

Attribute Set Reduction Using Filter Selection Method For Machine Learning Model To Diagnose Malnutrition In Preschool Children

Mr. Amol Avinash Shinde⁽¹⁾, Dr. D.V.Sahasrabuddhe⁽²⁾

¹Research Scholar

²BVDU, Institute of Management and Rural Development Administration, Sangli

Abstract

A machine learning model built to detect malnutrition in a preschool child may aid parents in early detection and timely intervention for recovering child's health, especially in situations where experts' consultation is not available. It becomes challenging to diagnose malnutrition in a preschool child based on simple health attribute values. The article explains the steps undertaken for attribute selection playing major role in the malnutrition detection.

Symptoms of malnutrition in a preschool child using Machine Learning algorithm is a quite challenging task which will work and in absence of expert, will help in the early detection of type of Malnutrition the child is suffering with. The process follows the steps defined as Exploratory Data Analysis (EDA), which is considered to be a crucial stage in development of any machine learning model.

METHODS: Based on a sample data collected for preschool children with the help of medical experts different statistical algorithms were executed to find the contribution of each attribute in the detection process. The majorly used techniques were analysing and understanding data using different graphical representations and correlation.

Keywords: Malnutrition Detection, Preschool Children, Machine Learning, EDA, Feature Selection, Filter Method.

1. INTRODUCTION

Malnutrition in preschool children remains a pressing global health concern, particularly in regions with limited access to healthcare and nutritional resources. Early diagnosis of malnutrition is critical to prevent long-term developmental and health consequences. However, identifying malnutrition in children at an early stage can be challenging, especially in areas lacking expert medical supervision. With the advancement of artificial intelligence and data-driven healthcare, machine learning (ML) offers a powerful solution for automated and accurate diagnosis of malnutrition.

This paper presents a comprehensive approach to attribute selection for developing an efficient machine learning model tailored to diagnose malnutrition in preschool children. The study emphasizes the role of Exploratory Data Analysis (EDA), data visualization, and correlation-based filter selection techniques to reduce dimensionality while retaining the most relevant health indicators. By combining medical expertise and statistical analysis, the final attribute set enhances the model's predictive accuracy while ensuring clinical relevance. The research aims to build a model that not only improves diagnostic efficiency but also serves as a practical decision-support system for caregivers and frontline health workers.

2. LITERATURE REVIEW:

The development of a machine learning model for malnutrition detection requires careful attribute selection, guided by prior research. Following research articles like Islam et al. (2022)[1] used ML models to predict malnutrition in women in Bangladesh based on factors like age, BMI, education, income, dietary diversity, number of children, healthcare access, maternal health, and employment, with Random Forest yielding the best results. Ula et al. (2021)[2] applied K-NN for diagnosing stunting in children using features such as weight, height, age, gender, BMI, and dietary intake, focusing on indices like W/A, H/A, and W/H. In a later study, Ula et al. (2022)[2] used Complete Linkage Agglomerative Hierarchical Clustering (CLAHC) to classify children's nutritional status based on biometric and dietary data. Khare et al. (2017)[3] analyzed Indian DHS data using SVM, Decision Trees, and Random Forest to identify malnutrition predictors including birth weight, income, sanitation, immunization, and maternal health. Duraisamy & Sudha (2014)[4] used data mining to categorize children into malnutrition categories using

age, weight, height, BMI, income, and parental education. Najafloo et al. (2021)[5] developed a recommendation system in Iran for ages 6–12 using decision trees and SVM on data like BMI, health status, and diet. Wajgi & Wajgi (2022)[6] used classification algorithms to detect infant malnutrition using physical and dietary parameters. Ahmadi et al. (2017)[7] focused on predicting low birth weight using RF and LR based on maternal factors like BMI, education, and prenatal care. Sharma et al. (2020)[8] highlighted the use of ML in clinical settings for malnutrition detection using intake, biomarkers, and medical history, showing ML's potential to improve diagnosis and treatment.

3. Attributes considered for ML Model for Malnutrition-

This review helped in recognizing gaps in existing studies and guided the selection of attributes that are relevant and effective for diagnosing malnutrition, especially in preschool children. The final list of Attributes was determined by analysing their usage in previous research and consulting with medical professionals to validate their importance from a clinical perspective.

In previous research, some Attributes were either used only in **statistical analyses** (e.g., SPSS, chi-square tests) or were reviewed without being integrated into predictive models. These underutilized attributes hold potential for improving the performance of machine learning-based malnutrition models. Features like Mid Upper Arm Circumference (MUAC), Muscle Circumference, Skinfold and Triceps Fold, Thickness, Peripheral oedema, Hair Colour and Type Changes, Stomatitis (Mouth Sores) have been shortlisted for future model integration due to their potential significance, and as per literature review and expert advice **28 Attributes** were finalized for future use. Table 1 describes Attribute List with their Data types-

Table 1-Attribute List with Data types

Data Type	List of Attributes
Categorical	Gender, Feeding Duration, Birth dieses, Pinch Test, Swelling, Mouth Ulcer, Feeding, Milk (Cow/Buffalo/Goat), Rice, Lentils, Clarified Butter (Ghee), Semolina, Fruits, Egg, All other Foods, Non-veg, Mother Education, Vaccination, Readymade Powder, Ragi, Cooked Fruits,
Continuous	Height, Weight, MUAC, Head Circumference, Family Income
Discrete	Age, Children Count

These attributes, although not heavily used in existing ML-based models, are considered clinically meaningful. Their inclusion could enhance the accuracy of malnutrition detection, especially when combined with standard anthropometric indicators. The aim is to integrate these variables into future implementations and test their impact using supervised machine learning algorithms.

4. Data Collection and Data Understanding-

Considering the above attributes a questionnaire was designed for data collection from three key sources viz. hospitals, nursery schools and in villages using interview method during home visits by Accredited Social Health Activists (ASHA). This multi-source, multi-method strategy ensured that the dataset represented diverse socioeconomic and health conditions, increasing the reliability of the machine learning model in detecting malnutrition.

Total data collected was of 1560 children from those data for 550 children were used for sample model development.

• Data Preparation Steps

The data was directly collected from the sources, primary data, there was not any missing data and the data was consistent. Hence there was not any need to pre-process the data. Each record from the collected raw data was labelled by a medical expert based on their clinical judgment and the health indicators provided. As per the attribute values every record from data was classified as either **Normal**, **Under Nutrition**, **Micro Nutrition**, or **Over Nutrition**. This expert-driven labelling ensured high accuracy and medical validity of the dataset. All collected data was later digitized into Excel sheets and converted to CSV format for model training. For analysing gender wise distribution of data over types of malnutrition, the data was represented using bar diagram as shown in Figure 1 where gender value '0' represents male and '1' represents female.

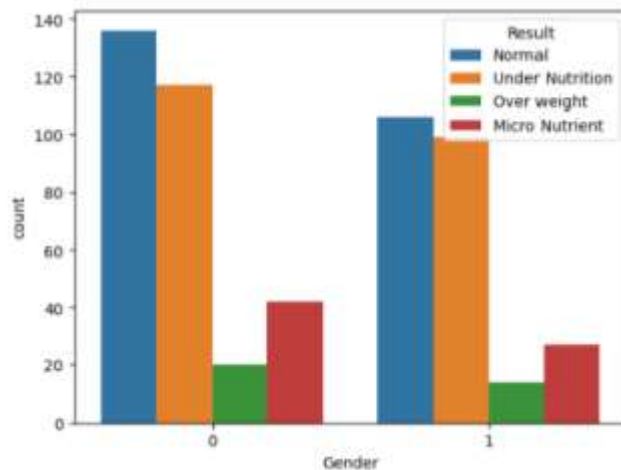


Fig 1-Bar chart representation for Gender for Malnutrition Type

The next step before further step of Exploratory Data Analysis is preparing data for next steps. The data collected is of different types and hence need to be encoded. Encoding for different attributes is explained in Table 2.

Table 2-Attribute Encoding

Sr. No.	Attribute Name	Values	Encoding
1.	Gender	Male	0
		Female	1
2.	Mother Education	Un Educated	0
		Under Matriculation	1
		Matriculation	2
		Higher Secondary	3
		Graduate	4
		Post-Graduate	5
3.	Feeding Duration	6 Months	1
		1 Year	2
		1.5 Year	3
		2 Year	4
4.	Vaccination	Complete Vaccinated	2
		Average Vaccinated	1
		Not Vaccinated	0
5.	Pinch Test	Skin regenerates (in 2 seconds)	0
		Skin does not regenerate (within 2 seconds)	1
6.	Swelling (Frequency)	More	3
		Moderate	2
		Small.	1
		No	0
7.	Mouth Ulcer (Frequency)	Twice a month	2
		Once a month	1
		Frequently	0
8.	Feeding, Readymade Baby Food, Milk (Cow/Buffalo/Goat), Rice, Lentils, Clarified Butter (Ghee), Semolina, Fruits, Eggs,	Yes	1
		No	0

	Ragi, Non-veg, Cooked Fruits, All other Foods, Birth disease		
--	--	--	--

5. Feature Reduction and Correlation Check

Initially Data collected for these 28 attributes was used build a pilot machine learning model using the **Logistic Model Tree (LMT)** and **Simple Logistic Regression** algorithms from literature which were found well suited classification algorithms for this dataset. Though LMT and Simple Logistic **Regression** showed high accuracy, using all 28 attributes at once increased model complexity, which could potentially lead to overfitting and reduced generalizability. As illustrated in **Figure 4** and **Figure 5**, these algorithms tend to internally select a subset of important attributes, ignoring the less relevant ones. This indicates the need for a attributes reduction mechanism to reduce dimensionality while retaining prediction accuracy.

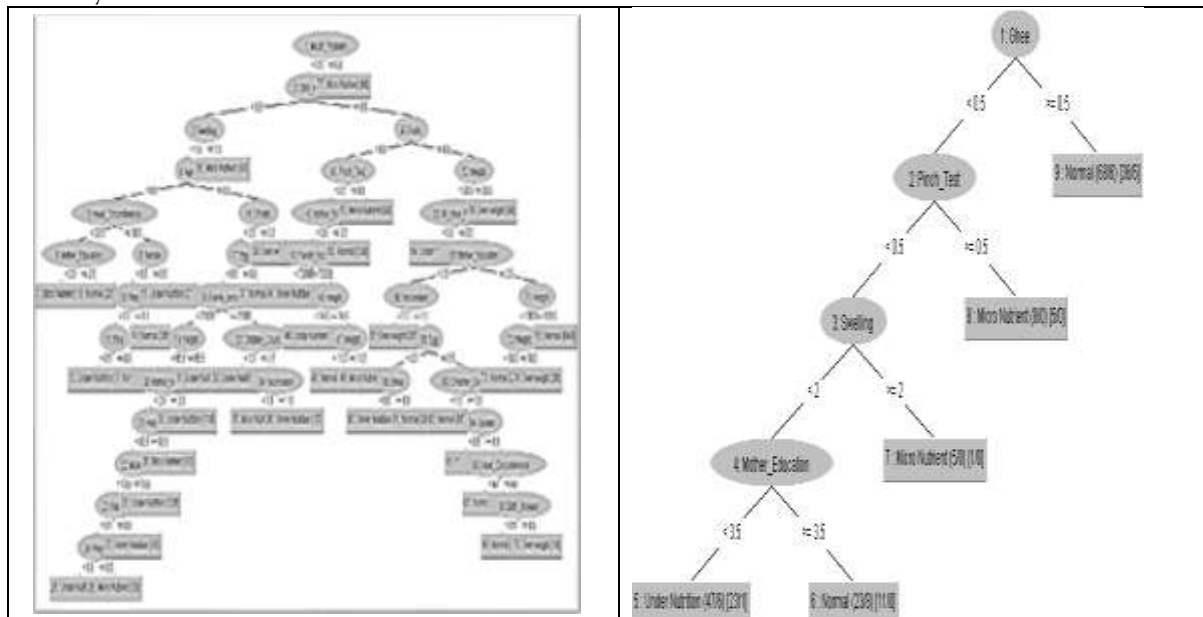


Figure 4 and 5 : Decision Tree Representation of Model

Correlation of Data

To support the development of a more accurate and interpretable machine learning model for predicting children's nutritional status, two correlation heat maps were generated and analysed. The first, shown in **Figure 6 – Full Correlation Heatmap**, presents the complete correlation matrix of all 28 attributes in the dataset. This visualization was used to examine the internal relationships between all features, helping to identify redundant variables and observe which attributes may influence the target class (Result). While Figure 10 offered a broad overview, it also contained many weak or insignificant relationships, making it less effective for direct feature selection.

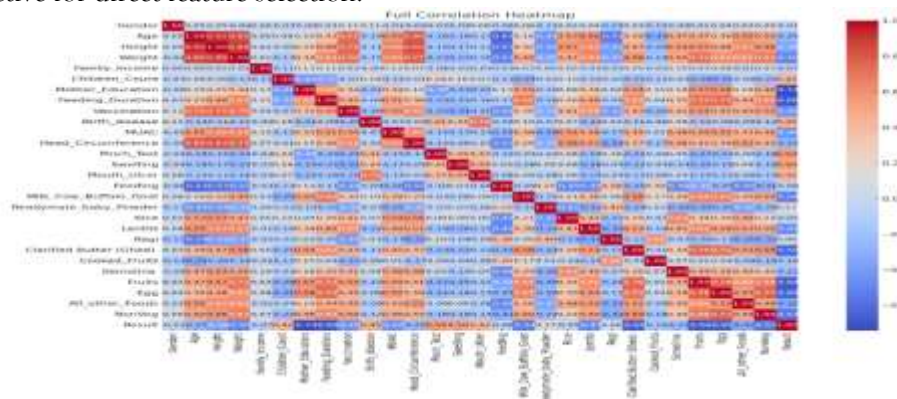


Fig.6- Full Correlation Heatmap

To refine the focus, a second heatmap—Figure 7 – *Strong Correlations with Result*—was created, highlighting only those attributes with strong or moderate correlation values linked directly to the Result variable. This targeted view revealed that attributes such as Mother Education ($r = -0.74$), Feeding Duration ($r = -0.68$), Fruits ($r = -0.68$), Egg consumption ($r = -0.66$), Clarified Butter (Ghee) ($r = -0.66$), Non-Vegetarian Food ($r = -0.57$), and Milk from Cow/Buffalo/Goat ($r = -0.54$) were strongly negatively correlated with poor nutritional outcomes, indicating their protective impact. In contrast, the Pinch Test ($r = +0.54$) showed a positive correlation, marking it as a clinical symptom strongly associated with malnutrition.

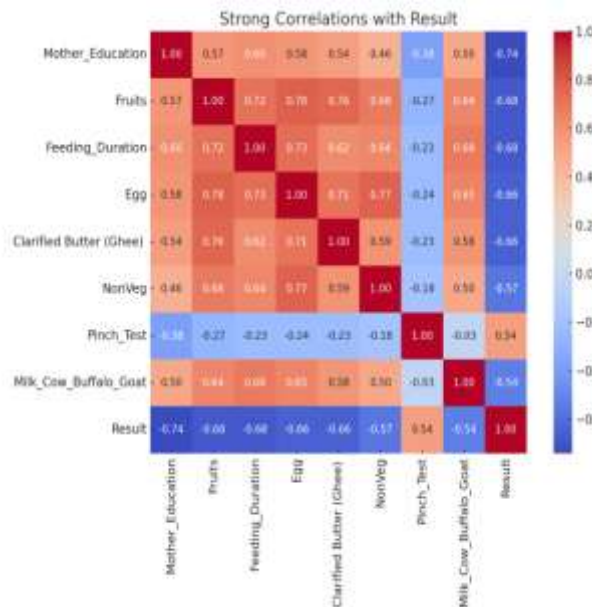


Fig. 7 -Strong Correlation with Results
Attribute Prioritization

Based on the correlation matrix and expert recommendations, a refined list of features was created. Features with consistently low or non-significant correlations were marked for removal, while those with strong correlations or expert-supported relevance were retained for model training.

Selected due to high correlation	Selected due to high correlation and by Expert Suggestion	Selected Expert Suggestion	Rejected
Lentils Clarified Butter (Ghee) Fruits Egg	Age Height Weight Feeding Duration Head Circumference Milk (Cow/Buffalo/Goat) Rice Semolina All other Foods Non veg	Gender Birth dieses Mid-Upper Arm Circumference(MUAC) Pinch Test Swelling Mouth Ulcer Feeding	Family Income Children Count Mother Education Vaccination Readymade Powder Ragi Cooked Fruits

Table 3: List of selected and rejected attributes

In Table number 5 the final lists all **21 selected** and **7 rejected** attributes for the machine learning model. The modified dataset will now be used for model training, testing, and evaluation. The selected features ensure that the model is trained on variables that are both statistically meaningful and clinically justified. This structured and data-driven approach to attribute selection significantly improves the reliability, interpretability, and accuracy of the malnutrition prediction model.

6. CONCLUSION –

This study highlights the value of selecting and reducing attributes to build an efficient machine learning model for diagnosing malnutrition in preschool children. Using correlation-based methods and expert input, 21 key features were chosen from 28, improving performance without losing accuracy. The model supports early detection of various malnutrition types and offers a low-cost, scalable solution for community healthcare. Future improvements may include image-based or deep learning approaches, strengthening its role in public health.

REFERENCES-

1. Islam MM, Rahman MJ, Islam MM, Roy DC, Ahmed NF, Hussain S, Amanullah M, Abedin MM, Maniruzzaman M. Application of machine learning based algorithm for prediction of malnutrition among women in Bangladesh. *International Journal of Cognitive Computing in Engineering*. 2022 Jun 1;3:46-57.
2. Ula, M., Ulva, A. F., Mauliza, M., Sahputra, I., & Ridwan, R. (2021). Implementation of Machine Learning in Determining Nutritional Status using the Complete Linkage Agglomerative Hierarchical Clustering Method. *Jurnal Mantik*, 5(3), 1910-1914.
3. Khare, S., Kavyashree, S., Gupta, D., & Jyotishi, A. (2017). Investigation of nutritional status of children based on machine learning techniques using Indian demographic and health survey data. *Procedia computer science*, 115, 338-349.
4. Thangamani, D., & Sudha, P. (2014). Identification of malnutrition with use of supervised datamining techniques–decision trees and artificial neural networks. *Int J Eng Comput Sci*, 3(09).
5. Najaflou, S., & Rabiei, M. (2021). Recommended system for controlling malnutrition in Iranian children 6 to 12 years old using machine learning algorithms. *International Journal of Web Research*, 4(1), 27-33.
6. Wajgi, R., & Wajgi, D. (2022, March). Malnutrition detection in infants using machine learning approach. In *AIP Conference Proceedings* (Vol. 2424, No. 1). AIP Publishing.
7. Ahmadi, P., Alavimajd, H., Khodakarim, S., Tapak, L., Kariman, N., Amini, P., & Pazhuheian, F. (2017). Prediction of low birth weight using Random Forest: A comparison with Logistic Regression. *Archives of Advances in Biosciences*, 8(3), 36-43.
8. Sharma, V., Sharma, V., Khan, A., Wassmer, D. J., Schoenholtz, M. D., Hontecillas, R., ... & Abedi, V. (2020). Malnutrition, health and the role of machine learning in clinical setting. *Frontiers in nutrition*, 7, 44.