Dissecting malnutrition and mal-governing - predicting level of malnutrition in Sahel region with machine learning methods

Team 23

Abstract

Severe malnutrition is a major problem in developing countries affecting more than 45M children aged 5-69 months worldwide. Acute malnutrition - wasting can lead to weight loss, weakness and fatigue, ultimately leading to death. Estimating the burden of malnutrition is difficult, currently it is measured by obtaining prevalence and incidence rates. The incidence correction factor (ICF) is commonly used to estimate incidents number, but it is argued that it is highly correlated with the geographical region under study, leading to possible under- or overestimation of burden. In this study we propose a new approach using machine classifiers to link the burden of malnutrition to known factors contributing to malnutrition, such as inadequate access to food, diseases and socioeconomic factors. Due to moderate quality dataset we made an extensive effort to clean the data and perform feature engineering. We compare it with a baseline methods using ICF and various naive and machine learning models. The best performing approach was Light Gradient-Boosting Machine (LGBM), with 65% of accuracy in predicting four levels of malnutrition, which is better than current approach. It is then explained using Explanatory AI (XAI) methods to identify key factors that influence the burden of malnutrition in different regions. We provide a set of policies that could help in lowering the level of malnutrition.

1 Introduction

For developing countries, malnutrition is still a lingering problem that varies by region and population. One of the results of malnutrition is wasting. It is a condition where a person's body is not getting enough nutrients, leading to a loss of muscle mass and body fat. It is characterized by a significant decrease in body weight, weakness, and fatigue. The United Nations reports that more than 45M children under the age of 5 were wasted globally in 2020 alone, where the majority of this occur in low- and middle-income countries. Such a process has a strong impact on children's development

opportunities, their ability to acquire knowledge, and ultimately can lead to their death (45% of the deaths are linked to the under-nutrition cases).

Uncovering the underlying processes that lead to malnutrition can increase awareness of the problem, leading to effective interventions and better policy-making. Unfortunately, the level of wasting incidents is hard to obtain, therefore researchers usually estimate it. Current framework calculates incidences, based on the prevalence rate, which is easier to obtain. The coefficient of this relationship is usually proposed to be 1.6 [1–3] and is called Incidence Correction Factor (ICF). The burden of wasting is then estimated as the sum of prevalent cases at the beginning of the observed period and incident cases over the entire period. The most common argument in the literature is that this factor is closely correlated with the geographical region under study, therefore the burden itself might be under- or over-estimated. Alternative methods, such as the capture-recapture method or Bayesian models, have been proposed to improve accuracy in estimating the burden of malnutrition.

What we propose is a novel approach, that assumes that the burden itself can be related to other, known factors that contribute to malnutrition, such as aforementioned regional and population-specific factors. These could be mainly inadequate access to food, poor eating habits and conditions of primary healthcare provision, but also socio-economic factors. Using data from a large scale survey, enriched with publicly available data on sociodemographics of the studied regions of the Sahel, we have build a set of machine learning multiclass classifiers. Our approach takes into account both the prevalence and incidence of malnutrition, as well as factors such as disease prevalence and demographic data. We compare it with the baseline method employing ICF, as well as different approaches to composing machine learning model. Based on the top performing approach we focus on exhaustive explanation of the model implications using explainable AI (XAI) methods, providing an explanation to what are the key factors influencing burden of malnutrition.

Based on an initial literature review, the hypotheses that we state in this study are:

1. machine learning based methods are able to significantly outperform traditional approaches (based on ICF = 1.6) when predicting burden of wasting for a given region.
2. female employment has a significant, negative relationship with the predicted burden of wasting for a given region.
3. democratization of a nation state has a significant negative effect on the predicted burden of wasting for a given region belonging to the nation state.

2 Literature review

The literature on malnutrition is very exhaustive. There are numerous studies related to both medical conditions that underlay stunning and acute malnutrition among children. When it comes to the estimation of the emerging hotspots the literature suggests using the ICF - incidence correction factor [1, 4].

However, the main flaw of such approach is that the ICF should be dependent on the particular region, as the risk of an incident can be dependent on the country or region specific variables. A study by [1] aimed to estimate the incidence correction

factor to improve program planning and inform the approach to burden estimation for severe wasting. The researchers calculated correction factors from 352 sites in 20 countries and found that the burden of severe wasting is often underestimated whe using the standard recommended ICF equal to 1.6. The study recommends the application of updated incidence correction factors as a simple way to improve program planning when incidence data are not available and to inform the approach to burden estimation.

Another study [5] compares the effectiveness of two methods of estimation of incidents in northwestern Nigeria, using active and adaptive case finding (AACF) and capture-recapture design. The sensitivity of AACF was found to be 69.5%, and 91.9% with capture-recapture case finding. The study highlights the importance of adequate estimation of number of incidents, given the potential impact of incomplete (underestimated) number of cases, since a speedy alerting can significantly reduce the risk of death [3].

There are numerous studies reporting significance of socio-economical and medical variables on the level of nutrition among children. For example, a study by [6] examines gender inequality in nutritional status among children under five years of age in a rural Bangladesh. The study found that 33% of children were severely malnourished, with 54.2% being female and 45.8% male. Existing gender gap persisted even after controlling for other variables, with female children being 1.44 times more likely to be severely malnourished.

Another study by [7] indicates that wealth, education, and ethnicity have significant impact on the malnutrition in ten Latin American countries, based on nationally representative surveys conducted between 2005 and 2017. Socially disadvantaged groups had higher prevalence of anaemia, while overweight was equally distributed among children, and education was a protective factor among adult women.

From the healthcare point of view, there is also an abundant literature regarding disease related influence on malnutrition. For example [8] aimed to identify the influential risk factors associated with physical and mental development in infants born to women with HIV virus in Tanzania. Preterm birth, child HIV infection, stunting, and wasting were independently associated with lower child development scores. Their recommendations are that the policymakers should focus on preventing these factors to improve child development. Other researchers [9] test the relationship between infection and malnutrition. They mention undernourishment can lead to poor growth, impaired intellect, increased mortality, and susceptibility to infection, while malnutrition (on overall) is a significant and independent contributor to morbidity and mortality in individuals with HIV. Yet another study [10] tackles anaemia, fever and malaria occurrences in relation with malnutrition in Burkina Faso. The prevalence of malnutrition and anaemia was high, and exclusive breastfeeding was found to be associated with a lower prevalence of malaria.

All of these researchers put emphasis on the effect of malnutrition of the pregnant women and their children on the human capital and development of the researched regions. In particular, [11] review the links between maternal and child malnutrition on adult health outcomes, using data from five long-standing cohort studies in developing countries. Malnutrition has been found to be associated with lower adult height,

lower economic productivity and lower offspring birth weight, while lower birth weight and malnutrition in childhood are risk factors for high glucose concentrations, blood pressure and harmful lipid profiles. The authors conclude that malnutrition in early childhood can lead to permanent impairments and chronic diseases and that prevention could bring significant health, educational and economic benefits.

According to the previous research, the quality of governance and democratization are significant drivers of the welfare of children [12]. We believe that it is hence expected that democratization should be a significant factor in predicting the burden of wasting in a given region. As democratization and the quality of governance are innately multidimensional and complex measures, we use several other indicators to describe them in this paper. The indicators we have decided to include in place of those composite measures are the corruption level indicator, rule of law indicator and freedom of discussion indicator among others.

3 Data and Transformations

The data that we used in this study comes from the nutritional surveys that were conducted in the Sahel region of Africa, enriched with the data from the publicly available INFORM dataset [13]. The target variable is the level of malnutrition in the region of Sahel, encoded as:

- Low $\rightarrow$ 0,
- Medium $\rightarrow$ 1,
- High $\rightarrow$ 2,
- Very High $\rightarrow$ 3.

All the variables are encoded on three administrative levels: country, admin level 1 and 2. The data was merged based on these administrative indexes and a year, for which we have observations throughout 2020-2022.

Fig. 1: Histogram of target variable distribution throughout the years

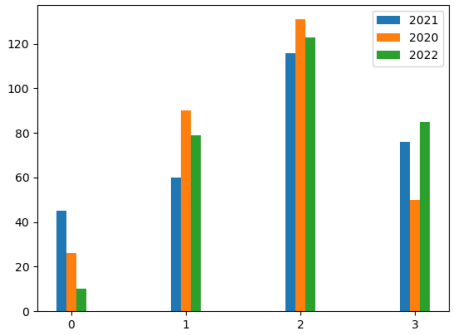


Table 1: Mean and mode for studied countries throughout the years

year	country	mean	mode
2020	burkina faso	1.33	1
	chad	2.19	2
	mali	2.09	2
	mauritania	0.95	1
	niger	1.97	2
	senegal	0.71	1
2021	burkina faso	1.29	1
	chad	1.57	2
	mali	1.84	2
	mauritania	1.93	2
	niger	2.15	3
	senegal	1.07	2
2022	burkina faso	1.84	1
	chad	1.97	2
	mali	2.02	2
	mauritania	1.84	2
	niger	2.21	3
	senegal	1.14	2

Based on figure 1 and table 1, we can clearly notice that throughout the years the overall burden of malnutrition has increased. It might be caused by COVID pandemic crisis and the war in the Ukraine. However there are few regions that have lower mean compared to previous years.

The data that we obtained was heavily preprocessed due to multiple issues and mistakes in the data.

We started with cleanup of the initial index, including:

1. typos correction - we've corrected *ADMIN_2_Admnistratif*, there were similar name matching, differences in letter cases. Thus, we changed all values to lowercase. Additionally blanks spaces were removed and replaced with dash if necessary,
2. removal of data for Nigeria and Sierra Leone, for those two countries the data was poor quality and 2020 was the only available year. Due to this data shortage and lack of values for multiple columns we decided to remove it,
3. for Chad and few other regions, we noticed that for 2020 all admin2 subregions were merged into one region at admin1 level. We exploded data for 2020 into multiple admin2 regions,
4. for Chad there were two columns - *mam_659_mois_prevalence*, *facteur_de_correction_incidence_mam* for year 2021 with multiple outliers with highly outstanding values. We replaced those values with mean for the year 2020 of the corresponding admin2 region. We didn't take 2022 year into mean to avoid data leakage.

5. due to very high correlation (above 95%) we removed one of two columns that are most probably duplicates - *infrastructure*, *socio_economic_vulnerability, infrastructure, lack_of_coping_capacity*

At the end, all the data was aggregated on the level named *ADMIN_2_Admnistratif*, so that all other indices were corrected and fixed on that level, aggregating by mean if there were duplicates caused by other admin2 indices. We have expanded the feature set by firstly adding new data and then performing transformations mentioned in table 2. New variables that were added include: democratization level, percentage of population with access to fresh water access and sanitary appliances, gender inequality, fetched from V-DEM dataset [14].

To cover larger data sample, we tackled the issue of missing data. We have excluded every variable that had fewer than 50% of observations, every case with more than 50% of observations had the missing values imputed with mean for the year for a specific country, if it was numerical variable and with mean for categorical variables. For every model training and evaluation, the dataset has been divided into training and testing samples and the missing for both training and testing were imputed using mean from training data.

Finally Spearman correlation, mutual information and Boruta algorithm [15] have been applied to measure the impact of explanatory variables on the target variable. The results are also presented in table 2, where feature selection is ordered from the most impactful column to the least. After data cleaning and preprocessing, we were left out with 297 observations for each year, which is 891 in total.

We applied logarithm transformation on following columns: *population_totale*, *population_6.59_month*, *gam_prevalence*, *sam_prevalence*, *mam_prevalence*. *u5_mortality_rate* was binned due to high concentration of values (0.10 and 0.20). We added lagged target variable to cover autoregressive factors. Lastly we embedded categorical variables using CatBoost encoder.

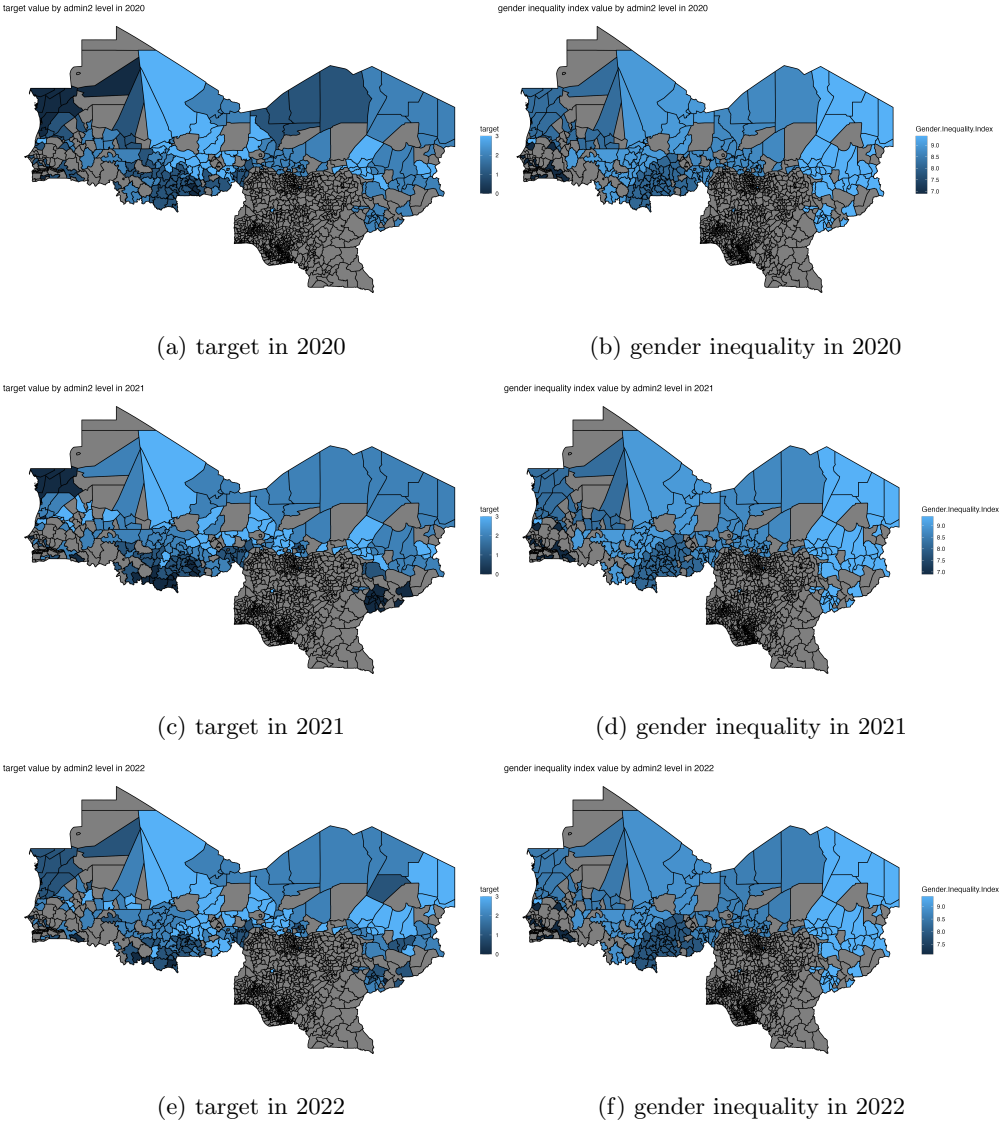
We tried to lower the dimensionality of the data using PCA algorithm, creating one variable from each group

1. health_care
2. prevalence_factor
3. infrastructure
4. governance
5. democratization

However, the final variables significantly decreased model predictive power on the training data and for that reason we decided to not use variables created by PCA algorithm.

After data cleaning we can have one last look onto the geographical distribution of the target variable, as well as gender inequality in figure 2. From the below figure it is not clear that target variable has strong relationship with gender inequality. There are many regions in which target increased, but gender inequality decreased.

Fig. 2: Geographical distribution of target variable and gender inequality in the studied region. Lighter colors indicate higher values.



(a) target in 2020

(b) gender inequality in 2020

(c) target in 2021

(d) gender inequality in 2021

(e) target in 2022

(f) gender inequality in 2022

Table 2: List of variables used in the study with their feature importance index and transformations

variable name	original dataset	added feature	embedding	binarized	logarithmized	lagged	feature selection
country	x		x				-
ADMIN_2_Admnistratif	x		x				-
year	x						22
admin1	x		x				-
lagged$_t target$	x		x			x	4
population_totale	x				x		39
population_6.59_month	x				x		21
gam_prevalence	x				x		-
sam_prevalence	x				x		-
mam_prevalence	x				x		-
mam_6.59_mois_prevalence	x				x		32
facteur_de_correction.incidence_mas	x						-
facteur_de_correction.incidence_mam	x						33
sam_u5_rate	x						34
u5_mortality_rate	x		x	x			23
diarrhee	x						35
malaria_fever	x						30
acces_au_structures_sante	x						14
pct_ch.ph_3.5	x						40
infrastructure	x						17
lack_of_coping_capacity	x						19
gam_rate_enfants_u5	x						38
stunting_rate_u5	x						36
Food_Insecurity_Probability	x						-
Physical_exposure_to_flood	x						-
Land_Degradation	x						-
Droughts_probability_and_historical_impact	x						15
Natural	x						-
Political_violence	x						1
Conflict_probability	x						9
HAZARD	x						20
Development_._Deprivation	x						16
Inequality	x						-
Aid_Dependency	x						37
Socio.Economic_Vulnerability	x						6
Uprooted_people	x						-
Health_Conditions	x						12
Children_U5	x						18
Malnutrition	x						25
Recent_Shocks	x						-
Food_Security	x						31
Other_Vulnerable_Groups	x						-
Vulnerable_Groups	x						24
VULNERABILITY	x						-
DRR	x						7
Governance	x						29
Institutional	x						27
Communication	x						11
Physical_infrastructure	x						-
Access_to_health_care	x						13
LACK_OF_COPING_CAPACITY	x						-
RISK	x						-
Mortality_rate_under.5	x						-
ACLED	x						5
Conflict_Intensity	x						26
Improved.Sanitation.Facilities		x					-
Improved.Water.Source		x					-
Gender.Inequality.Index		x					28
v2x_corr		x					8
v2x_polyarchy		x					3
v2x_rule		x					10
v2xcl_disc		x					35
v2xcl_prpty		x					2

4 Models

The problem of predicting burden is complicated and it is hard to assume a priori decision boundaries for changing the level of possible malnutrition. Thus we made a broad and exhaustive research of different econometric and machine learning models. In this section we present a brief description of the models we used in this research, our approach to the modelling and performance criteria.

4.1 Algorithms

4.1.1 Naive approaches

Naive approaches do not assume usage of any statistical models. We created benchmark results to be able to compare our results in the most objective way. First naive approach assumes just random class drawing for every prediction, with estimated accuracy being 25%. Another two approaches assume taking results from the preceding year - first simply per administrative region, second using mode of the level of malnutrition for a specific country. Last naive approach is the benchmark approach that is currently used to estimate number of incidents with ICF [1–3]. We used following equation:

$$Burden = SAM_Prevalence * (1 + ICF) * Population_{6-59months} \tag{1}$$

to calculate numerical burden estimate. To obtain simplified categorization we mapped it to 0-1 scale and binned the space in the following way:
- $[0.00 - 0.25) \rightarrow 0(Low)$,
- $[0.25 - 0.50) \rightarrow 1(Medium)$,
- $[0.50 - 0.75) \rightarrow 2(High)$,
- $[0.75 - 1.00] \rightarrow 3(VeryHigh)$.

4.1.2 Ordered logit/probit

We used two types of ordinal regression - the ordered logit and ordered probit model. Ordinal regression is a statistical model used to analyze and predict the results of ordinal dependent variables. It is commonly used if a dependent variable has ordered levels – in our case “low”, “medium”, “high” and “very high”. The model estimates the probability of an event occurring in one of the categories, given a set of independent variables. Ordinal regression assumes that the levels of the target variable are ordered in the way that the distance between the categories is equal. From the output of the regression we can obtain estimates of the coefficients for each independent variable, as well as the odds ratios and confidence intervals. Ordinal regression to model malnutrition was used for example by [16] to estimate the key risk factors of malnutrition in Bangladesh.

4.1.3 Multinomial logistic regression

Multinomial logistic regression is an approach that estimates the choice of a categorical outcome variable, which can't be ordered in a logical way e.g. colours. It is also used for data, when the distance between the levels is unknown – we don't have the specific distance between burden levels, which is a key factor for which we have estimated this model. Multinomial logistic regression has been used by [17] to estimate malnutrition in India.

4.1.4 Random forest

Random forest is a first of a set of machine learning algorithms used in this study. It employs an ensemble of decision trees to make predictions. It works by creating multiple decision trees on randomly selected subsets of the data and then combining their predictions to make a final prediction. This helps to reduce over-fitting and increase the accuracy of the model. Random forest can be used for both classification and regression tasks and is a popular algorithm for solving complex problems Random forest has been also used to estimate malnutrition, as well as it is well acknowledged in various fields such as finance, healthcare, and marketing, specifically in medical studies [18–20]

4.1.5 Naive Bayes

Naive Bayes methods are a set of supervised learning algorithms that are based on Bayes theorem with the "naive" assumption of conditional independence between every pair of variables given the value of the target variable. The main difference between naive Bayes classifiers is the assumption made regarding the distribution. In our research we chose The Gaussian Naive Bayes algorithm, which is used for classification, where the likelihood of the features is assumed to be Gaussian.

4.1.6 Support Vector Machines

The Support Vector Machines (SVM) is one of the fundamental non-parametric machine learning algorithms. The main author of this model is Professor Vladimir Vapnik [21]. The general idea of SVM is as follows: in a multi-dimensional space there exists a hyperplane which separates the classes in optimal way. The goal of SVM is to find the hyperplane, which maximizes the minimum distance (margin) between this hyperplane and observations from both classes. SVM for malnutrition estimation problem has been used for example by [22] to predict the malnutrition levels in Bangladesh.

4.1.7 K-nearest neighbours

The K-nearest neighbours (KNN) algorithm is a basic and probably the simplest supervised machine learning algorithm for both classification and regression problems. Its based on a simple idea - the best prediction for a certain observation is the known target value (label) for the observation from the training set that is most similar to the observation for which we are predicting. KNN's main advantage is that it's

non-parametric (it does not require the assumption of a sample distribution) and instance-based (it does not carry out the learning process directly - it remembers the training set and creates predictions on the basis of it on an ongoing basis). It was also used to prediction of malnutrition in Bangladesh [22, 23].

4.1.8 CatBoost and NGBoost

CatBoost is a machine learning algorithm for gradient boosting on decision trees. The algorithm was constructed in the way it is able to handle missing values, feature interactions, and high-dimensional data. Another approach that we used was NGBoost, which also is a boosting algorithm that uses natural gradient descent to optimize the model. The difference between the two is that CatBoost estimates the expected probability for every sample, while NGBoost is able to provide a probability distribution for the whole population, due to usage of Kullback-Leiber deviation. Both the algorithms are well-recognized in the machine learning state-of-the-art [24, 25].

4.1.9 AutoML

To reduce the time to provide reliable machine learning model, we also used AutoML approaches. They are based on the preparation of many, automated models providing an ensemble of the best predictors. PyCaret [26] is an open-source, low-code machine learning library that provides AutoML models, another one is MLJar [27]. Since we do not have many data points we do not focus on usage of neural networks, argumenting that they need many samples to provide reliable estimator [28]. Therefore best algorithms from AutoML packages were supposed to provide boosting, ensemble methods.

4.2 Modelling approach

The prepared modeling methodology results from both: (a) modeling goals - the task of predicting categorical target variable (multiclass classification problem) for next year and (b) type of data we are dealing with - short panel data (as to time dimension – just 3 available years). Accordingly, we developed the following evaluation strategy:

1. We select appropriate econometric and machine learning model architectures that are best suited to the given research problem (based on literature and expert knowledge).
2. We distinguish two pairs of training and evaluation (testing) dataset:
   - PAIR (1):
     - training dataset (year = 2020)
     - testing dataset (year = 2021)
   - PAIR (2):
     - training dataset (year = 2020 2021)
     - testing dataset (year = 2022)

   Please notice that this approach insures us against the risk of data leakage and potential overfitting in case of temporal dimension.
3. Per each model architecture we perform the following operations on the training set:

- preparation of variables specific to a particular model architecture and point in time
- model fitting
- model stability validation using cross-validation
- model hyperparameters tuning using cross-validation

The final product of step 2 is to generate the best model in its class for future prediction. For cross validation, we decided to use the Shuffle Split Cross Validation approach (the number of splits is 10), which allows us to introduce high randomization during the procedure. This choice seems to be a good trade-off between aware data leakage (training dataset (year = 2020 2021)) and potential overfitting within a single year.

4. Per each model architecture we perform following steps:
   - we create prediction on testing dataset with models trained in step 2
   - we score our testing dataset results with our novel scoring methodology
   - we perform model inspections for the impact of individual variables on the outcome using Explainable AI methods

   Please notice that we treat the test set as an out-of-sample data, to which we have no access until the final inference.

4.3 Performance criteria

The problem we are addressing is not a typical business problem - human lives may depend on the quality of our forecast, so the selection of evaluation metrics had to be done responsibly and meticulously. In our opinion, we should primarily maximize the number of observations that have been classified positively and any overclassing in the forecast will not be a big burden for us. Willing to apply our models to government recommendations, we prefer to make the mistake of overestimating the level of the burden target variable rather than underestimating it - an overestimation may result in the introduction of stronger regional policies but should not have as negative an effect on the people of that region (analogy to models indicating the spread of the COVID-19 pandemic). In contrast, underestimating the problem could lead to starvation and the death of millions of lives.

Taking this into account, we decided to approach the evaluation problem with a two-pronged approach:

- using common-known metrics:
  - Weigted F1-score,
  - Accuracy,
  - Weigted Precision Recall,
- deriving our own evaluation metrics:
  - driven underestimation punished accuracy (DUPA) – see Algorithm 4.3.1,
  - driven underestimation punished f-score (DUPF1) – see Algorithm 4.3.2.

4.3.1 Driven underestimation punished accuracy

Input:

- y_{true}: list or array of true values

- y_{pred}: list or array of predicted values
- *higher_pen*: float representing penalty weight for true values greater than predicted values (default 1.0)
- *lower_pen*: float representing penalty weight for true values less than predicted values (default 0.5)

Algorithm:

1. Calculate the number of y_{true} values less than y_{pred} and multiply by *lower_pen*.
2. Calculate the number of y_{true} values greater than y_{pred} and multiply by *higher_pen*.
3. Add the results of step 1 and step 2 together.
4. Multiply the sum from step 3 by -1.

4.3.2 Driven underestimation punished f-score

Input:

- y_{true}: list or array of true values
- y_{pred}: list or array of predicted values
- *beta*: float representing beta value for f-beta score calculation (default 2)
- *large_penalty*: float representing penalty multiplier for f-beta scores below threshold (default 0.5)

Algorithm:

1. Calculate the confusion matrix using y_{true} and y_{pred} and store in variable cm.
2. Calculate the true positives and store in variable tp.
3. Calculate the false positives and store in variable fp.
4. Calculate the false negatives and store in variable fn.
5. Calculate the precision values and store in variable precision, using the formula $tp/(tp + fp + 1e - 9)$.
6. Calculate the recall values and store in variable recall, using the formula $tp/(tp + fn + 1e - 9)$.
7. Calculate the f-beta scores and store in variable f_beta, using the formula $(1 + beta^2) * precision * recall/(beta^2 * precision + recall + 1e - 9)$.
8. For each element in f_beta, if it is less than $large_penalty$, multiply it by $large\_penalty$.
9. Calculate the weights for each class and store in variable w.
10. Calculate the weighted f-beta score and store in variable $weighted\_f\_beta$

5 Results

5.1 Models performance

Table 3: Results of the metrics for all the models trained in the study

	2021 testing						2022 testing				
Model	**DUPA**	**DUPF1**	**f1-score**	**accuracy**	**precision**	**recall**	**DUPA**	**DUPF1**	**f1-score**	**accuracy**	**precision**
Naive model - random class	-166	0.15	0.3	0.3	0.32	0.3	-178	0.12	0.26	0.24	0.3
Naive model - martignal (values from prev. Year)	-120.5	0.40	0.49	0.49	0.52	0.49	-119	0.46	0.53	0.52	0.57
Naive model - mode for country (values from pre. Years)	-160	0.25	0.24	0.32	0.2	0.32	-134.5	0.34	0.33	0.43	0.29
Naive model - based on classic approach with 1.6 ICF	-232	0.13	0.13	0.22	0.31	0.22	-106	0.19	0.13	0.29	0.08
Ordered probit	-130.5	0.31	0.4	0.4	0.41	0.4	-124.5	0.36	0.46	0.47	0.49
Ordered logit	-149	0.19	0.38	0.38	0.39	0.38	-124.5	0.36	0.47	0.47	0.5
Multinomial logit	-139.5	0.26	0.32	0.34	0.33	0.34	-113.5	0.40	0.42	0.44	0.47
Naive Bayes	-156	0.22	0.32	0.35	0.36	0.35	-166	0.25	0.34	0.35	0.46
KNN	-134	0.21	0.41	0.42	0.44	0.42	-125.5	0.33	0.48	0.48	0.5
SVM	-143	0.18	0.35	0.37	0.4	0.37	-130.5	0.29	0.42	0.42	0.44
Random Forest	-126.5	0.32	0.44	0.44	0.46	0.44	-96.5	0.51	0.59	0.59	0.62
CatBoost	-120.5	0.33	0.45	0.46	0.47	0.46	-118.5	0.33	0.45	0.44	0.5
NGBoost	-105	0.44	0.49	0.49	0.51	0.49	-99	0.59	0.6	0.6	0.63
ml-jar (AutoML - ensemble weighted model from CatBoost, XGBoost, Random Forest)	-109	**0.50**	0.52	0.52	0.54	0.52	-122.5	0.23	0.48	0.47	0.5
PyCaret (AutoML - LightGBM model)	**-101**	0.45	**0.53**	**0.53**	**0.56**	**0.53**	**-74.5**	**0.63**	**0.65**	**0.65**	**0.69**

Fig. 3: DUPA score for different models

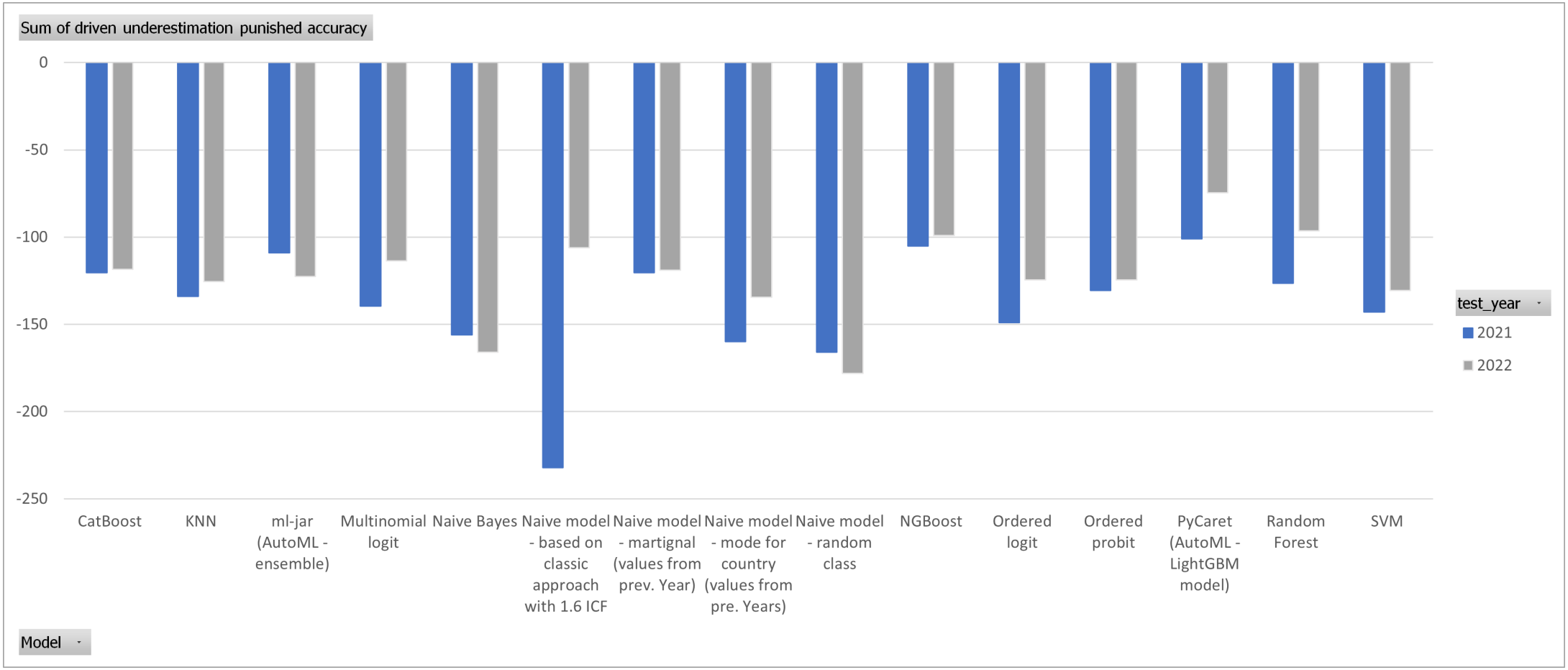


Fig. 4: F1 score for different models

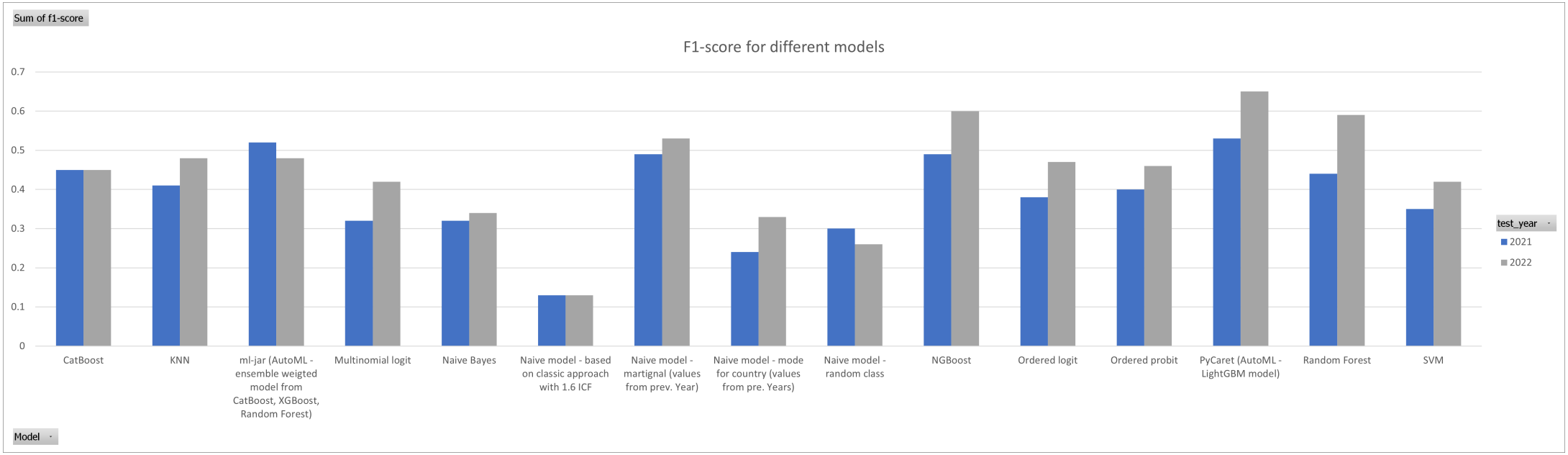


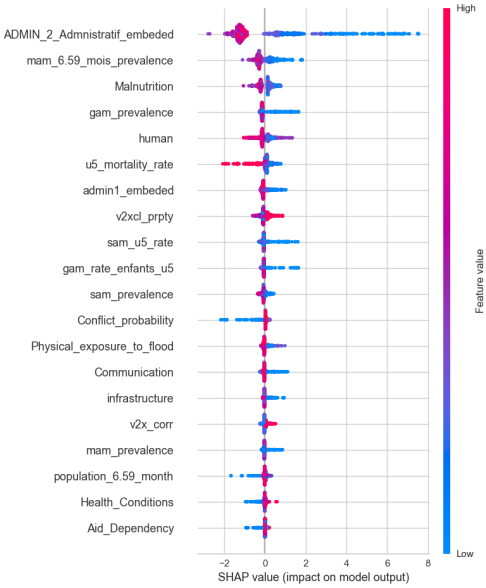
Table 3 and figures 3, 4 contain the results of our models on two out-of-sample test sets. We unequivocally state that the best results on both sets on almost all evaluation metrics were achieved by the LightGBM model (created with the AutoML PyCaret library) - on our metrics: DUPA/DUPF1/F1-score we get the following results: -101/0.45/0.53 (year 2021)) and -74.5/0.63/0.65 (year 2022). These results seem to be very satisfactory in the context of the performance of other architectures of statistical models and machine learning. It should be noted that we managed to outperform all naive benchmark models in this way (interestingly, the best of them is martignal (values from prev. Year), and the weakest is a fully random model. Moreover, it is worth noting that the class of models that best coped with this problem were tree models based on the idea of bagging (Random Forest) and boosting (CatBoost, NGBoost, LightGBM, XGBoost). In addition to PyCaret, we tried the ml-jar package, which selected the model ensemble weighted model from CatBoost, XGBoost, Random Forest as the best - however, as it turned out during the test, such greedy assembling led to strong overfitting and cannot be used in a practical problem.Finally, based on our empirical results, we conclude that the LightGBM model preceded by our proprietary preprocessing is the best solution to the research problem under consideration.

5.2 Model drivers and insights - XAI

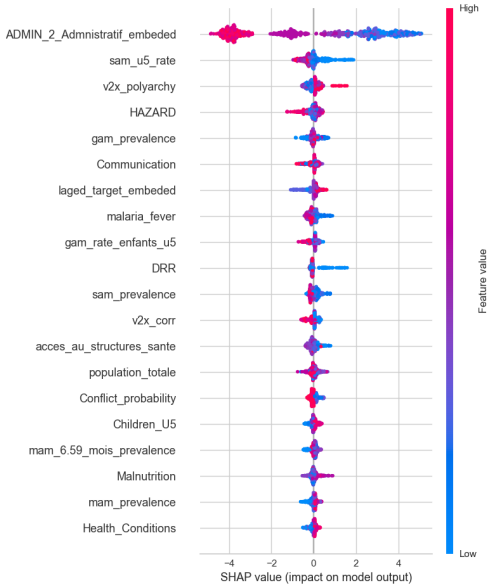
For an ordered choice problem, the explainable AI (XAI) methods are still underdeveloped. The publicly available XAI tools do not distinguish between ordered and unordered choice problems. All of the tools suggest performing the analysis with one-versus-rest approach, meaning that the choice problem is divided into n separate binary classification problems, where n is the number of label levels in the original problem. This means that the interpretation of XAI method results can be unintuitive and time consuming, especially when compared to the interpretation process in regression or binary classification problems. Nevertheless, for ordered choice problems XAI is still an excellent tool to derive meaningful insights from black box models of this paper, which is key for the shareholders, here policymakers. We verify Hypothesis 2 and Hypothesis 3, as well as derive insights about the prediction strategy of the best performing model, using the available XAI tools from the shap library in Python.

Figure 5 shows the magnitude of information to be gained from using the XAI methods combined with black box algorithms. The variables in the figures are ordered by having the variables with most significant impact on the final label value being at the top, and the ones with the least impact being at the bottom. When analyzing the plots, we find for example that one of the most significant predictors when considering the label = 0, which coincides with low burden of wasting levels, is v2xcl_prpty, which is an indicator of how much the right to private property is respected in the region. The figure shows that the better the right to private property is established in the nation, the higher the chance of burden of wasting being low, which is in line with what was expected. The figure for the predicted label = 2, meaning high burden of wasting, show that among the most significant drivers behind predicting this value of label are u5_mortality_rate meaning the mortality rate of children under 5 years old, as well are the Physical_infrastructure indicator. For the mortality rate of children under 5 years old, it seems that as the mortality rises, so does the burden of wasting.

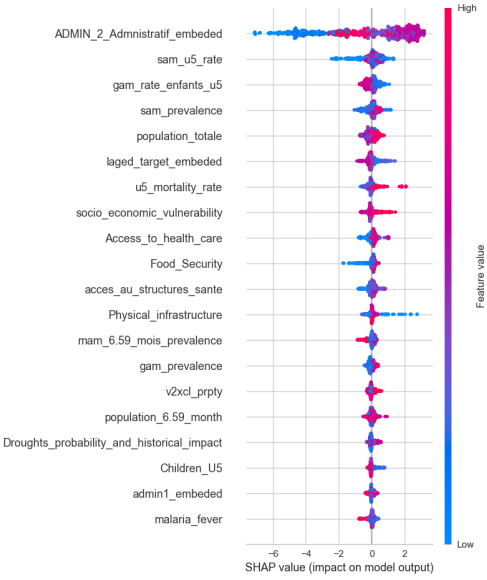
Fig. 5: SHAP values per target category



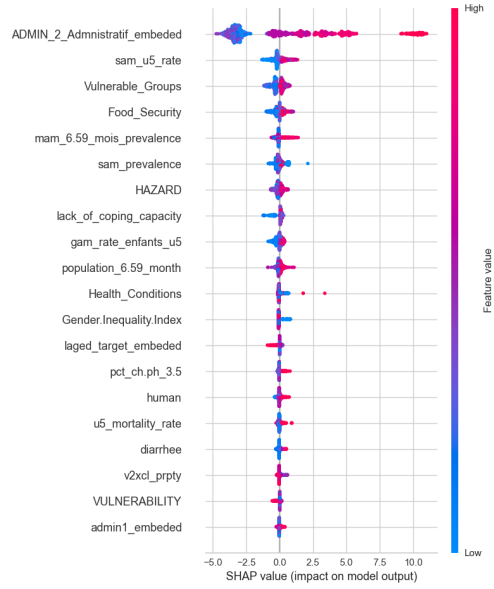
(a) target = 0



(b) target = 1

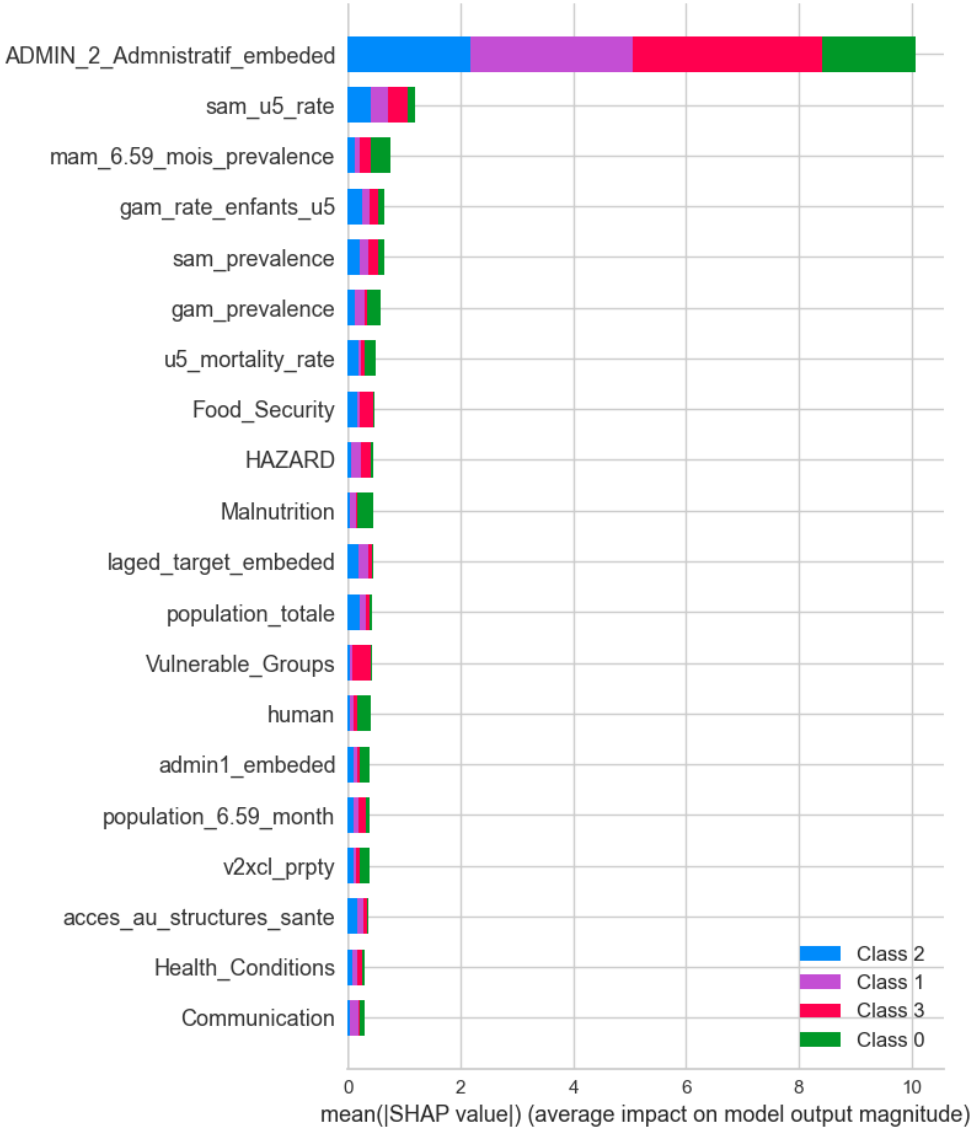


(c) target = 2



(d) target = 3

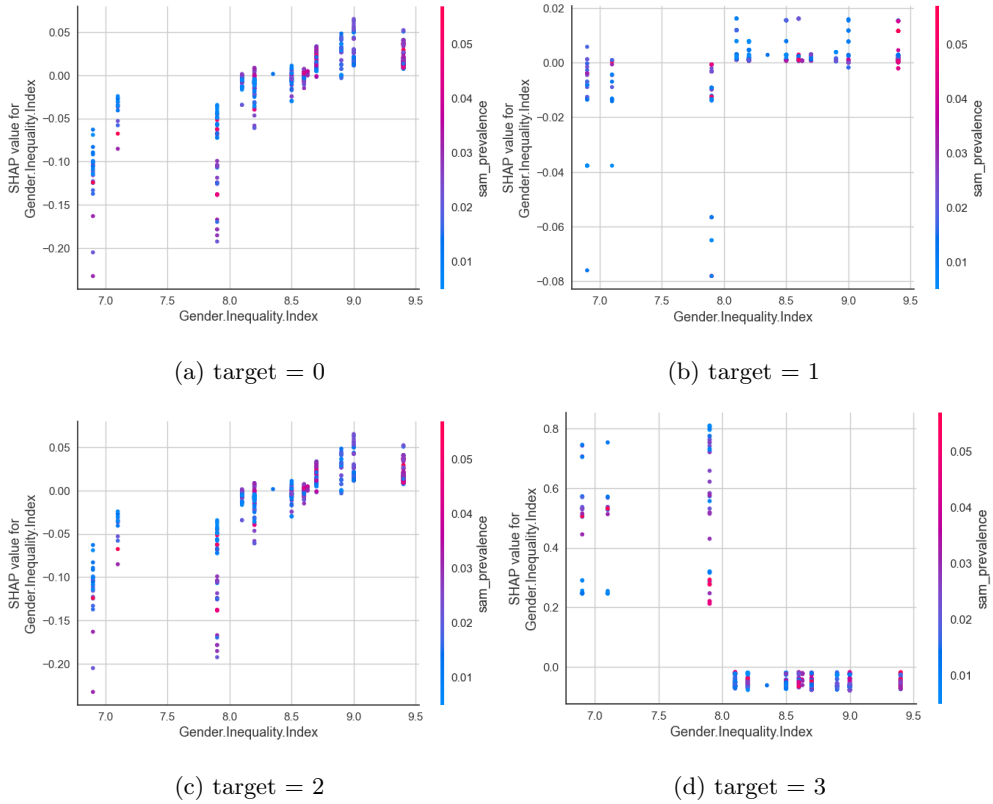
Fig. 6: SHAP values for best model without distinction



With the physical infrastructure the relationship is opposite. Several democratization indicators e.g. v2x_polyarchy (the extent to which electoral democracy applies to the country) and v2x_corr (political corruption index) also have a relationship that is implied by the model in line with Hypothesis 3.

The combined SHAP value plot suggests two conclusions about the performance and predictive strategy of our best performing model. For one, the most variable

Fig. 7: Partial Dependence Plots for variable GII values in distinction for target values



(a) target = 0

(b) target = 1

(c) target = 2

(d) target = 3

with the highest impact on the overall input is the embedded value of the admin 2 level name of the region. This suggests that this variable inherited the explanatory power of some of the variables not included in the model that are time invariant at least in the investigated period, such as the amount of arable land. This results in the variable containing the explanatory power of the fixed effects of the regions. The second important takeaway is that the variables most impacting the predicted burden of wasting are related to the prevalence of wasting in the preceding period, for example sam_u5_rate, sam_prevalence and gam_prevalence. This is in line with both intuition as the changes in quality of life are expected to change gradually over the years, and with the current scientific consensus, as the currently most popular burden model is based on the prevalence of the disease from the preceding period.

In order to focus on one or two given predictors, which is most likely the most useful for the policymakers, a partial dependence plot can be constructed. The partial dependence plot shows more precisely how the level of a given predictor affects the final prediction. In line with the shap plots, the partial dependence profiles have to be plotted with the one-versus-rest approach.

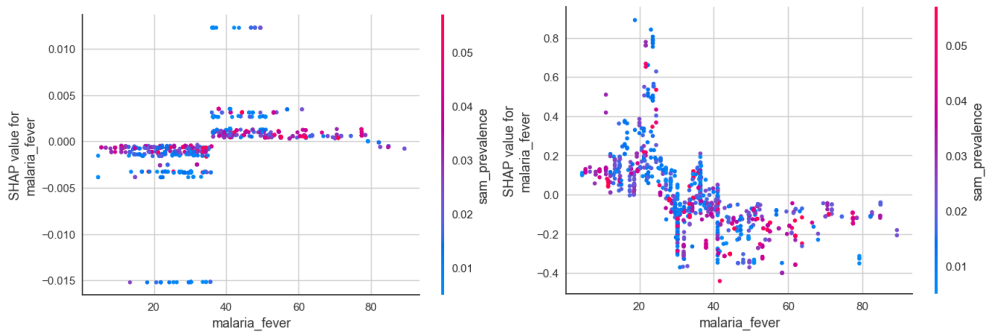
Figure 7 shows the partial dependence profiles of Gender.Inequality.Index (GII), being our stand in for female employment in this research, as the female employment data was not available for the scope of our research. The general idea underlying the construction of PD profiles is to show how does the expected value of model prediction behave as a function of a selected explanatory variable. This means that the if the slope of the best line in the plot would be positive, the variable has a positive relationship with the label. With GII being the highest in regions with the most unequal treatment based on gender, the results seem not to be intuitive. The model interprets the high values of GII with low levels of burden of wasting for label values corresponding to low and medium burden. For the value of label corresponding to high burden, the relationship between GII and the probability of predicting high burden is positive, in line with Hypothesis 2, however for the values corresponding to very high burden the relationship flips to negative. This means that the implied relationship between GII and burden of wasting is U-shaped, which is not in line with the scientific consensus. The difference between the scientific consensus about the relationship and the results of the model suggests that the situation in Sahel due to social and cultural factors is far different from the rest of the world, and should be investigated further. It is likely that due to limited scope of data the model used for the implied relationship captures the variance between different countries, rather than between a given country over the years. GII is not a variable that changes a lot over time, so if the data was sufficient to perform a robust panel analysis, it is likely that within-country variations in GII should be significant enough to derive meaningful insights for the policy makers.

Figure 8 shows the impact of rate of deaths associated with malaria fever (malaria_fever) on the burden of wasting. Here, the scientific consensus is that the relationship is significant and positive, which is in line with what the partial dependence plot show. For the plot corresponding to very high burden, the impact of malaria_fever is clearly positive, as in the higher the value of malaria_fever the higher the probability of predicting the burden as very high.

6 Discussion and recommendations

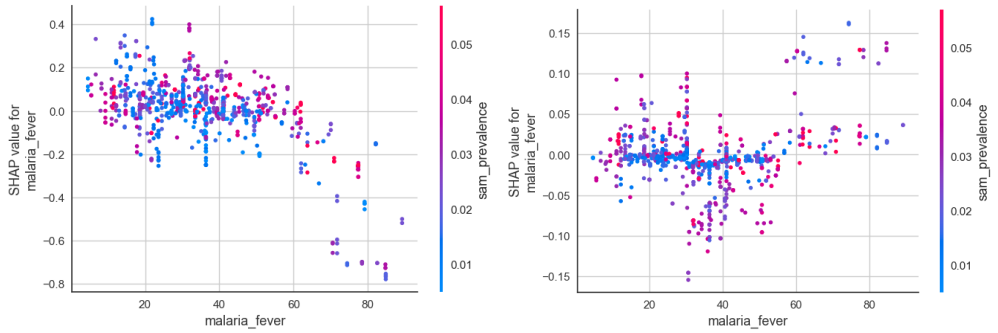
This paper addressed the issue of predicting the burden of wasting in the Sahel region, based on the data obtained between 2019 and 2021. The analysis performed show that the method of calculating burden that is prevailing in contemporary literature is lacking in predictive power when compared to more modern, machine learning based approaches, which is in line with our first hypothesis. We have also proposed a framework for developing models and two new model performance evaluation metrics, fit for the problem of ordered choice modeling. One of the limitations of the machine learning based approaches used to be the leak of model interpretability, however basing on the recent developments in the field of explainable AI, we have shown that even the most complex models can be used to derive meaningful and intuitive insights, that may aid the policymakers. We argue that due to leveraging the developments in data science, we don't have enough evidence to reject our main hypothesis. In that vein, we have performed an analysis of impact of gender inequality index on the predicted burden of wasting, and have concluded that in order to achieve robust results of such

Fig. 8: Partial Dependence Plots for variable malaria_fever values in distinction for target values



(a) target = 0

(b) target = 1



(c) target = 2

(d) target = 3

an analysis long-term data is required, as the relationship implied by our model is opposite to the scientific consensus. This is also opposite to what we assumed in our second hypothesis, hence we have to reject it, but we underline the significance of further research on this topic. In order to test our third hypothesis, we have analyzed the impact of democratization and quality of governance indicators on the prediction, concluding that in line with contemporary literature, the more democratic a country is, the lower the predicted burden of wasting. Thus, we don't have enough evidence to reject our third hypothesis.

A major limitation of our analysis is that we are trying to model dynamic effects, for example the impact of GII on burden of wasting, without any reliable source of panel or time-series data. Ideally, we would have representative survey results over time following the information on the regions over the long term so that we could properly assess the predictors of burden of wasting over time with a panel dataset. While we have the data regarding the years 2020-2022, not only is this period too short to observe the impact of overarching social changes, but also includes several economic shocks that affect the malnutrition levels in Sahel, such as the Covid-19

pandemic and the war in Ukraine - one of the main grain exporters to the northern parts of Africa. Furthermore, the data is of rather poor quality, containing multiple missing values and inconsistencies, which is often the case regarding data obtained in less developed countries.

Another limiting factor in this analysis is that the burden of wasting is presented as an ordered choice problem, when it can be more robustly analyzed as a regression problem. The available methods for regression problems are more developed on each stage of model development, as the problem is much more often occurring. The impact of the ordered choice methods being underdeveloped compared to those of regression can be most felt at the model construction and the XAI analysis steps of the modeling process. Therefore, it may be beneficial to perform a similar analysis predicting a continuous variable in order to provide better insights about drivers of burden of wasting.

Further study of the burden of wasting drivers is crucial to accurately assessing the welfare impacts of many public policies. This research suggests that policies propagating democracy and limiting corruption may be useful as tools of reducing the burden of wasting. Furthermore, the study shows the need for additional research of this data, and rough direction that additional research may proceed in on this topic.

References

[1] Isanaka, S., Andersen, C.T., Cousens, S., Myatt, M., Briend, A., Krasevec, J., Hayashi, C., Mayberry, A., Mwirigi, L., Guerrero, S.: Improving estimates of the burden of severe wasting: analysis of secondary prevalence and incidence data from 352 sites. BMJ Global Health **6**(3), 004342 (2021) https://doi.org/10.1136/bmjgh-2020-004342 . Accessed 2023-04-19

[2] Dale, N.M., Myatt, M., Prudhon, C., Briend, A.: Using cross-sectional surveys to estimate the number of severely malnourished children needing to be enrolled in specific treatment programmes. Public Health Nutrition **20**(8), 1362–1366 (2017) https://doi.org/10.1017/S1368980016003578 . Publisher: Cambridge University Press. Accessed 2023-04-19

[3] Bulti, A., Briend, A., Dale, N.M., De Wagt, A., Chiwile, F., Chitekwe, S., Isokpunwu, C., Myatt, M.: Improving estimates of the burden of severe acute malnutrition and predictions of caseload for programs treating severe acute malnutrition: experiences from Nigeria. Archives of Public Health = Archives Belges De Sante Publique **75**, 66 (2017) https://doi.org/10.1186/s13690-017-0234-4

[4] Kodish, S., Allen, B., Salou, H., Schwendler, T., Isanaka, S.: Conceptualizing factors impacting nutrition services coverage of treatment for acute malnutrition in children – an application of the Three Delays Model in Niger. Public Health Nutrition, 1–21 (2021) https://doi.org/10.1017/S1368980021004286

[5] Isanaka, S., Hedt-Gauthier, B.L., Salou, H., Berthé, F., Grais, R.F., Allen, B.G.S.: Active and adaptive case finding to estimate therapeutic program coverage

for severe acute malnutrition: a capture-recapture study. BMC Health Services Research **19**, 967 (2019) https://doi.org/10.1186/s12913-019-4791-9 . Accessed 2023-04-19

[6] Choudhury, K.K., Hanifi, M.A., Rasheed, S., Bhuiya, A.: Gender Inequality and Severe Malnutrition among Children in a Remote Rural Area of Bangladesh. Journal of Health, Population and Nutrition **18**(3), 123–130 (2000). Publisher: icddr,b. Accessed 2023-04-19

[7] Batis, C., Mazariegos, M., Martorell, R., Gil, A., Rivera, J.A.: Malnutrition in all its forms by wealth, education and ethnicity in Latin America: who are more affected? Public Health Nutrition **23**(S1), 1–12 (2020) https://doi.org/10.1017/S136898001900466X . Publisher: Cambridge University Press. Accessed 2023-04-19

[8] McDonald, C.M., Manji, K.P., Kupka, R., Bellinger, D.C., Spiegelman, D., Kisenge, R., Msamanga, G., Fawzi, W.W., Duggan, C.P.: Stunting and Wasting Are Associated with Poorer Psychomotor and Mental Development in HIV-Exposed Tanzanian Infants12. The Journal of Nutrition **143**(2), 204–214 (2013) https://doi.org/10.3945/jn.112.168682 . Accessed 2023-04-19

[9] Katona, P., Katona-Apte, J.: The Interaction between Nutrition and Infection. Clinical Infectious Diseases **46**(10), 1582–1588 (2008) https://doi.org/10.1086/587658 . Accessed 2023-04-19

[10] Asoba, G.N., Sumbele, I.U.N., Anchang-Kimbi, J.K., Metuge, S., Teh, R.N.: Influence of infant feeding practices on the occurrence of malnutrition, malaria and anaemia in children 5 years in the Mount Cameroon area: A cross sectional study. PLoS ONE **14**(7), 0219386 (2019) https://doi.org/10.1371/journal.pone.0219386 . Accessed 2023-04-19

[11] Victora, C.G., Adair, L., Fall, C., Hallal, P.C., Martorell, R., Richter, L., Sachdev, H.S., Maternal and Child Undernutrition Study Group: Maternal and child undernutrition: consequences for adult health and human capital. Lancet (London, England) **371**(9609), 340–357 (2008) https://doi.org/10.1016/S0140-6736(07)61692-4

[12] Halleröd, B., Rothstein, B., Daoud, A., Nandy, S.: Bad Governance and Poor Children: A Comparative Analysis of Government Efficiency and Severe Child Deprivation in 68 Low- and Middle-income Countries. World Development **48**, 19–31 (2013) https://doi.org/10.1016/j.worlddev.2013.03.007 . Accessed 2023-04-20

[13] DRMKC - Disaster Risk Management Knowledge Centre (last): INFORM subnational model of Sahel. https://drmkc.jrc.ec.europa.eu/inform-index/INFORM-Subnational-Risk/Sahel/moduleId/1798/id/383/controller/Admin/action/Results Accessed 2023-04-19

[14] The V-Dem Dataset – V-Dem. https://v-dem.net/data/the-v-dem-dataset/ Accessed 2023-04-20

[15] Kursa, M.B., Rudnicki, W.R.: Feature Selection with the Boruta Package. Journal of Statistical Software **36**, 1–13 (2010) https://doi.org/10.18637/jss.v036.i11 . Accessed 2023-04-20

[16] Das, S., Rahman, R.M.: Application of ordinal logistic regression analysis in determining risk factors of child malnutrition in Bangladesh. Nutrition Journal **10**(1), 124 (2011) https://doi.org/10.1186/1475-2891-10-124 . Accessed 2023-04-20

[17] Sarkar, S.: Cross-sectional study of child malnutrition and associated risk factors among children aged under five in West Bengal, India. International Journal of Population Studies **2** (2016) https://doi.org/10.18063/IJPS.2016.01.003

[18] Browne, C., Matteson, D.S., McBride, L., Hu, L., Liu, Y., Sun, Y., Wen, J., Barrett, C.B.: Multivariate random forest prediction of poverty and malnutrition prevalence. PLOS ONE **16**(9), 0255519 (2021) https://doi.org/10.1371/journal.pone.0255519 . Publisher: Public Library of Science. Accessed 2023-04-20

[19] Alam, M.Z., Rahman, M.S., Rahman, M.S.: A Random Forest based predictor for medical data classification using feature ranking. Informatics in Medicine Unlocked **15**, 100180 (2019) https://doi.org/10.1016/j.imu.2019.100180 . Accessed 2023-04-20

[20] Pauly, O.: Random Forests for Medical Applications. PhD thesis, Technische Universität München (2012). https://mediatum.ub.tum.de/1094727 Accessed 2023-04-20

[21] Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3), 273–297 (1995) https://doi.org/10.1007/BF00994018 . Accessed 2023-04-20

[22] Talukder, A., Ahammed, B.: Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. Nutrition **78**, 110861 (2020) https://doi.org/10.1016/j.nut.2020.110861 . Accessed 2023-04-20

[23] Islam, M.M., Rahman, M.J., Islam, M.M., Roy, D.C., Ahmed, N.A.M.F., Hussain, S., Amanullah, M., Abedin, M.M., Maniruzzaman, M.: Application of machine learning based algorithm for prediction of malnutrition among women in Bangladesh. International Journal of Cognitive Computing in Engineering **3**, 46–57 (2022) https://doi.org/10.1016/j.ijcce.2022.02.002 . Accessed 2023-04-20

[24] Duan, T., Avati, A., Ding, D.Y., Thai, K.K., Basu, S., Ng, A.Y., Schuler, A.: NGBoost: Natural Gradient Boosting for Probabilistic Prediction. arXiv. arXiv:1910.03225 [cs, stat] (2020). https://doi.org/10.48550/arXiv.1910.03225 . http://arxiv.org/abs/1910.03225 Accessed 2023-04-20

[25] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: CatBoost: unbiased boosting with categorical features. arXiv. arXiv:1706.09516 [cs] (2019). https://doi.org/10.48550/arXiv.1706.09516 . http://arxiv.org/abs/1706.09516 Accessed 2023-04-20

[26] Ali, M.: PyCaret: An Open Source, Low-code Machine Learning Library in Python. (2020). PyCaret version 1.0.0. https://www.pycaret.org

[27] Płońska, A., Płoński, P.: MLJAR: State-of-the-art Automated Machine Learning Framework for Tabular Data. Version 0.10.3. MLJAR Sp. z o.o., Łapy, Poland (2021). https://github.com/mljar/mljar-supervised

[28] Goodfellow, I.J., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge, MA, USA (2016). http://www.deeplearningbook.org