International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 937-944

Catboost Modeling In The Classification Of Factors Influencing Nutritional Status In Toddlers

Muhammad Amin[1], Fadhillah Fitri[2]

[1,2]Department of Statistics, Faculty of Mathematics and Natural Sciences

Coresponding Author: Muhammad Amin, aminmuhammad0301@gmail.com

Abstract– Childhood malnutrition remains a major public health challenge in Indonesia due to the complex interactions among maternal, socioeconomic, and demographic factors influencing nutritional status. This study aimed to develop a CatBoost-based classification model to identify the determinants of nutritional status among children under five in West Sumatra, Indonesia. A quantitative predictive research design was employed using secondary survey data. Data preprocessing included cleaning, handling missing values, feature selection, and partitioning the dataset into training and testing subsets. The CatBoost classifier was trained through hyperparameter optimization and evaluated using accuracy, precision, recall, F1-score, confusion matrix, and receiver operating characteristic (ROC) analysis. The optimal model was obtained with a tree depth of four and 400 boosting iterations, achieving an accuracy of 45% and a Macro F1-score of 0.4282. Feature importance analysis identified maternal nutritional knowledge as the most influential predictor, followed by maternal education, household per-capita income, maternal employment status, and maternal age. These findings demonstrate that CatBoost is a promising machine learning approach for identifying key determinants of childhood nutritional status and can support evidence-based nutritional surveillance and targeted public health interventions.

Keywords: Chatboost, Child Nutritional Status, Predictive Modeling, West Sumatera.

I. Introduction

Childhood nutritional status remains one of the most important indicators of population health and human capital development. Adequate nutrition during the first 1,000 days of life is essential for optimal physical growth, cognitive development, immune function, and long-term productivity. Conversely, malnutrition during early childhood may result in irreversible consequences, including impaired neurodevelopment, reduced educational attainment, increased susceptibility to infectious diseases, and a higher risk of chronic non-communicable diseases in adulthood. Despite substantial global efforts to improve maternal and child nutrition, malnutrition continues to pose a significant public health challenge, particularly in low- and middle-income countries (LMICs)[1], [2], [3].

The World Health Organization (WHO) classifies childhood malnutrition into several major forms, including stunting, wasting, underweight, overweight, and obesity, each representing different biological and epidemiological conditions [4]. According to recent global estimates, millions of children under five continue to experience one or more forms of malnutrition, while the coexistence of undernutrition and overnutrition has created a double burden of malnutrition in many developing countries [5], [6]. Indonesia remains among the countries facing this challenge. The 2024 Indonesian Nutritional Status Survey reported that the prevalence of stunting, wasting, and underweight among children under five remained at levels requiring sustained public health interventions [7]. Moreover, several provinces, including West Sumatra, continue to report nutritional indicators exceeding internationally recommended thresholds, highlighting the need for more effective identification of children at nutritional risk and evidence-based intervention strategies[8].

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 937-944

Nutritional status is influenced by a complex interaction of biological, behavioral, socioeconomic, environmental, and healthcare-related factors. Previous studies have demonstrated that maternal education, maternal age, nutritional knowledge, household income, breastfeeding practices, sanitation, food security, and healthcare accessibility all contribute to childhood nutritional outcomes [4], [9], [10], [11]. These determinants rarely operate independently; instead, they interact through complex and often non-linear relationships. Consequently, identifying the most influential determinants remains challenging when conventional statistical techniques are applied to heterogeneous population data.

Multiple regression has traditionally been employed to investigate determinants of childhood nutritional status because of its ability to estimate associations between predictor variables and categorical outcomes. However, regression-based approaches require several statistical assumptions, including linearity between predictors and the log odds, independence of observations, limited multicollinearity, and adequate sample size. Violation of these assumptions may reduce model reliability and predictive performance, particularly when datasets contain numerous categorical variables, missing values, complex interactions, or non-linear relationships. These limitations have encouraged researchers to explore machine learning approaches that prioritize predictive performance while accommodating complex data structures [12][13].

Recent advances in machine learning have demonstrated considerable potential in nutritional epidemiology and public health research. Algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks have been successfully applied to disease prediction, nutritional risk assessment, obesity classification, and child health surveillance [14], [15], [16]. Among these algorithms, CatBoost (Categorical Boosting), developed by Yandex, has emerged as one of the most promising gradient boosting techniques for datasets containing numerous categorical variables. Unlike many conventional machine learning algorithms, CatBoost directly processes categorical features through ordered target statistics while minimizing target leakage and overfitting. Furthermore, the algorithm efficiently handles missing values, automatically captures interactions among variables, and frequently demonstrates competitive predictive performance with minimal feature engineering [17], [18], [19], [20], [21].

Given these methodological advantages, CatBoost offers considerable potential for identifying determinants of nutritional status among children under five using multidimensional health datasets. Nevertheless, its application within nutritional epidemiology, particularly in Indonesia, remains limited. Therefore, this study aims to develop a CatBoost classification model to identify the factors influencing the nutritional status of toddlers in West Sumatra, Indonesia. In addition to evaluating model performance using multiple classification metrics, this study also identifies the most influential predictors through feature importance analysis, thereby providing evidence that may support targeted nutritional intervention programs and data-driven public health policy.

II. MÉTHODOLOGIE

2.1 Study Design

This quantitative study employed a predictive modeling approach using supervised machine learning to classify the determinants of nutritional status among children under five. The research was conducted using secondary data collected by the Center for Food, Nutrition, Family and Community Empowerment Research, Universitas Negeri Padang, in 2022. The study focused on identifying the most influential factors associated with children's nutritional status through the CatBoost classification algorithm. A quantitative predictive design is appropriate because the primary objective was to develop a classification model capable of learning patterns from labeled observations and predicting nutritional status based on multiple demographic, socioeconomic, and maternal characteristics [4], [9], [10], [11], [22].

2.2 Data Collection and Study Variables

The study utilized a secondary dataset obtained from household nutritional surveys conducted in West Sumatra Province. The dataset contains individual records of children under five and their caregivers, including demographic characteristics, maternal characteristics, socioeconomic conditions, and nutritional information. Prior to model development, all records were anonymized to ensure confidentiality.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 937-944

The response variable was children's nutritional status, classified according to the WHO Child Growth Standards based on anthropometric assessment. Predictor variables consisted of maternal age, maternal education, maternal occupation, maternal nutritional knowledge, household income per capita, and several additional demographic variables available in the dataset. These variables were selected because previous epidemiological studies consistently identified them as important determinants of childhood nutritional status.

Table 1. Description of study variables

Variable	Role
Maternal age	X1
Maternal Job	X2
Maternal education	X3
Maternal nutritional knowledge	X4
Household income per capita	X5
Nutritional status	X6

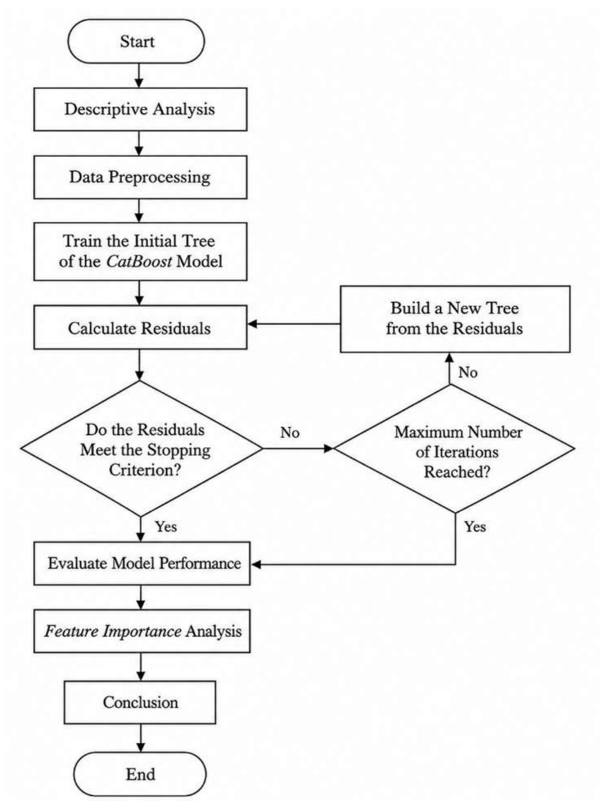

Figure 1. Overall research workflow

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 937-944

2.3 Data Preprocessing

Data preprocessing was performed before model construction to improve data quality and predictive performance. Duplicate observations were removed, while missing values were identified and treated using appropriate imputation methods. Previous studies have demonstrated that careful preprocessing substantially improves machine learning performance and model stability.

The dataset was subsequently divided into training and testing subsets. The training dataset was used to develop the classification model, whereas the testing dataset was reserved for independent performance evaluation. Separating training and testing data prevents optimistic bias and provides a more realistic estimate of model performance on unseen observations.

2.4 Model Evaluation

The predictive performance of the CatBoost model was evaluated using the independent testing dataset. Classification performance was assessed using confusion matrix, accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These evaluation metrics provide complementary information regarding overall predictive accuracy, class discrimination ability, and model robustness, particularly for multiclass classification problems [23].

The confusion matrix summarizes correctly and incorrectly classified observations for each nutritional status category. Precision measures the proportion of correctly predicted positive observations, recall evaluates the proportion of actual positive observations correctly identified by the model, whereas the F1-score represents the harmonic mean of precision and recall. Meanwhile, AUC evaluates the model's ability to discriminate among nutritional status classes across different classification thresholds [23].

Finally, feature importance generated by CatBoost was analyzed to identify the variables contributing most strongly to nutritional status classification. Feature importance provides interpretable information regarding the relative contribution of each predictor and has become an important component of explainable machine learning in healthcare research.

III. RÉSULTS AND ANALYSES

3.1 Characteristics of the Classification Model

This study used secondary data from the Center for Food, Nutrition, Family, and Community Empowerment Research at Padang State University in 2022. After the data cleaning process, 634 sets of data on toddlers eligible for analysis were obtained from the initial 644 data sets. The application of CatBoost in this study was carried out in several stages, namely data cleaning, splitting the data into training and test sets, model training, performance evaluation, and interpretation of key factors.

The CatBoost classifier was developed to classify the nutritional status of children under five using maternal and household characteristics as predictor variables. Hyperparameter tuning was performed by evaluating several combinations of tree depth and boosting iterations. Among the evaluated models, the optimal configuration consisted of a tree depth of **4** and **400 boosting iterations**, which produced the highest Macro F1-score (**0.4282**). Increasing tree depth beyond four levels or increasing the number of iterations up to 1,000 did not improve model performance, indicating that a more complex model did not necessarily enhance classification accuracy.

3.2 Classification Performance

The predictive performance of the selected CatBoost model was evaluated using an independent testing dataset. Overall, the model achieved an **accuracy of 45%**, indicating that nearly half of the nutritional status categories were correctly classified. The weighted F1-score (0.46) was slightly higher than the macro F1-score (0.43), suggesting differences in predictive performance across nutritional status categories.

The class-wise evaluation demonstrated that the model performed best in identifying children with **normal nutritional status**, yielding a precision of **0.60** and an F1-score of **0.52**. Lower predictive performance was observed for the **undernutrition** and **overnutrition** categories, with F1-scores of **0.38** and **0.39**, respectively.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
SCHOLAR AI Be Smart
Vol. 58 No. 2 August 2026, pp. 937-944

3.3 Confusion Matrix Analysis

The confusion matrix revealed substantial variation in prediction accuracy across nutritional status categories. Among **71 children** with normal nutritional status, **32** were correctly classified, whereas the remaining observations were assigned to other nutritional categories. Likewise, among **37 children** classified as overnutrition, only **16** were correctly predicted. These findings indicate that the model experienced considerable difficulty distinguishing between adjacent nutritional categories, particularly normal and overnutrition.

3.4 Receiver Operating Characteristic Analysis

Model discrimination was further evaluated using Receiver Operating Characteristic (ROC) analysis. The highest discriminative performance was obtained for the **undernutrition** class, with an Area Under the Curve (AUC) of **0.66**, followed by the **overnutrition** class (**AUC = 0.62**). In contrast, the **normal nutritional status** category achieved an AUC of **0.50**, indicating limited discriminative ability.

Figure 1. Kurva ROC - AUC

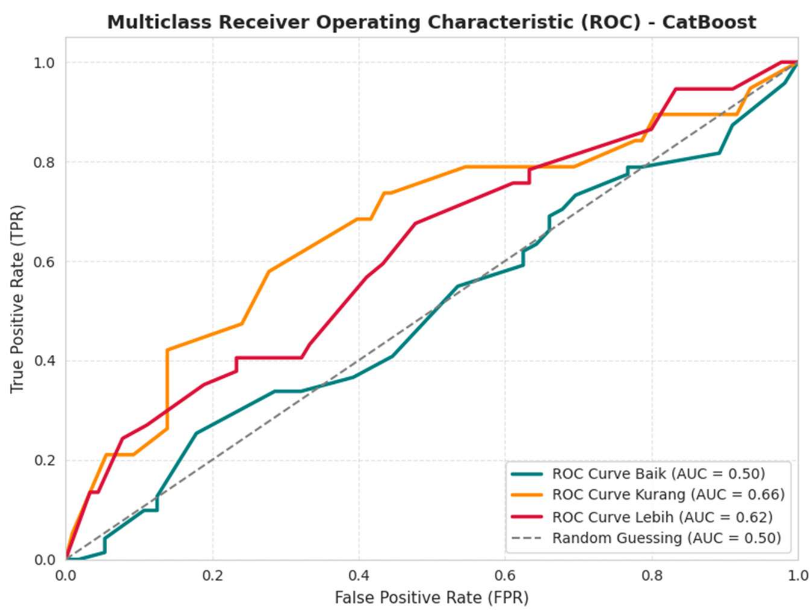

Figure 1 demonstrates that the CatBoost model discriminated undernutrition more effectively than the remaining nutritional status categories.

3.5 Feature Importance

Feature importance analysis identified **maternal nutritional knowledge** as the most influential predictor of children's nutritional status. This variable was followed by **maternal education**, **household per-capita income**, **maternal employment status**, and **maternal age**, respectively.

International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

https://ijpsat.org/
Vol. 58 No. 2 August 2026, pp. 937-944

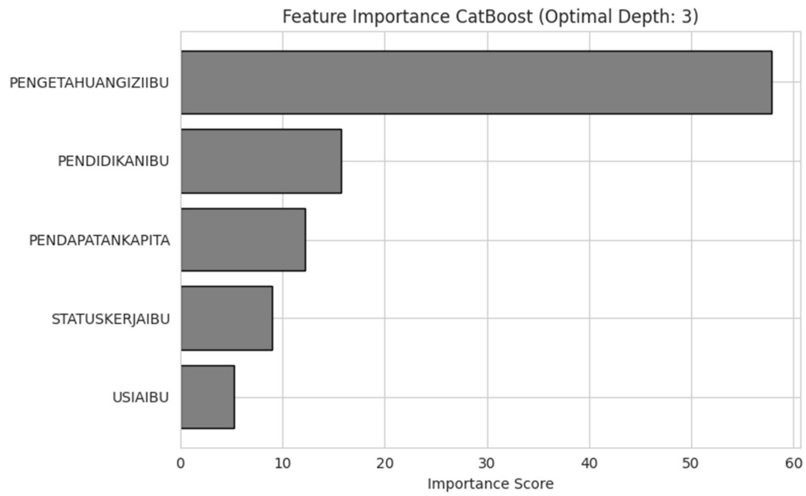

Figure 2. Ranking of predictor importance generated by CatBoost

Figure 2. presents the relative importance of predictor variables generated by the CatBoost algorithm. Maternal nutritional knowledge contributed the greatest proportion to the classification model, followed by educational and socioeconomic factors.

Overall, the findings demonstrate that CatBoost successfully identified several important determinants associated with nutritional status among children under five. Although the model achieved moderate predictive performance, the feature importance analysis consistently highlighted maternal knowledge and socioeconomic characteristics as the principal contributors to nutritional status classification.

IV. CONCLUSION

This study developed a CatBoost-based classification model to identify the determinants of nutritional status among children under five in West Sumatra, Indonesia. The findings demonstrate that the proposed model was able to classify nutritional status with moderate predictive performance while effectively identifying the relative importance of the predictor variables. Hyperparameter optimization resulted in the best-performing model with a tree depth of four and 400 boosting iterations, indicating that an appropriately tuned CatBoost model can capture complex relationships among demographic, socioeconomic, and maternal factors without excessive model complexity.

Feature importance analysis revealed that maternal nutritional knowledge was the strongest predictor of children's nutritional status, followed by maternal education, household per-capita income, maternal employment status, and maternal age. These findings suggest that maternal knowledge and socioeconomic conditions play a central role in determining child nutritional outcomes and should therefore be prioritized in nutrition intervention programs and public health policies. Overall, the research objective was successfully achieved. The CatBoost algorithm not only provided a predictive classification model but also generated interpretable information regarding the factors contributing most to nutritional status. These results demonstrate the potential of machine learning approaches to complement conventional epidemiological analyses by supporting evidence-based decision-making in child nutrition programs.

Future studies should consider larger and more geographically diverse datasets, incorporate additional environmental and behavioral determinants, and compare CatBoost with other state-of-the-art machine learning algorithms to further improve predictive performance and model generalizability.

References

[1] P. R. Indonesia., "Undang-Undang Republik Indonesia Nomor 59 Tahun 2024 tentang Rencana Pembangunan Jangka Panjang Nasional Tahun 2025–2045," *https://jdih*, 2024.

[2] S. W. P. R. Indonesia., "Strategi Nasional Percepatan Pencegahan dan Penurunan Stunting 2025–2029," *https://stunting*, 2024.

[3] M. B. Ali, R. Tuhin, M. A. Alim, M. Rokonuzzaman, S. M. Rahman, and M. Nuruzzaman, "Acceptance and use of ICT in tourism: the modified UTAUT model," *J. Tour. Futur.*, vol. 10, no. 2, pp. 334–349, 2024, doi: 10.1108/JTF-06-2021-0137.

[4] World Health Organization, "Malnutrition," 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/malnutrition/

[5] A. T. Abeng, D. Ismail, and E. Huriyati, "Sanitasi, infeksi, dan status gizi anak balita di Kecamatan Tenggarong Kabupaten Kutai Kartanegara," *J. Gizi Klin. Indones.*, vol. 10, no. 3, p. 159, 2014, doi: 10.22146/ijcn.18867.

[6] E. Suminar and A. R. Wibowo, "The Correlation Between Infection Diseases History and Nutritional Status in Toddler," *Fundam. Manag. Nurs. J.*, vol. 4, no. 1, p. 18, 2021, doi: 10.20473/fmnj.v4i1.21587.

[7] Kemenkes RI, "Hasil Survei Status Gizi Indonesia (SSGI) 2022," *Kemenkes*, pp. 1–150, 2022.

[8] WHO., "Physical status: The use and interpretation of anthropometry," *WHO Tech. Rep. Ser. No*, 1995.

[9] L. A. Budiman Rosiyana, R., Sari, A. S., Safitri, S. J., Prasetyo, R. D., Rizqina, H. A., Neng I Kasim, I. S., & Indriany Korwa, V. M., "Analisis Status Gizi Menggunakan Pengukuran Indeks Massa Tubuh dan Beban Kerja dengan Metode 10 Denyut pada Tenaga Kesehatan," *Nutr. Nutr. Res. Dev. J.*, 2021.

[10] I. S. Paramita, H. Atasasih, and D. Rahayu, *Penilaian Status Gizi Antropometri Pada Balita*, 1st ed. CV. Sarana Ilmu Indonesia, 2024.

[11] F. K. Lailani, Yuliana, and A. Yulastri, "Literature Riview : Masalah Terkait Malnutrisi: Penyebab, Akibat, dan Penanggulangannya," *JGK J. Gizi dan Kesehat.*, vol. 2, no. 2, pp. 129–138, 2022, doi: 10.36086/jgk.v2i2.1503.

[12] N. E. Rosuliana, T. Nurhayati, E. Mawaddah, and M. U. Ningsih, *Buku saku deteksi dini dan perawatan balita pneumonia*. Cipedes: Perkumpulan Rumah Cemerlang Indonesia ANGGOTA IKAPI JAWA BARAT Pondok, 2024.

[13] A. A. Reddy *et al.*, "Perspectives on Forest governance among the indigenous communities of India's Eastern Ghats," *For. Policy Econ.*, vol. 169, p. 103350, 2024, doi: https://doi.org/10.1016/j.forpol.2024.103350.

[14] R. Hidayat Mahdiana, D., & Fergina, A., "Comparative Analysis of Logistic Regression, SVM, Xgboost, and Random Forest Algorithms for Diabetes Classification," *J. Teknol. Sist. Inf. Dan Apl.*, 2024.

[15] S. Ramadhani & Wayahdi, M. R., "K-Nearest Neighbor and Random Forest Algorithms in

Loan Approval Prediction," *J. Minfo Polgan*, 2024.

[16] Y. F. Zamzam Saragih, T. H., Herteno, R., Muliadi, Nugrahadi, D. T., & Huynh, P.-H., "Comparison of Catboost and Random Forest Methods for Lung Cancer Classification using Hyperparameter Tuning Bayesian Optimization-based," *J. Electron.*, 2024.

[17] L. Prokhorenkova Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A., "Catboost: Unbiased boosting with categorical features," *32nd Conf. Neural Inf. Process. Syst.*, 2018.

[18] J. T. Hancock & Khoshgoftaar, T. M., "Catboost for big data: An interdisciplinary review," *J. Big Data*, 2020.

[19] A. Maulana Afidh, R. P. F., Maulydia, N. B., Idroes, G. M., & Rahimah, S., "Predicting Obesity Levels with High Accuracy: Insights from a Catboost Machine Learning Model," *Infolitika J. Data Sci.*, 2024.

[20] S. S. M. Alqrinawi Burhanuddin, M. A., & Salahuddin, L., "Catboost Model for Enhanced Treatment Prediction in Type 2 Diabetes Patients," *J. Inf. Syst. Eng. Manag.*, 2025.

[21] Y. Zhang Zhang, H., Wang, D., Li, N., Lv, H., & Zhang, G., "Development of a 5-Year Risk Prediction Model for Transition From Prediabetes to Diabetes Using Machine Learning: Retrospective Cohort Study," *J. Med. Internet Res.*, 2025.

[22] B. Mutonhodza, M. G. Manzeke-Kangara, E. H. Bailey, T. M. Matsungo, and P. Chopera, "Maternal selenium deficiency was positively associated with the risk of selenium deficiency in children aged 6–59 months in rural Zimbabwe," *PLOS Glob. Public Heal.*, vol. 4, no. 7, p. e0003376, 2024, doi: 10.1371/journal.pgph.0003376.

[23] D. T. Wilujeng Fatekurohman, M., & Tirta, I. M., "Analisis Risiko Kredit Perbankan Menggunakan Algoritma K-Nearest Neighbor dan Nearest Weighted K-Nearest Neighbor," *Indones. J. Appl. Stat.*, 2023.