 exhibits the out-of-sample performances evaluated on the data from 2020 of these three individual experts.

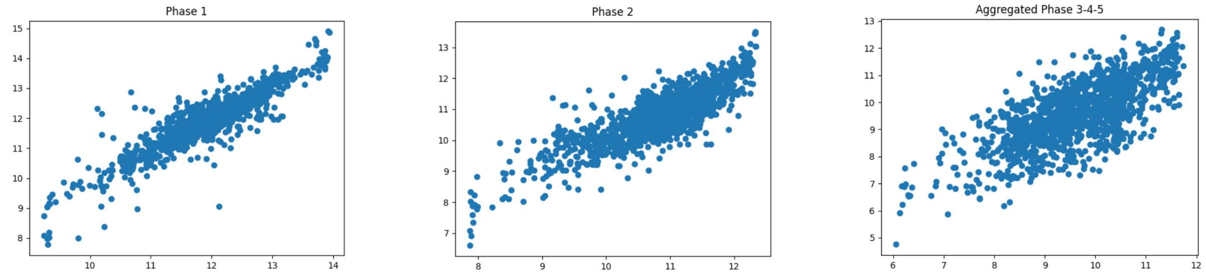


Figure 4: Out-of-Sample performance of three of our experts

For these experts we used a Histogram-based Gradient Boosting regression. We also included as experts the actual prediction of the future phase, so four experts that have as a target the relevant category of risk. In other words we use as target the proxy variable y_t^* defined as follows:

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

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

We use all the data up to 2019 to tune and cross validate the model. We then use the data in 2020 as an out-of-sample test set to validate our results. Once we validated our approach we can join the dataset to train the whole engine on a bigger dataset and use this final model on the data from 2021 to produce a one-step ahead prediction and discuss the implications of these predictions.

We will report two types of accuracy: exact accuracy and a fuzzy accuracy measure. Our fuzzy accuracy measure will define the predictions that fall in an interval ±1 from the true value as correct. This measure gives good feedback about how useful our point prediction actually is. 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 phase 2 and an observation marked as phase 3 is much less relevant than being able to distinguish between a phase 1 and a phase 3. The fuzzy accuracy measure is also appropriate in this setting because the phase class is constructed by simply taking the largest value of the phase counts. If the two largest groups are fairly close, it is surmountable to predict the phase class below (or above) the true class.

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.

4 Results

Our model allows us to identify with a reasonable amount of certainty the level of risk in a given area. The results on the test sample (2020) show good exact accuracy. Moreover, the performance under the Fuzzy Accuracy performance measure allows us to claim that our forecast has a high level of reliability for policy. In terms of policy intervention, the worst-case scenarios happen where an intervention is needed but it does not happen, or when it is not needed but we actually intervene. The almost sure level of certainty about the general level of risk makes sure that we rule out these scenarios. The results are summarized in Table 1. See Figure 5 for the confusion matrix of our predictions on the test sample.

	Exact Accuracy	Fuzzy Accuracy
MoE Model	63%	99 %

Table 1: Performance of our model

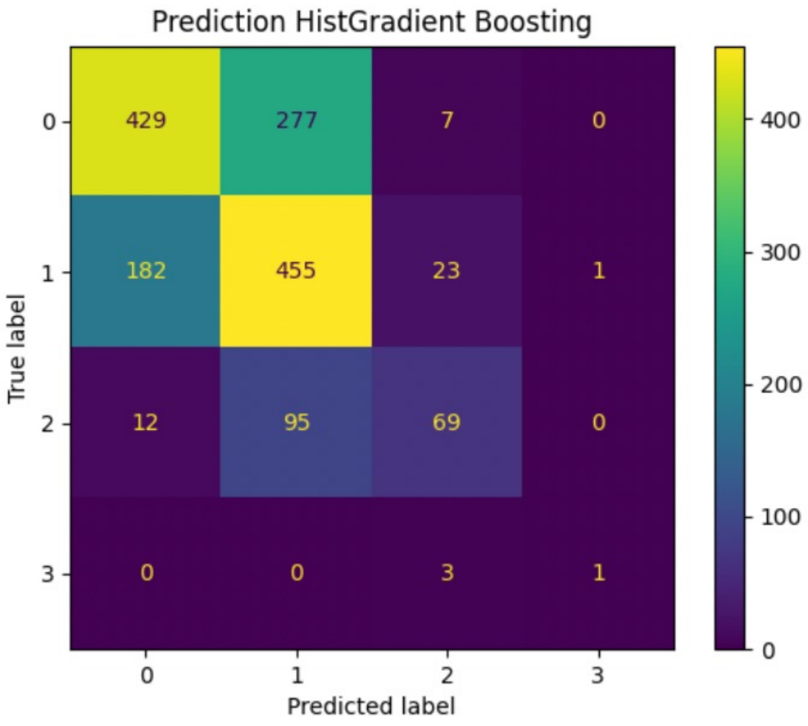


Figure 5: Confusion matrix of the forecasts on the out-of-sample test set

Another interesting point of view is to perform seasonal forecasts. In settings where the food production/distribution chain is not so stable, there can be important differences between the spring and the autumn period. The data allows us to identify the next season and we can use our engine to perform a shorter horizon forecast. The results are in line with what we showed before but we achieve a higher exact accuracy arriving to a 66% level, as one can see in Table 2. The confusion matrix of our seasonal predictions on the test sample is exhibited in Figure 6.

	Exact Accuracy	Fuzzy Accuracy
MoE Model	66%	99 %

Table 2: Performance of our model

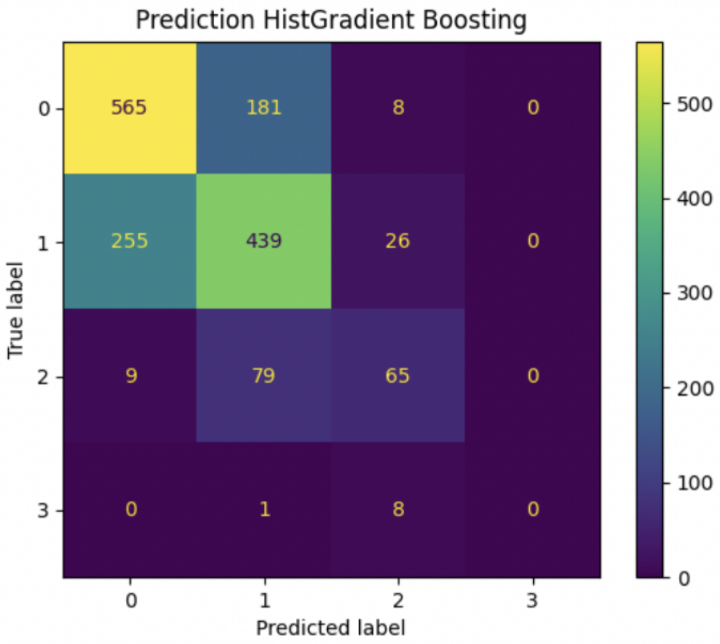


Figure 6: Confusion matrix of the seasonal-forecasts on the out-of-sample test set

Our model then allows us to produce geographical predictions of areas of risk by combining the engine and the information about the administrative levels. We present aggregated information at the regional level, compared to the predictions we presented in the confusion matrix there is some degree of approximation given by the aggregation.

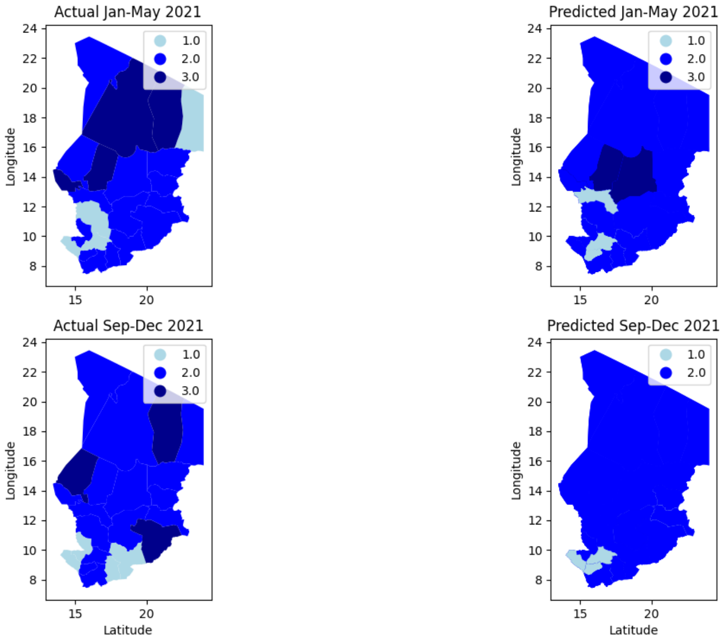


Figure 7: Geographical Representation of the Level of Risks

Figure 7 shows us the forecasts we produced for our out-of-sample validation sample. Having the actual data to validate the predictions we can compare the forecasts with the actual values of risk. In the map, it is possible to see how the engine identifies the relevant level of risks with a certain degree of approximation.

Another relevant step we can do towards policy intervention is to determine the most important factors of malnutrition risk. 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 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. Figure 8 show all features with a non-zero permutation score for our model. The most important variable was the region variable, indicating that it might be worth considering having different policies for different regions. Variables describing the food prices were also influential when making predictions. The higher the prices the more difficult it might be for people to afford food. Furthermore, while it is logical that the previous phase levels are very relevant for the next step prediction it is interesting to see how different phases have different weight and this could be due to different level of persistence. It is likely that it is hard to revert from a phase 3 setting so that is why this feature gets the most weight our of the three. Lastly, the inclusion of the 'year' variable suggests that food insecurity also differed between the years.

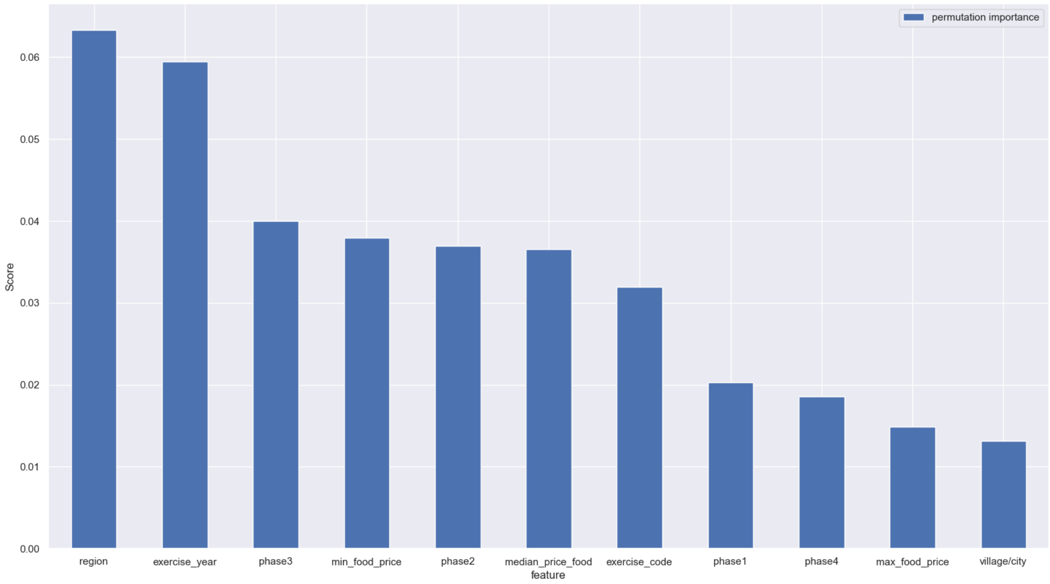


Figure 8: Permutation feature importance plot for Chad

5 Conclusion and policy recommendations

We will now conclude our findings and provide policy recommendations.

5.1 Concluding remarks

This paper proposes state-of-the-art machine learning algorithms to predict the hotspots of food insecurity in Chad as accurately as possible. Because the data was rather messy and contained many missing values, a (histogram) gradient boosting algorithm is exploited to find as many useful patterns within the data as possible, on a yearly and seasonal level. Some of the advantages of the gradient boosting method are its flexibility, also when it comes to large datasets, handling missing values and prevention of multicollinearity.

The results show that the proposed method is very reliable in terms of policy interventions. The exact accuracy of our model is 63% on the yearly level, and 66% on the seasonal level. This means that most of the forecasts are predicted correctly. Even more impressive is the fuzzy accuracy of the MoE histogram gradient boosting algorithm, which is 99% for both levels. 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, the phase class is constructed by simply taking the largest value of the phase counts. If the two largest groups are fairly close, it is surmountable to predict the phase class below (or above) the true class. The fuzzy accuracy measure accounts for this and is merely affected when the difference is larger than 1.

In our analysis, we also elaborate on the feature importance. We can categorize the most important features in three categories: Geographical location, food prices, and phases. These features might give insights on how to conduct policy, as will be discussed in the next section.

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 predict the hotspots of food insecurity. With that in mind, as discussed in 3, we recommend using either seasonal histogram gradient boosting model, for two reasons. First of all, the exact and fuzzy accuracy scores are quite impressive, meaning that the model is quite capable of predicting the level of food insecurity accurately. Second, focusing on the seasonal level, instead of a yearly level, is more interesting for policy intervention, because it is more efficient to intervene in seasons when it is actually necessary instead of the full year.

Besides predicting next year's or next seasonal's hotspot, it is essential to identify the factors determining the risk of food insecurity. From the permutation importance plots, we learn that the most important factors are 'phase1', 'phase2', 'phase3', food prices, and the geographical information. This implies that the count data of the previous period is very important and gives a good indication of which areas will be in danger in the next period. Also, our results indicate that policy should be tailored to specific areas since the severity of malnutrition differs between regions and cities.

Figure 9 clearly visualizes the high-risk areas in terms of food insecurity. As explained before, the severity of food insecurity differs per region and season. Therefore, we recommend using our prediction as an indication of the next period's risk areas in order to provide help as efficiently as possible.

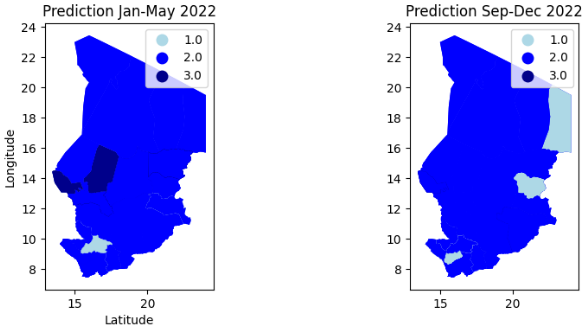


Figure 9: Geographical Prediction fro 2022

Lastly, we should note that we would 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.

References

Dasgupta, S. and Robinson, E. J. (2021). Food insecurity, safety nets, and coping strategies during the covid-19 pandemic: Multi-country evidence from sub-saharan africa. *International Journal of Environmental Research and Public Health*, 18(19):9997.

Fortin, S., Kameli, Y., Ouattara, A., Castan, F., Perenze, M. L., Kankouan, J., Traore, A., Kouanda, S., Conte, A., Martin-Prével, Y., and et al. (2015). Targeting vulnerable households in urban burkina faso: Effectiveness of geographical criteria but not of proxy-means testing. *Health Policy and Planning*, 31(5):573–581.

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.

Mokgomo, M. N., Chagwiza, C., and Tshilowa, P. F. (2022). The impact of government agricultural development support on agricultural income, production and food security of beneficiary small-scale farmers in south africa. *Agriculture*, 12(11):1760.

WFP (2023). Chad: World food programme.