DS5220 Predicting undernutrition among under-five children

Bezawit Ayalew
Hwijong Im
Anish Rao

Abstract

This study aims to predict the risk factors of undernutrition among under-five children in Ethiopia using machine learning algorithms. We extracted our dataset from the 2016 Ethiopian Demographic and Health Survey (DHS) child recode and individual recode datasets. To predict the risk factors, we decided to combine three indicators: stunting, wasting, and underweight, as the target variable to determine whether a child is malnourished. We evaluated several machine learning algorithms to predict the determinants of undernutrition and our final results show that the Naive Bayes model outperformed the other models in predicting the indicators. This study provides insight into potential risk factors for undernutrition in Ethiopia, which could inform potential interventions and future policy decisions to address this critical public health issue.

Introduction

Undernutrition is a critical public health challenge that affects millions of children worldwide. According to the latest data from the WHO about under-five children, it is estimated that around 150 million are stunted, 45 million are wasted and around 340 million are suffering from some kind of micronutrient deficiency.

Predicting the risk factors of undernutrition is essential in developing effective interventions to address this issue. Our study aims to use machine learning algorithms to predict the risk factors of undernutrition among children under the age of five in Ethiopia. Our study uses data from the 2016 Ethiopian Demographic and Health Survey (DHS). The DHS is a national survey conducted every five years that is designed to collect valuable information on a population's health and well-being in order to make further improvements to their quality of life. The Ethiopian DHS has multiple datasets about men, women, households, etc. For the data collection of each of these datasets, the DHS employed multi-stage stratification to stratify the data into 624 clusters. Since we only needed the samples of under-5 children, we extracted 9471 relevant samples from the 'Individual Recode' dataset. This dataset contains many features pertaining to health information such as height, weight, and age. It also has various features for socio-economic and demographic data, such as maternal education, wealth index, etc. In order to predict which features were the most significant determinants of undernutrition we employed four different machine learning algorithms: Naive Bayes, Random Forest, MLP, and SVM. We computed the label for our models by comparing the standard measures of the height-for-age, weight-for-height, and weight-for-age indicators provided by WHO with the measures of each child in Ethiopia provided by the child recode dataset. If all three indicators are a significant amount below the standard measure by WHO, then the label variable for that child would be binary encoded as 1, and 0 otherwise.

Background

Some background information regarding the data preprocessing and model selection techniques is needed to understand some of the methods used in this study.

Data Preprocessing

We needed to have prior knowledge of multiple preprocessing techniques to make the data set we're working with most consistent. Some of the techniques we used include:

Data Cleaning: Data cleaning involves dealing with missing or inconsistent data from the dataset. For example, in our dataset, all of the feature names were encoded so we used some data cleaning methods to decode the feature names

One-Hot-Encoding: This is a technique used to represent categorical variables as binary vectors. The size of the vector is determined by the number of unique categories. Each position in the vector corresponds to a possible value of the categorical variable. As many features in our dataset use categorical variables, this technique was highly beneficial.

SMOTE: The Synthetic Minority Oversampling Technique is a preprocessing technique that is used when dealing with the class imbalance in your dataset. SMOTE generates synthetic samples for the minority class by creating 'synthetic' examples that are interpolated between existing ones. This was an essential technique as our dataset was showing class imbalance issues.(Satpathy, 2020)

Model Selection

We evaluated and compared four different machine learning algorithms. In order to do this we required prior knowledge of these how these algorithms worked and concepts like precision, recall, etc. to evaluate the performance of these models. (Wakefield)

Naive Bayes: The Naive Bayes Classifier is a supervised learning algorithm that is based on Bayes' theorem and used for classification. It treats every value as independent and predicts its category using probability.

SVM: The Support Vector Machine algorithm is a supervised learning model that is used for classification and regression analysis. The algorithm first sorts the data into different categories by analyzing a group of training examples marked for one category or the other, it then builds a model for assigning new values to categories.

Random Forests: Random Forests are ensemble learning methods that combine different algorithms to achieve better results in classification and regression problems. Each classifier is individually is weak, but when combined with others, it can produce excellent results. The algorithm essentially starts with a decision tree and segments data into smaller sets based on specific variables.

Multi-layer Perceptron: Multi-layer perceptrons (MLPs) are a type of neural network that is inspired by the function of neurons. They are widely used for supervised machine learning problems such as classification and regression. MLPs are composed of multiple layers of nodes, each layer being fully connected to the next.

Additionally, concepts such as precision and recall are important metrics used to evaluate the performance of machine learning models in classification problems. By using these metrics, we are able to compare the performance of the four different models used in our study and identify the one that meets the requirements of the problem we're trying to solve.

Related Work

This project is based on a paper written by the Cambridge Press for Public Health Nutrition "Predicting Undernutrition in Ethiopia" (H Bitew 2021). The study draws on data from the Ethiopian Demographic and Health Survey. Uses Five ML algorithms including eXtreme gradient boost, k-nearest neighbors(k-NN), random forest, neural network, and generalized models to predict risk factors of undernutrition. The study uses geographic location to see the differences in undernutrition in different regions in Ethiopia.

Although our project is based on the Cambridge study, the model selection process is different from the study. Our project uses four different ML algorithms instead of five. The models our project uses are Naive Bayes, Random Forest, MLP, and SVM. In addition, the survey data used in our project is different as our features only include information from the Child and Individual data frames, whereas the Cambridge study includes features from the household data frame which contains specific features about the household of a child such as "Time to water source". In addition, the Cambridge study includes a breakdown of the relevancy of each feature by the three categories stunting, wasting, and underweight. Our project is different from categorizing all three categories as it focuses on combining the three variables as indicators of undernutrition.

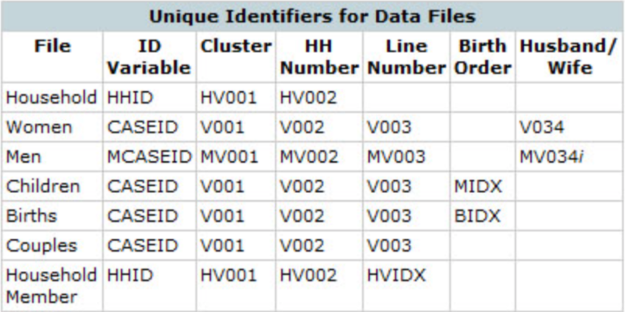
The methods that could have been applied to our project include using regional heat maps to further investigate the distribution of undernutrition amongst different regions in Ethiopia. In addition, including a household data frame in our dataset relates to our methods as it would have given us more features to compare, which might have improved the accuracy of the models, and given us further insight. The reason we did not include both of these in our method is due to the feasibility study we did in reaching our goals to complete the project. Taking into account how the household recodes file and geographical data were organized in the DHS dataset, the Household datasets were difficult to interpret and translate in a short turnaround. Our feasibility study included ensuring our exploratory data analysis, data preprocessing, and data cleaning were coherent, and integrating the household data frame and geographical data into the process may have hindered our quality assurance in the methods we used.

Project Description

Extracting dataset

From the 2016 DHS Ethiopian Survey category, we downloaded two DTA files, one being the Child Recode (used to extract and compute the label for our models) which included information about each child such as the height, weight, age, etc. Secondly, we downloaded an individual recode (used to extract features for our model) dataset which included 20 features that pertained to each child such as wealth index, maternal education, maternal underweight status, etc.

Each recode file contents were encoded using specific codes assigned by the DHS. In order to understand and interpret the contents of the data we used the DHS code handbook which included the interpretation of each variable that was in the code and dataset hierarchy table below(Unique Identifiers for Data Files) provided by DHS to derive meaning to each variable name in the datasets.



Unique Identifiers for Data Files						
File	ID Variable	Cluster	HH Number	Line Number	Birth Order	Husband/ Wife
Household	HHID	HV001	HV002			
Women	CASEID	V001	V002	V003		V034
Men	MCASEID	MV001	MV002	MV003		MV034*i*
Children	CASEID	V001	V002	V003	MIDX	
Births	CASEID	V001	V002	V003	BIDX	
Couples	CASEID	V001	V002	V003		
Household Member	HHID	HV001	HV002	HVIDX		

After interpreting the meaning of the unique identifiers for the data files, we imported both the child dataset and individual dataset into Python and converted the Stata files into a data frame using the p.read Stata file function.

Formula Computation For Label Variable

Starting with the Child Recode file, our first goal was to extract variables from the dataset to compute our label undernutrition.

According to WHO, this is how undernutrition is calculated: undernutrition indicators are determined by the following standard measures: stunting: height-for-age < –2 sd; wasting: weight-for-height < –2 sd and underweight: weight-for-age < –2 sd of the WHO Child Growth Standards median. Severe stunting, wasting, and underweight were those children whose height-for-age, weight-for-height, and weight-for-age *Z*-score was below minus 3 (−3) sd. This project, thus, considered all three
3 (−3) sd. This project, thus, considered all three undernutrition indicators to predict childhood undernutrition determinants (H Bitew 2021).

The Child recode file included height for age, weight for height, and weight for age two standard deviations away. We used the Child data frame to create three new columns, stunting: column calculated as the value in the ‘height_age_sd’ column less than -100, wasting: column calculated as the value in the ‘weight_height_sd’ column being less than -100, and underweight: column calculated as the value in the ‘weight_age_sd’ column being less than -100. After creating those three columns for stunting, wasting, and underweight, we divided the values by 10 to remove the implied decimal place. We then assigned new values to the target column of the Child data frame based on the three conditions: if the stunting column is less than -100, the wasting value is less than -100.

In this regard, the label/target was undernutrition: binary coded as 1 for stunted, wasted and underweight if the standard was met, else 0 for not stunted, not wasted and not underweight(H Bitew 2021).

Merging Data for Features

After computing the label the child data frame consisted of two columns one which was the “caseid” (a primary key to identify each child) and the target column coded as 1 or 0 to indicate the undernutrition of each child.

The next step was joining the Individual data frame which included all the necessary features with the Child data frame. We used a merge function (inner join on case_id) to combine the two data frames into one data frame.

Data Preprocessing

As mentioned above, since all column names in the individual data frame were stored in code and not the actual feature name, we implemented column rename functions to replace each variable code with the correct feature names to later help us identify and interpret the important features that determine and influence child undernutrition.

In the second step in pre-processing, we used data one-hot encoding to convert categorical values into numerical representations that can easily be understood by the algorithms. For example, for the feature “Region” we changed “Tigray” to be coded as 1 and “Afar” to be coded as 2, “Amhara” to be coded as 3, and “Oromia” to be coded. We used one-hot encoding on other variables in a similar way to enhance model performance, as it allows us capture non-linear relationships among categories, by representing categorical variables as binary vectors, one hot encoding provides a way for our models to learn the complex patterns and interactions between the variables in our dataset.

Since our dataset showed a class imbalance issue we decided to use the SMOTE preprocessing technique to oversample the minority class. We set the hyperparameter “random_state” = 22, as the seed for the random generator to control for the randomness of the oversampling process, which best fit our dataset and made it reproducible.

Since the dataset was collected from large surveys there was an importance in dealing with the noise of the dataset and ensuring the missing data was taken into account before building any type of model. For the missing data we used an inner join to get disclude

Model Building

Our first primary selection for a model was Random Forest. This is because of its feature selecting ability needed for our type of dataset and the goal of selecting the best features to predict undernutrition.

The data set was split, 80 % training set and 20% test set. Using the training set we built the first Random Forest Model. We used the confusion matrix to review the accuracy of the model. At first, the result we got below showed 0 cases for the number of False Negatives in the confusion matrix, after we encountered that issue we decided to use the sampled dataset using SMOTE to fix the class imbalance problem. After getting reasonable measurements in the confusion matrix we used the feature importance function available for Random Forest and built a data visualization that ranked the most important features at the top and the least important features at the bottom.

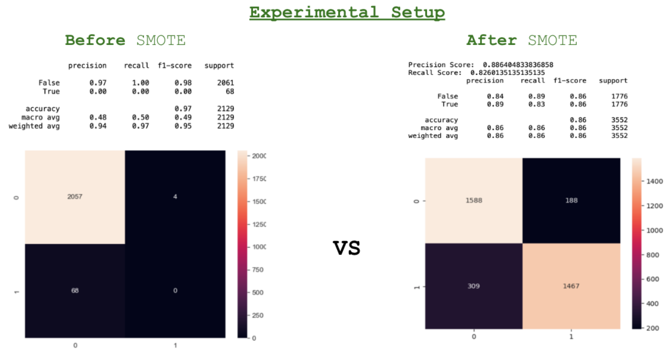
After building the Random Forest model, we then used the same set of tanning and test samples to build SVM, MLP, and Naive Bayes models. For each of the models we also built a confusion matrix to showcase the accuracy, precision, and recall of each model to compare which one performed the best for our specific data type.

Finally, we used the heat map function to further visualize our confusion matrix and we set the hyperparameters “annot” = True, to allow the confusion matrix values to be displayed on top of the heat map cells.

After reviewing our scores using the heat map, further hyperparameter tuning was conducted to maximize the accuracy of our models.

Empirical Results

The results from our accuracy test when we first ran the models showcased a class imbalance issue.



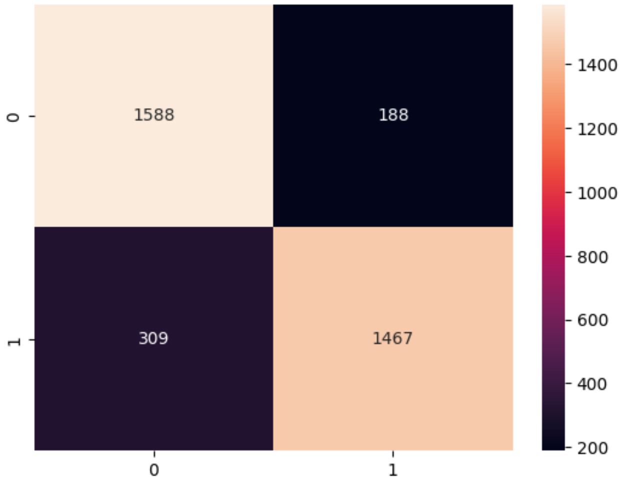
```
              precision    recall  f1-score   support

       False       0.97      1.00      0.98      2061
        True       0.00      0.00      0.00        68

    accuracy                           0.97      2129
   macro avg       0.48      0.50      0.49      2129
weighted avg       0.94      0.97      0.95      2129
```

```
Precision Score:  0.886404833836858
Recall Score:  0.8260135135135135
              precision    recall  f1-score   support

       False       0.84      0.89      0.86      1776
        True       0.89      0.83      0.86      1776

    accuracy                           0.86      3552
   macro avg       0.86      0.86      0.86      3552
weighted avg       0.86      0.86      0.86      3552
```

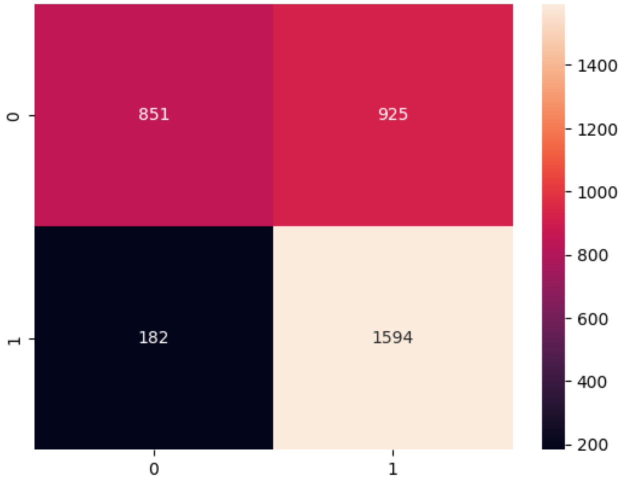
The initial results of our models before the SMOTE technique was applied showed low accuracy, a lack of true values of the target variable, we faced this low accuracy on all the models that we trained. The SMOTE technique improved the number of true values in the target variable. As we can see from the experimental setup figure above the total number of true negatives increased by 1467. Our results show that the technique enhanced our model sensitivity and reduced overfitting.

Classification Models

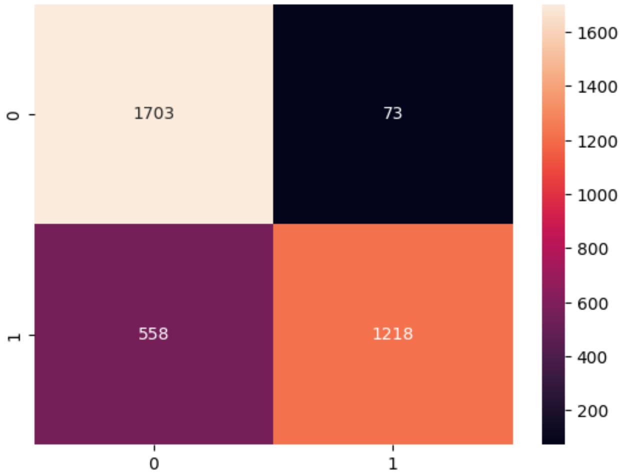
To classify children who are malnourished, we used models Random Forest, Support Vector Machine, Multi-Layer Perceptron, and Naive Bayes. The F-1 score of all models was intermediate, however, there were different strengths in the accuracy of each model. To be specific, one model had a high precision score other than other models, another model had a high recall score, and the other model had the highest F-1 score. The reason that we used Random Forest was to get the feature importances of our data set for our future direction of study, and with the perspective of the medical classification model, we thought that it would be important to classify the malnourished children without missing any cases. For this reason, we decided to choose which has the highest recall score, the Naive Bayes. The scores of the confusion matrix are as follows.



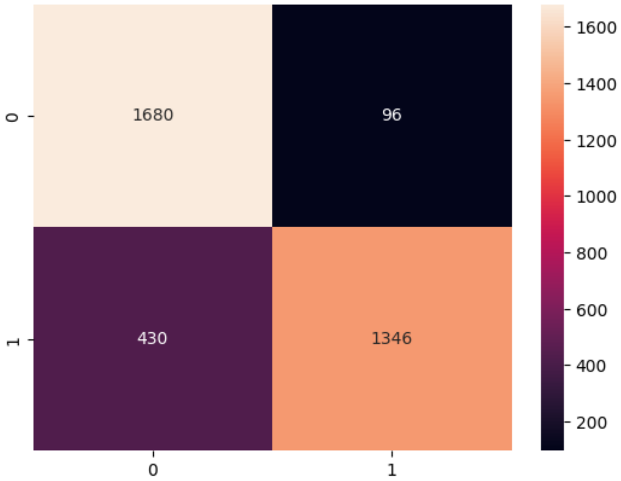
The Confusion Matrix of Radom Forest



The Confusion Matrix of Naive Bayes



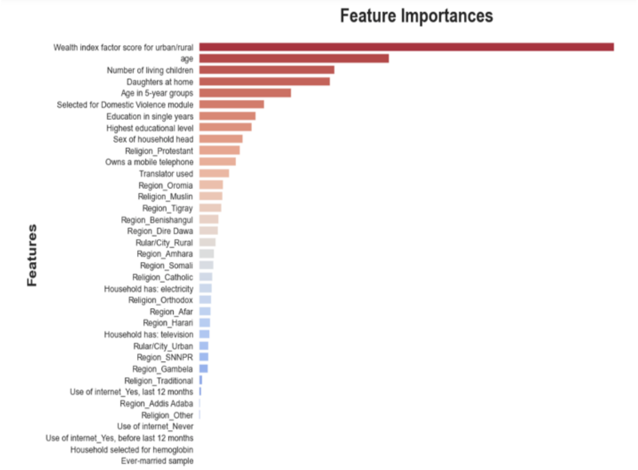
Support Vector Machine (SVM)



Multi-Layer Perceptron (MLP)

Feature Importance

We used our final data frame that included the target variable and features to build a feature importance graph using the Random Forest model. We can see from our results below that the top features that contribute to undernutrition are “wealth, age, number of living children, and daughters at home”. The features with less significance compared to the other features are “Ever married, household selected for hemoglobin and region Addis Ababa”. From these results obtained, we can deduce our model is performing well, as the Addis Ababa region is the capital of Ethiopia, and it's safe to assume that if the child can afford to live in the city region they may have better access to resources and therefore are less likely to be categorized as malnutrition. In addition, wealth being the highest importance shows our classifier is performing well since the likelihood of a child coming from a wealthy family being malnourished is low.



Feature importances of malnutrition children

Algorithm Comparison

The overall confusion matrix of the accuracy of models was high on average, however, it is essential to classify the children who are malnutrition and classify missing without any cases. Therefore, we concluded to use the model, which has the highest recall score, Naive Bayes. While we discovered and conclude the model we trained, it was a meaningful challenge for us to manage all the journeys we stepped. Even though we spend lots of time completing the data set, we learned how to break through and make the Target feature from the raw data set. Also, how to solve the imbalance data problem. Lastly, We learned how to decide on the models we designed and select appropriate models according to the circumstances.

Conclusion

In conclusion, our project aimed to predict the determinants of undernutrition among under 5 children in Ethiopia using a combination of socio-economic, demographic, and health-related features. While all the models we tested, including MLP, Random Forest, and SVM, performed well, the **Naive Bayes** model outperformed them all. This was primarily due to its better recall rate, which is essential for our data type since it's more important to identify a child that is malnutrition and not miss any cases**.**

Future Direction

Through our analysis, we extracted many important features that could be potential predictors of malnutrition. While some of these features were expected, such as 'Wealth score,' we also discovered many unexpected features, such as ‘religion_protestant’ and ‘Daughters at home’. This kind of information could be valuable for policymakers in identifying at-risk populations and developing appropriate interventions.

In the future direction, we are going to discover more unexpected features from the feature importance in Random Forest. This is because features, which we cannot predict and give rise to the number of malnutrition children, might help the policymakers who want to prevent malnutrition children as pragmatic information.

To future DS5220 students, we want to advise that before starting the project, students have to set the goal clearly and parently. This is because the project we are doing is not just making a trained model, which has high accuracy. The models that we made have to be on purpose and help in positive ways.

References

Demographic Health Survey
Ethiopian Demographic Survey 2016
https://dhsprogram.com/data/

Public Health Nutrition: Predicting Undernutrition
Fikerwold H Bitew 2021
https://rb.gy/blux8

Overcoming Class Imbalance using SMOTE
Swastik Satpathy 2020
https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

A Guide to the Types of Machine Learning Algorithms and their Applications
KatrinaWakefield 2023
https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html

Link to Git Repository

https://github.com/Bellwood22/DS5220_Supervised-Machine-Learning/tree/main/Group%20Project