Predictive modelling techniques to estimate the burden of acute malnutrition and severity mapping in the absence of anthropometric data

Use of machine learning

Why was the project needed?

- In South Sudan, anthropometric data is regularly collected in bi-annual, nation-wide cross-sectional surveys under the Food Security and Nutrition Monitoring System (FSNMS) in order to estimate the burden of acute malnutrition (caseload) and severity mapping for the coming year.
- Due to the COVID-19 pandemic, relevant global guidance suspended the collection of this anthropometric data in the most recent round of FSNMS due to the concern of spreading the virus with equipment such as height boards and weighting scales.
- In 2020, the Nutrition Cluster was faced with the challenge of no anthropometric data for estimating caseloads and planning in 2021.

The **South Sudan Nutrition Cluster**, with support from **REACH South Sudan, UNICEF and WFP** through the **Nutrition Information Working Group**, were involved in this project.

What was the purpose/scope of the project?

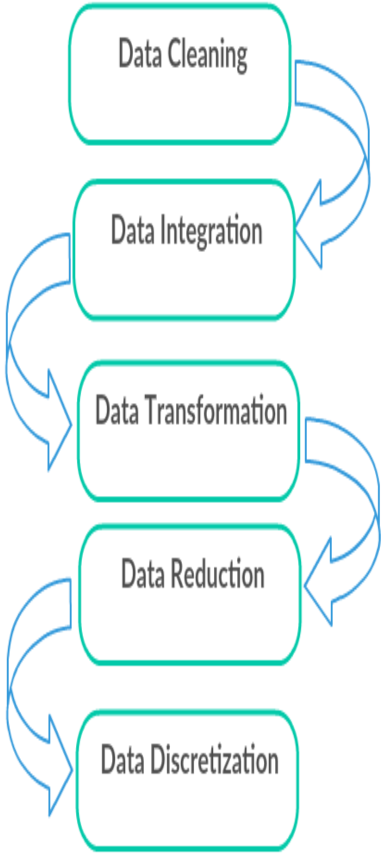
To address the problem, the Nutrition Cluster (NC) used predictive modelling techniques to estimate the burden of acute malnutrition and severity mapping in the absence of anthropometric data. The primary purpose was:

- To estimate nutritional status of children in a current national household dataset that did not have anthropometric measurements.
- To aggregate these results up to the county level for proxy GAM and SAM estimates.
- To use the proxy GAM estimates to inform severity classification and burden calculations for the HNO 2021.

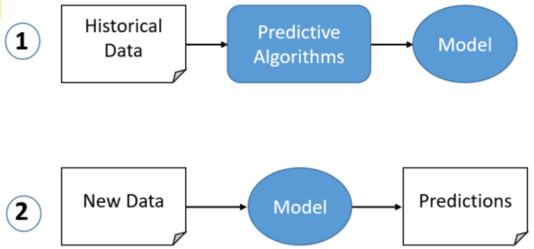
What was the data used for the project and its quality?

Training data: Common indicators of Health, WASH, FSL, Nutrition, Livelihoods and Shocks in FSNMS rounds 23, 24, and 25 were used and harmonized into a single, child-level dataset containing approximately 30,000 observations of children with anthropometric data.

Predicting data: Common indicators in FSNMS round 26, that were also in rounds 23, 24, and 25 were used for a single, child-level dataset without anthropometric data.



Methodology Overview



1

- Historical Data (FSNMS 23, 24, 25)
- Predictive Algorithms (regression, RF, SVM, QDA, LDA)
- Predictive Model (National Level)

2

- New Data (FSNMS 26)
- Predictive Model (National Level)
- Child-Level Predictions on Nutritional Status (classification)

3

- Predicted Child-Level Nutritional Status
- Aggregation of Predictions to County-Level
- Proxy GAM to feed into annual burden estimates/Severity mapping

Models and Key Outcomes Measured

Key Outcome Predicted:
- Classification of GAM and SAM by WHZ (1/0)

Models Tested:
- Logistic Regression
- Random Forest
- SVM
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis

Feature Selection - Several methods in sequence;
- removing high collinearity and low/zero variance variables;
- Backward and forward stepwise selection
- Lasso regression

Performance Metrics
- Accuracy, sensitivity, specificity

Results Evaluation:
- The final results of the **best performing model** were also reviewed by a group of in-country and global support nutrition analysts to provide a gut-check on the results within the context of South Sudan.
- County GAM and SAM prevalence that were deemed to be unrealistic by the group were discarded, and those observations were instead replaced using **hot-deck imputation techniques**

Key Results - Performance Metrics for Child Nutritional Status (WHZ)

	Algorithms				
	Logistic Regression	Random Forest	support vector machine (SVM)	Linear discriminant analysis (LDA)	Quadratic Discriminant Analysis (QDA)
Accuracy (95% CI)	0.6085 (0.5992, 0.6177)	0.862 (0.8553, 0.8684)	0.7332 (0.7247, 0.7415)	0.6077 (0.5984, 0.617)	0.5872 (0.5778, 0.5965)
Sensitivity	0.6239	0.8347	0.7385	0.6241	0.7768
Specificity	0.5930	0.8892	0.7279	0.5913	0.3976
Kappa	0.217	0.7239	0.4663	0.2155	0.1744
Detection Rate	0.3119	0.4173	0.3692	0.3120	0.3883

Selected County GAM Predictions

State (admin 1)	County (admin 2)	Predicted Proxy GAM - FSNMS Round 26	Round 23 Result	Round 24 Result	Round 25 Result
Warrap	Gogrial East	16.6%	14.9%	22.2%	16.4%
Warrap	Gogrial West	20.1%	14.9%	20.1%	16.4%
Warrap	Tonj East	15.6%	11.6%	9.5%	11.0%
Warrap	Tonj North	15.3%	9.5%	9.5%	11.1%
Warrap	Tonj South	19.2%	9.5%	9.5%	11.1%
Warrap	Twic	22.2%	24.1%	22.2%	16.4%
Western Equatoria	Ezo	3.8%	2.1%	1.6%	3.9%
Western Equatoria	Ibba	1.3%	6.9%	6.4%	5.2%
Western Equatoria	Maridi	3.4%	5.5%	6.4%	5.2%
Western Equatoria	Mundri East	8.3%	5.7%	6.4%	5.2%
Western Equatoria	Mundri West	2.5%	5.7%	6.4%	5.2%
Western Equatoria	Mvolo	10.9%	8.6%	5.5%	4.2%
Western Equatoria	Nagero	2.8%	5.2%	13.1%	3.6%
Western Equatoria	Nzara	4.8%	2.1%	1.6%	3.9%
Western Equatoria	Tambura	2.6%	5.2%	13.1%	3.6%
Western Equatoria	Yambio	2.3%	2.1%	1.6%	3.9%

Key Strengths and Weaknesses

Strengths

- Machine learning can help understand complex, non-linear relationships causing malnutrition, particularly methods such as random forest and neural networks.
- Wide variety of multi-sectoral household and individual level indicators available for exploration.
- Large sample sizes from historical FSNMS.
- Objective, impartial method for proxy GAM estimate, compared to group consensus methods.

Weaknesses

- Team was limited in terms of time and workload.
- Feature selection methods was limited and needs refinement.
- In areas where household FSNMS data was poor quality, the predictions tended to be poor quality.
- Comparing results at county level, but usually not representative in FSNMS. May be better to use domain results for historical comparison.

Future Plans and Potential Applications

Potential Applications

- **Imputation of child nutritional status** when poor quality anthro data.
- **Prediction of malnutrition through remote household surveys** such as mVAM or other modalities.
- With further advances in explainable artificial intelligence (xai), **identification of causal pathways of malnutrition reflected** within the complex modelling.

Next Steps

- **Re-run analyses** with clean datasets, refined feature selection strategy, sub-national disaggregation.
- **Explore additional methods with random forest to improve performance**, such as xgboost, model pruning, etc.
- Explore performance of **neural network** models and **dimensionality reduction techniques (PCA)**.
- If possible, **test or validate models** on SMART or household survey data.

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI


THANK
YOU