FTL Global ML bootcamp 3 | Capstone project | Group 13

Malnutrition Risk Prediction ML Model

Literature Review | *Data Research* | *Technology Review*

Group members:

- Fatima Nasraldin
- Lina Ahmed
- Linda Adil
- Maha Abdalfedil
- Nada Ali

Contents:

Literature Review for Malnutrition Prediction Using Machine Learning and Time Series Data

1. Introduction:

Malnutrition remains a critical issue in developing regions like Africa, contributing to high child mortality and long-term health impacts. Effective prevention and intervention strategies require accurate forecasting of malnutrition risk. Our project aims to develop a machine learning model that predicts malnutrition trends over time using historical health and demographic data.

2. Organization:

This systematic review is organized thematically: work on the prediction of child malnutrition in forecasting with data sources; studies that apply machine learning algorithms; thematic grouping allows a side-to-side comparison to be made of methodologies and results from similar research areas. It also emphasizes the significance of access to clean water, food security, and socio-economic well-being in improving childhood nutritional status in .

3. Summary and Synthesis:

Paper 1:

Predictions of Child Malnutrition in Developing Countries

Key Findings: The paper discusses the integration of several data sources to incorporate health, nutrition, and socio-economic data for forecasting child malnutrition rates in developing countries. It comes up that combining diverse data sets greatly increases the accuracy of predictions on the children's malnutrition status. It was shown that with the use of a wide range of data sources, a more in-depth understanding of the determinants of child malnutrition is formulated, thereby leading to interventions that are much better targeted.

Methodology: The authors used a variety of statistical and econometric models to analyze and forecast trends in malnutrition. In simple terms, the methodology was based on data collection and the integration of different data sets from government and other non-governmental institutions, followed by implementing forecasting models that consider historical behavior and current data input as part of estimation.

Contribution to the Field: The current work is a large contribution to the field in showing the potential of multi-source data integration to forecast malnutrition. It lays the groundwork for future work on more advanced predictive models, such as those based on machine learning.

Paper 2:

Machine Learning Algorithms for Predicting Undernutrition Among Under-Five Children in Ethiopia

Key Findings: The paper investigates the application of machine learning algorithms in predicting undernutrition among children below the age of five in Ethiopia. It has been indicated that machine learning models, especially decision trees and the support vector machine, improve predictive accuracy for undernutrition compared to the ordinary statistical method. This study emphasizes possible machine learning, which correctly and efficiently points out children at risk of undernutrition compared to the conventional approach.

Methodology: Data used was from Ethiopian Demographic Health Survey (DHS) and varied machine learning methods encompassing decision trees, random forests, to support vector machines. The models were prepared and validated by using cross-validation techniques and compared according to the performance measurement using metrics like accuracy, precision, and recall.

Contribution to the Field: This makes a very valuable contribution, demonstrating the effectiveness of machine learning algorithms in public health with regard to child malnutrition. New vistas of research are opened with this application of AI and machine learning in global health challenges.

Comparison and Contrast: The need for advanced predictive methodologies to address the burden of child malnutrition is a common factor in both the papers. The first paper integrates various data sources for better prediction accuracy with traditional econometric modeling, and the second one adopts new machine learning algorithms to show higher accuracy in use. What is more, the one methodology integrates diversity of data sources with classical models while the other adopts machine learning techniques for the sake of prediction.

4. Conclusion:

The literature reviewed only testifies to the importance of accurate forecasting and prediction in addressing child malnutrition in developing countries. As evidenced through this first paper's focus on integrating disparate data sources and the second paper's use of machine learning algorithms, opportunities for advanced improvements in precision within the identification of at-risk children and intervention strategies are greatly promising. In this research, we leverage multi-source data integration and machine learning techniques for the development of a robust, scalable model of child malnutrition prediction. This study utilizes multi-source data integration in combination with machine learning techniques to develop a very robust, scalable, and accurate model for prediction in child malnutrition.

5. Proper Citations:

[1] *"Forecasting Child Malnutrition in Developing Countries: Evidence from Multiple Data Sources"* International Food Policy Research Institute, 2000. [URL].

[2] *"Machine Learning Algorithms for Predicting Undernutrition Among Under-Five Children in Ethiopia"* ResearchGate, 2021. [URL].

Data Research for Malnutrition Prediction Using Time Series Data from World Bank Group

1. Introduction:

This is the case in many emerging areas, particularly Africa where malnutrition remains one of the most important factors explaining child mortality and long-term health. This challenge necessitates well-grounded and precise predictive tools that can predict malnutrition patterns to help streamline targeted actions.

We have announced plans to ***build our own data set*** and conduct extensive research in a more holistic effort of developing such a model. It begins with widespread searches on the web to data collection from all possible sources: health, demographic and environmental data.

2. *Organization:*

The following is the same thematic organization of this data research on how to create a custom dataset as explained for our needs.

1- ***Web Scraping :*** Data extraction from across the web sources finding befitting datasets

2- ***Data Integration :*** Processing and combining data from different sources to form a single, meaning full datasets.

3- ***Ensuring data is accurate, consistent and formatted for analysis.***

3. Data description:

For this project, we will use data from the ***World Bank Group*** *(Health Nutrition and Population Statistics)* ***[Link]*** and the ***Food and Agriculture Organization (FAO)*** ***[Link]*** to build a comprehensive dataset for predicting malnutrition trends.

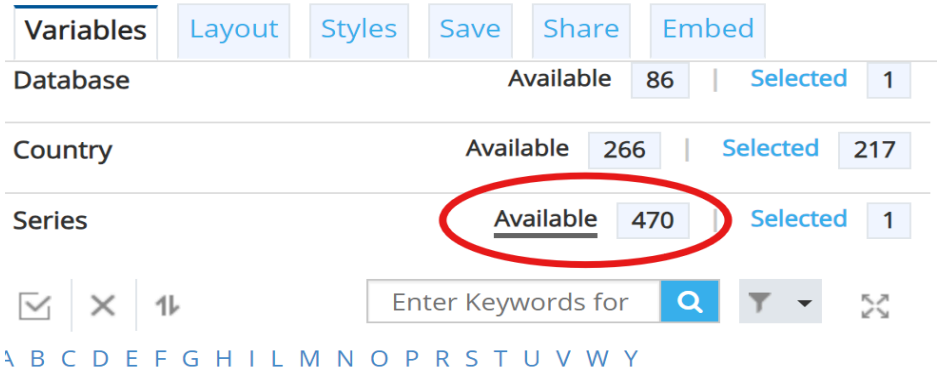
The World Bank Group provides *health and demographic data*, while the FAO offers insights into *agricultural production, food security, and environmental factors*.

Combining these sources will create a robust dataset capturing various factors affecting malnutrition. The next sections will detail the steps of extensive web research, data integration, and data cleaning and preprocessing to ensure the dataset is accurate and cohesive.

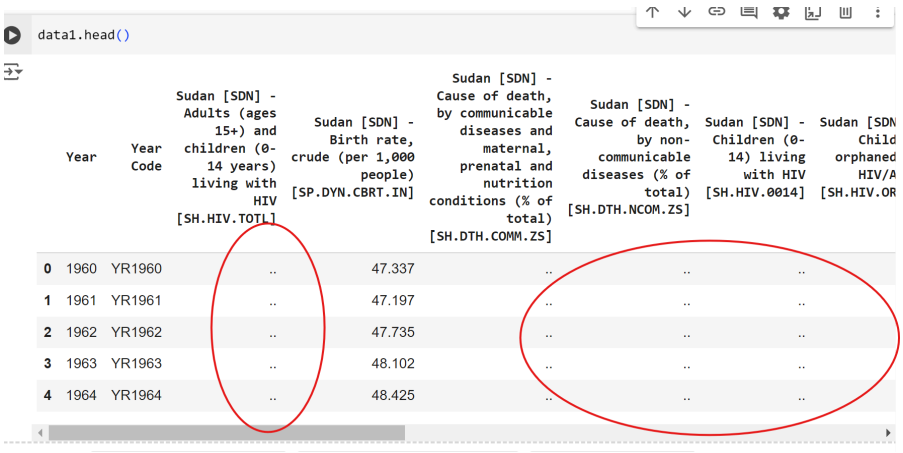
A. Extensive Web Research:

The primary step in the data collection is deep web research to find and collect related datasets. The World Bank Group gives a really big amount of information where one dataset ***has 470 features*** [Fig1] for each country (about 64 years) which goes empty specially in the first third with huge null values because measurements restarted in the 2000s in some diseases. [Fig2] Given the quantity and depth of this data,a thorough picking procedure is necessary.

We reviewed all available features thoroughly to identify what most affects malnutrition. Once we understood how good and relevant each feature was, we shortlisted around 30 key attributes (out of total 470) that provide better insights to predict malnutrition trends. By doing this we are able to keep our dataset manageable whilst also having a many True positives as possible.



"Fig1: The World Bank Group provides a vast dataset with 470 features for each country."



```
data1.head()
```

	Year	Year Code	Sudan [SDN] - Adults (ages 15+) and children (0-14 years) living with HIV [SH.HIV.TOTL]	Sudan [SDN] - Birth rate, crude (per 1,000 people) [SP.DYN.CBRT.IN]	Sudan [SDN] - Cause of death, by communicable diseases and maternal, prenatal and nutrition conditions (% of total) [SH.DTH.COMM.ZS]	Sudan [SDN] - Cause of death, by non-communicable diseases (% of total) [SH.DTH.NCOM.ZS]	Sudan [SDN] - Children (0-14) living with HIV [SH.HIV.0014]	Sudan [SDN Child orphaned HIV/A [SH.HIV.OR
0	1960	YR1960	..	47.337	..	..	..	
1	1961	YR1961	..	47.197	..	..	..	
2	1962	YR1962	..	47.735	..	..	..	
3	1963	YR1963	..	48.102	..	..	..	
4	1964	YR1964	..	48.425	..	..	..	

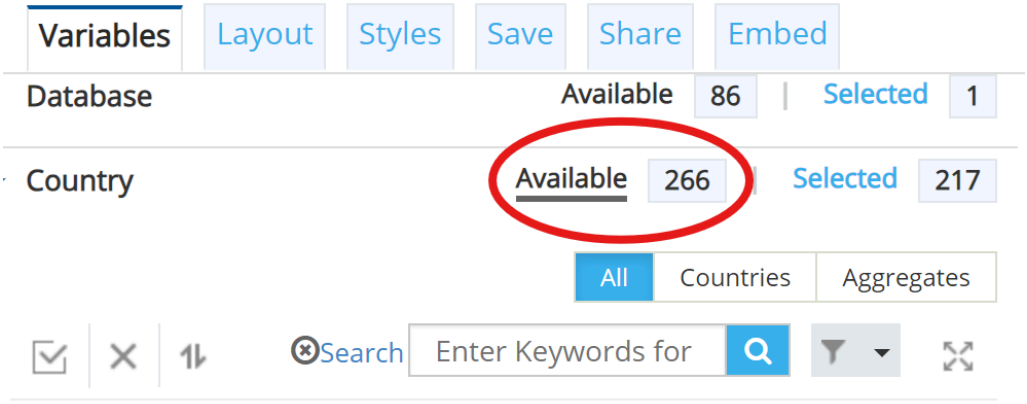
"Fig2: The dataset contains a significant number of null values, particularly in the first 30 years."

We performed equally comprehensive web searching for the FAO dataset, which contains over 50 datasets relating environmental factors to crop production. The columns already listed contain the broad scope of this data, so we carefully went over these and identified which would be advantageous for our use-case. We clipped the dataset to choose environmental and agricultural variables that potentially play an important role in driving malnutrition with environmentally associated

diseases. Having these targets in place we make sure that what we cover is from the right time period, and it also keeps our dataset relevant but most importantly manageable.

B. Data Integration:

The next step in our process is data integration, where we merge data from different Least Developed Countries (LDCs) using the selected datasets from both the World Bank Group and the FAO.



We first extracted and processed the data for each country individually. To create a cohesive dataset, we integrated these individual country datasets by merging them based on common variables, such as time period and relevant features. We added a column indicating the country of origin to distinguish between the different data sources. By repeating this process for each country, we expanded the dataset and increased the number of rows, resulting in a comprehensive dataset that captures the varied factors influencing malnutrition across multiple countries. This approach ensures that our integrated dataset is both detailed and representative of the broader regional context.

C. Data Cleaning and Preprocessing:

In the data cleaning and preprocessing phase, we focus on ensuring that the integrated dataset is accurate, consistent, and ready for analysis. This involves several key steps: first, we address missing values and inconsistencies by employing appropriate methods such as imputation or exclusion. Notably, the dataset contains a significant number of null values in the first 30 years due to the fact that measurements for certain diseases were not developed until the 2000s. Then We plan to standardize data formats and units to ensure uniformity across all features and identify and correct outliers and errors to prevent skewed analysis. Additionally, then we will normalize data where necessary to facilitate comparison and integration. By meticulously cleaning and preprocessing the data, we enhance its quality and reliability, thereby setting a solid foundation for accurate and effective predictive modeling.

Data creation step is done in a python notebook ***[Link]***

4. *Data Analysis and Insights:*

As we are still in the process of creating our dataset, we have not yet begun exploratory data analysis (EDA). However, based on our initial exploration and the dataset's structure, we anticipate several key patterns and trends:

- ***Descriptive Statistics:*** Indicators are expected to show consistent trends correlating with malnutrition rates.
- ***Seasonality and Trends:*** Time series analysis is anticipated to reveal seasonality in indicators like agricultural productivity, which is likely to align with variations in malnutrition rates.
- ***Correlations:*** We expect strong correlations between malnutrition levels and variables such as health expenditure and food access,

indicating that these factors are critical for predicting malnutrition trends.

- ***Data Gaps:*** We recognize that the dataset will include significant null values, particularly in the first 30 years, due to the lack of disease measurements prior to the 2000s. Additionally, some countries may have missing data for specific years, which will require imputation techniques or alternative methods to maintain model reliability.

  Once the dataset is fully integrated and cleaned, we will perform detailed EDA, utilizing visualizations such as line plots and heatmaps to illustrate trends and relationships. These insights will be crucial for guiding the feature engineering process and selecting the appropriate modeling techniques.

5. *Conclusion:*

In summary, the iterative and time-consuming process of creating and integrating our dataset is crucial for ensuring that it accurately represents the factors influencing malnutrition trends.

By carefully selecting relevant features from the World Bank Group and FAO datasets, and meticulously integrating and cleaning the data, we are building a robust dataset that meets our specific research needs. This effort is not only aimed at improving the accuracy of our predictive model but also at providing valuable insights into the broader issue of malnutrition.

Understanding the importance of each step in this process underscores our commitment to producing a high-quality dataset that will enhance the effectiveness of our model and contribute to meaningful interventions in the fight against malnutrition.

6. *Proper Citations:*

1- World Bank Group. (n.d.). Health, Nutrition, and Population Statistics. https://databank.worldbank.org/source/health-nutrition-and-population-statistics

2- Food and Agriculture Organization (FAO). (n.d.). FAOSTAT: Food and Agriculture Data. https://www.fao.org/faostat/en/#data

Technology Review for Malnutrition Prediction Using Machine Learning and Time Series Data

1. *Introduction:*

This technology review concentrates on machine learning methods and the tools needed to make predictions for malnutrition risk trends based on time series data. Given the complex and evolving nature of malnutrition determinants, selecting the right technologies is critical to ensuring accurate and timely predictions. This review aims to examine and assess tools and methods that could be used for the analysis of health and population data as part of a project aimed at developing predictive modeling in line with study objectives.

2. *Technology Overview:*

The tools reviewed, including machine learning algorithms and time series analysis techniques, are designed to predict malnutrition by handling complex, non-linear relationships and analyzing time-dependent trends.

The core technologies for this project include machine learning algorithms and time series analysis tools:

- ***Machine Learning Algorithms***: Random Forests, Gradient Boosting Machines (GBM), and CatBoost are the key models reviewed. These algorithms are widely used for classification and regression tasks and can handle complex relationships and non-linear data.
- ***Time Series Analysis Tools***: ARIMA (AutoRegressive Integrated Moving Average), Prophet, and LSTM Long Short-Term Memory (although not all chosen for this project) are popular for forecasting trends in time-dependent data. They provide techniques for handling seasonality, trend decomposition, and forecasting.
- ***Data Processing and Visualization Tools****:* Python libraries like pandas, scikit-learn, and matplotlib are used for data manipulation, model development, and result visualization. These tools are essential for

preprocessing time series data, feature engineering, and interpreting model outputs.

3. *Project Relevance :*

We spend time on these early steps, because we want to make sure that the resulting dataset meets our research objectives while at the same time lays a foundation for data about malnutrition in general.

We want to experiment with different models and see what the best model is for predicting undernutrition trends. Whilst we have not started testing the models, we thought that these sets of technologies would be a good bet to solve those problems for this project.

- ***Random Forests and Gradient Boosting*** offer robustness and interpretability, making them suitable for predicting malnutrition based on socio-economic and health factors.
- ***Time Series Tools*** are crucial for capturing temporal trends in malnutrition indicators and improving the accuracy of long-term forecasts.
- These technologies allow for efficient handling of structured time series data, enabling better risk stratification and intervention planning without the complexity and computational load of deep learning approaches.

4. *Comparison and Evaluation:*

- ***Random Forest vs. Gradient Boosting***: Both models are ensemble-based, but Gradient Boosting tends to provide better accuracy at the cost of longer training times. Random Forest is more interpretable and easier to tune.
- ***ARIMA vs. Prophet***: ARIMA is a traditional statistical model best suited for stationary time series data, while Prophet is more flexible and better at handling seasonality and missing data, making it a good fit for health-related forecasts.

- ***Ease of Use and Scalability***: Python libraries offer excellent scalability and integration with data science workflows, making them accessible and easy to deploy for similar projects.

5. ***Use Cases and Examples:***

- ***Case Study 1: UNICEF's Use of Random Forest for Nutrition Analysis****:* UNICEF leveraged Random Forest models to predict malnutrition levels in West Africa using socio-economic indicators, demonstrating the model's capacity to handle complex, non-linear relationships.
- ***Case Study 2: Time Series Forecasting for Health Data***: The World Health Organization (WHO) used ARIMA models to project trends in child mortality rates, highlighting the importance of temporal analysis in health interventions.

6. ***Gaps and Research Opportunities:***

One key limitation of the current machine learning models such as Random Forests is their inability to fully capture the dynamic relationships in evolving time series data. There is an opportunity to explore hybrid models that combine statistical approaches (e.g., ARIMA) with machine learning for better predictive performance. Additionally, the integration of external factors, such as climate conditions, remains an area for further exploration. Advanced feature engineering techniques could enhance the model's accuracy and applicability in real-world scenarios.

7. ***Conclusion:***

In summary, the combination of machine learning techniques and time series analysis tools provides a robust foundation for predicting malnutrition risks. The selected technologies balance accuracy, interpretability, and efficiency, making them well-suited for this project. By leveraging these tools, the project aims to deliver

actionable insights that can support effective health and nutrition interventions, particularly in regions most affected by malnutrition.

8. *Proper Citations:*

1- *"Predicting Micronutrients Using Neural Networks and Random Forest (Part 1)"* Towards Data Science, 2020. [URL]

2- "Time Series Forecasting for Healthcare Diagnosis and Prognostics with the Focus on Cardiovascular Diseases," ResearchGate, 2017. [URL]