Using Diet Analysis to Predict and Prevent Child Malnutrition

A Capstone Project Proposal by Team PySpark - HDSC Spring Cohort 2023

Malnutrition in children remains a critical global issue affecting millions of young lives, prompting concerted efforts by organizations like the UN and UNICEF. The significance of this problem lies not only in its prevalence but also in the challenge of early detection, particularly in children.

Project Aim

This project aims to leverage machine learning, data science, and data analysis to analyse diet patterns in children under 5 years old, in order to predict and thus subsequently prevent malnutrition.

Previous work

The research conducted by Team Seaborn of the HDSC 2023 Winter Cohort, has explored related areas using data from the Global Nutrition Report archives. They analysed the diet variables such as exclusive breastfeeding, early initiation, solid foods, etc. to predict the burdens of malnutrition such as stunting, wasting and overweight of children. They achieved this using Machine Learning and Deep Learning algorithms.

Our Contribution

Malnutrition can be caused by lack or excess of both macronutrients as well as the micronutrients. However, the contribution of micronutrients towards malnutrition was not considered in the work done during the previous cohort. We propose to build up on their research and study the effect of micronutrients along with the macronutrients on malnutrition and hence predict the burdens of malnutrition. We intend to predict two more burdens of malnutrition, namely; Severe wasting and underweight. Apart from this, we also intend to do time series prediction of the various indicators of malnutrition.

Datasets and Inputs

As specified in the problem statement which is to analyze the diet patterns in order to predict malnutrition, we will focus on diet and nutrition related features and download a customized dataset from the UNICEF data warehouse (**UNICEF Malnutrition Customized dataset**). This dataset contains new features compared to the dataset used in the previous cohort like different food groups, meal frequency, and features related to micronutrients intake.
We have also identified two additional datasets from the World Health Organization (WHO) database. This first dataset (**Global Database on Child Growth and Malnutrition**) contains data on indicators of Malnutrition such as Severe wasting, Wasting, Overweight, Stunting and Underweight. The data from the second dataset (**Vitamin and Mineral Nutrition Information System (VMNIS)**) can be used for assessing the prevalence of various vitamin and mineral deficiencies in populations. Other datasets that we are considering to use for our work are the datasets used by other researchers that are available on Kaggle (**Malnutrition across Globe dataset (Source: Kaggle)**) and Github (**Malnutrition dataset (Source: GitHub)**).

Since all these datasets contain Geographic area as a common feature, we will attempt to merge the data into a single dataset. If that is not possible then we will work on them independently and then use geographic areas to interpret the results.

Methodology

1. Data selection and preprocessing
   We will start by selecting datasets from the sources mentioned in the datasets and inputs section that cover a broader range of malnutrition indicators along with dietary information. After fixing structural errors and removing irrelevant observations, this data will be preprocessed to handle missing values, outliers and data normalization.
2. Feature Engineering and Selection
   In order to select the most relevant features, we will use feature importance techniques like L1 regularization (Lasso), L2 regularization (Ridge), Permutation Feature Importance, and Decision Tree as Feature Importance.
3. Machine Learning models
   Expanding the repertoire of machine learning models, we will include advanced regression models like ElasticNet, which combines L1 and L2 regularization to leverage the benefits of both Lasso and Ridge. Additionally, we will explore Gradient Boosting Machines (GBM), XGBoost, and LightGBM, which are powerful ensemble techniques for regression tasks. These models are known for their ability to handle complex interactions between features and provide accurate predictions. In addition to the deep learning regression model used previously, we will explore different neural network architectures like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) to extract more complex patterns from the data.
4. Time series forecasting visualized using dashboards
   Apart from the traditional machine learning and deep learning methods, we will also use time series forecasting to predict the future values of the indicators of malnutrition. This will be done subject to availability of yearly data in a particular geographic area. This will then be visualized using an interactive dashboard.
5. Comprehensive Diet Analysis and Nutritional Gap Identification
   To improve diet analysis, we will leverage advanced techniques like Non-negative Matrix Factorization (NMF) and topic modeling to extract meaningful information from dietary data. These methods will help identify nutritional gaps contributing to different forms of malnutrition. We will also explore the use of Natural Language Processing (NLP) to gain insights from textual data sources like dietary surveys and nutritional guidelines
6. Targeted Interventions
   Based on the outcomes of the predictive models and comprehensive diet analysis, we will propose targeted dietary interventions to prevent child malnutrition effectively

Conclusion

By enhancing the previous team's work to include undernutrition and micronutrient deficiencies and expanding the repertoire of machine learning models, this project aims to provide a more comprehensive understanding of child malnutrition. Time series forecasting, Targeted interventions and data visualization dashboards will make the research more user friendly.