Health and Socio-Demographic Risk Factors of Childhood Stunting: Assessing the Role of Factor Interactions Through the Development of an AI Predictive Model

Taqwa Hariguna[1,*], Sarmini[2], Abdul Azis[3]

[1,2,3] *Department of Information System, Universitas Amikom Purwokerto, Indonesia*

(Received: August 24, 2024; Revised: October 11, 2024; Accepted: October 21, 2024; Available online: December 14, 2024)

Abstract

Stunting is a significant global health problem, especially in developing countries such as Indonesia. This study aims to develop and evaluate an artificial intelligence (AI)-based predictive model to identify the risk of stunting in children using the CatBoost algorithm which is a combination of Weighted Apriori and XGBoost. This model is designed to utilize the advantages of each algorithm in handling data with variable weights to improve prediction accuracy. Feature analysis shows that "Height (cm) & Age (months)" are the main indicators in classifying children's nutritional status. Model evaluation shows high accuracy of 94.85%, precision of 95%, recall of 94.85%, and F1 Score of 94.84%. Kappa Coefficient and Matthews Correlation Coefficient (MCC) reached 93.13% and 93.19%, respectively, while ROC-AUC reached 99.70%. These findings indicate that the CatBoost model can provide highly accurate results in detecting the risk of stunting and offer in-depth insights into risk factors that can improve the effectiveness of health interventions. This study fills the gap in the literature by integrating the Weighted Apriori and XGBoost algorithms, providing a significant contribution to early detection of stunting and supporting government efforts to reduce the prevalence of stunting in Indonesia and other regions.

Keywords: Stunting, Artificial Intelligence, CatBoost, Weighted Apriori, XGBoost, Risk Prediction, Health Intervention

1. Introduction

Stunting is a global health problem that is still a major challenge, especially in developing countries. According to the latest report from the World Health Organization (WHO), more than 149 million children under the age of five experienced stunting in 2020, with South Asia and Sub-Saharan Africa being the regions with the highest prevalence [1]. Stunting is a form of chronic malnutrition that occurs due to a lack of nutritional intake over a long period of time. This condition not only affects physical growth but also has implications for cognitive development, educational performance, and individual productivity in adulthood [2].

In Indonesia, the prevalence of stunting is still high even though various intervention programs have been launched. Based on data from the 2018 Basic Health Research (Riskesdas), the prevalence of stunting in Indonesia reached 30.8% [3]. This figure is of concern to the government, which is targeting a reduction in the prevalence of stunting to 14% by 2024 [4]. Several studies have shown that the prevalence of stunting is closely related to factors such as access to proper sanitation, socio-economic status, and maternal knowledge of parenting patterns [5]. However, the main challenge in handling stunting is early identification of cases, because generally symptoms of stunting are only seen after significant physical impacts occur [6]. Therefore, an accurate and rapid early detection system is needed so that interventions can be carried out more effectively.

Artificial Intelligence (AI) and machine learning are increasingly popular approaches in the health sector to address a variety of health issues, including malnutrition and stunting [7]. These technologies enable more efficient and effective data analysis through pattern recognition from big data, which cannot be manually identified by humans [8]. Several machine learning algorithms such as Decision Tree, Random Forest, and Support Vector Machine (SVM) have been widely used for health prediction, including the risk of stunting in children [9], [10], [11]. For example, Khan et al. used Random Forest to predict the risk of stunting in Pakistan, with an accuracy rate of 85% [12]. Meanwhile, Lee et

*Corresponding author: Taqwa Hariguna (taqwa@amikompurwokerto.ac.id)

DOI: https://doi.org/10.47738/jads.v5i4.612

al. used XGBoost to predict stunting in Vietnam with quite promising results, demonstrating the potential of machine learning in accelerating stunting detection and intervention [13].

However, the standard apriori algorithm used in many studies still has limitations in handling heterogeneous data, especially in terms of variable weights [14]. The Weighted Apriori algorithm emerged as a solution to overcome these limitations, by taking into account variable weights in the analysis of associative patterns, so that the results obtained are more accurate in the context of health analysis [15]. Weighted Apriori has been applied in various fields, including e-commerce data analysis, but its use in the health context, especially for stunting prediction, is still limited [16]. In addition, the XGBoost algorithm has also been proven effective in handling complex data and has been widely used in various predictive applications, including in the health sector [17]. The development of the Weighted XGBoost algorithm offers better capabilities to overcome data problems that have variables with different weights [18].

Previous research by Citrakesumasari et al. [19] used Random Forest to predict malnutrition in children based on demographic factors, with results showing that environmental conditions such as access to clean water greatly influence the risk of malnutrition. Another study by Simamora et al. [20] used the Weighted Apriori method to predict the risk of malnutrition in India, with a higher level of accuracy compared to the standard apriori method. These studies demonstrate the great potential of machine learning algorithms to analyze complex health risk factors.

On the other hand, although various machine learning algorithms have been used for stunting prediction, studies that combine the Weighted Apriori and Weighted XGBoost algorithms are still rare in the literature. The combination of these two algorithms is expected to improve the accuracy of stunting prediction by utilizing the advantages of each algorithm. Weighted Apriori functions to find associative patterns from weighted data, while Weighted XGBoost is able to optimize predictions by taking into account important variables in heterogeneous data [21]. This study will fill this gap and provide a new approach to early detection of stunting that can be applied on a wider scale.

Several previous studies have applied machine learning in the health sector, especially to predict the risk of stunting. For example, Yunus et al. [12] using Support Vector Machine (SVM) and Random Forest to predict stunting based on socio-economic and environmental factors in Pakistan. Another study by Putri et al. [22] applied Decision Tree to predict stunting risk in Indonesia , which showed that maternal nutritional intake and health factors greatly influenced the risk of stunting in children [22]. However, a weakness of these studies is the lack of handling of variables with different weights, which are often important factors in health data.

In addition, research by Citrakesumasari et al. [19] showed that Random Forest and XGBoost are effective algorithms for analyzing health data with high prediction accuracy. However, these algorithms do not fully take into account the weight of each variable in the data [23]. Research by Simamora et al. [20] also showed that the use of Weighted Apriori was able to increase prediction accuracy by taking into account the weight of variables that had a greater influence on the final result.

This study provides novelty by combining two algorithms, Weighted Apriori and Weighted XGBoost, which have not been widely explored in the literature related to stunting detection in children. This combination of algorithms is expected to be able to capture associative patterns between variables better and provide more accurate prediction results by considering the weight of each variable in the dataset. In addition, this study also uses more comprehensive data, including demographic factors, nutritional intake, health environment, and child and family medical history [24].

With this approach, this study is expected to not only fill the gap in the literature but also provide a real contribution in accelerating early detection of stunting. The resulting predictive model is expected to provide more in-depth information on stunting risk factors, so that health interventions can be carried out more precisely and more effectively.

2. Methodology

The flow of this research is illustrated in figure 1 and the explanation below the figure explains how the process of identifying stunting in toddlers is carried out systematically and how further analysis using multi-class models and classifications is applied.

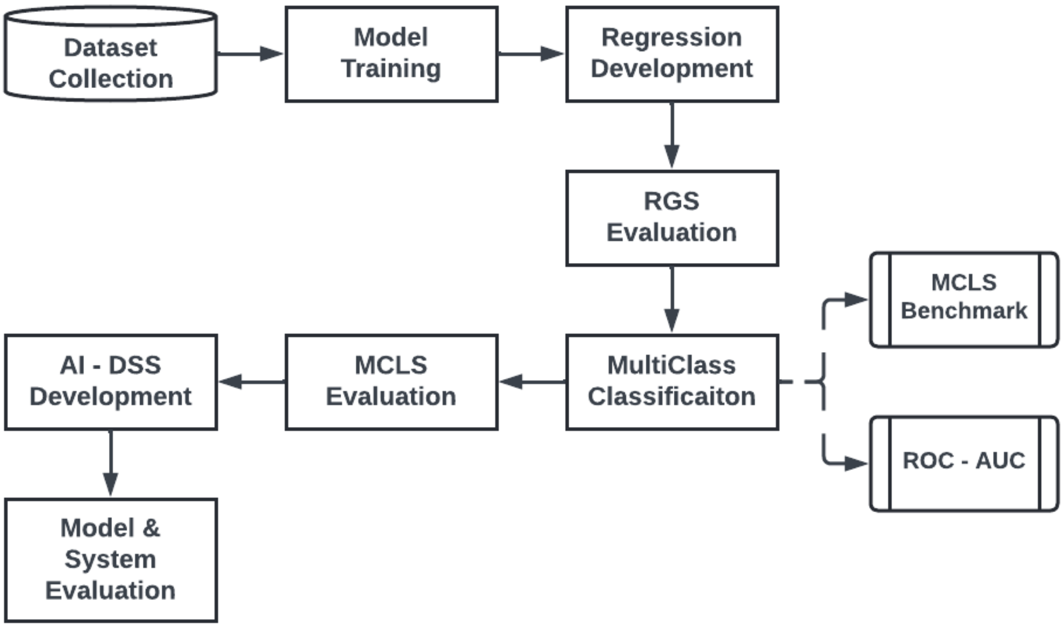


Figure 1. Research phase

2.1. Dataset Collection

The data collection process is an important initial step in this research. The dataset used comes from Kaggle, which is based on the z-score formula to determine stunting status according to WHO standards. This dataset is very rich in information, consisting of 121,000 rows of data, which includes important variables such as age, gender, height, and nutritional status. These variables reflect the main factors that affect the growth of children under five years of age. By using these indicators, researchers can map the health and growth conditions of toddlers for further prediction and intervention purposes.

The variables involved provide a solid basis for analysis. For example, the age column is used to measure the developmental stage of the toddler, while gender provides additional insights as gender differences may affect growth patterns. Height, as one of the main indicators of growth status, allows researchers to determine whether a child is experiencing growth retardation or not. Nutritional status, categorized into four levels (severely stunted, stunted, normal, tall), helps quickly identify children who need special attention.

2.2. Model Training

The model training stage begins by dividing the dataset into two parts, namely training data and test data (train-test split). This separation is important to ensure that the developed model does not overfit the training data and is able to generalize well to new data.

At this stage, two analytical approaches are applied: regression and multi-class classification. Regression is used to predict the relationship between input variables (such as age, gender, and height) with the output of nutritional status. The development of the regression model is carried out using various algorithms which are then compared based on their performance (RGS Benchmark). After that, an evaluation is carried out to assess the performance of the regression model on the test and validation data, including determining which features have the most influence on the prediction (feature importance).

2.3. MultiClass Classification

Next, a multi-class classification approach is applied to classify toddlers based on their nutritional status (severely stunted, stunted, normal, and tall). At this stage, benchmarking of different multi-class classification models is carried out to compare their performance. One of the evaluation metrics used is ROC - AUC, which helps in assessing how well the model is able to distinguish between different classes.

CatBoost (Categorical Boosting) is a gradient-based boosting algorithm developed to handle prediction problems, both regression and classification, with optimal performance in handling categorical data. Like XGBoost, CatBoost uses a boosting approach to combine predictions from multiple low-level decision tree models (weak learners) to improve accuracy. However, the main difference is in the way CatBoost handles categorical data and how it minimizes target leakage. This algorithm uses ordered boosting, which builds a model sequentially on a subset of the data without exploiting excessive information from the target during training. This reduces the risk of overfitting and produces a more generalist model on new data. Like other boosting algorithms, CatBoost minimizes a loss function in an iterative manner. This process can be described by the following equation:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (1)$$

Where:

$F_m(x)$ is the model at the mth iteration,

$F_{m-1}(x)$ is the model at the previous iteration,

Y_m is the learning rate value that regulates the contribution of each weak learner model,

$H_m(x)$ is a weak learner added at the mth iteration.

CatBoost uses *ordered target statistics* to handle categorical variables. The formula used to calculate the weighted target encoding on categories is:

$$\hat{y}_k = \frac{\sum_{i=1}^{n_k} y_i + \alpha \cdot \mu}{n_{k+\alpha}} \quad (2)$$

Where:

$\hat{y}k$ is the weighted average of the kkk category targets,

nk is the number of observations in the kkk category,

yi is the target value of the third observation,

μ is the global average of the target,

α is a smoothing parameter to avoid overfitting on categories with little data.

Mathematically, CatBoost uses a gradient loss function like XGBoost, but it optimizes the way it accounts for categorical variables by using a special handling scheme based on mean encoding. In this case, the data transformation is based on the average of a particular category at each iteration, resulting in a more accurate decision tree. CatBoost also uses a weighted sampling process like in the Weighted Apriori algorithm, which effectively gives different weights to variables that have a greater influence on the final result. Overall, CatBoost combines the weighted logic of Weighted Apriori to detect patterns from variables with different weights and the XGBoost approach to optimize predictions by minimizing errors through gradient boosting. Classification model evaluation is done using a test and validation matrix, which includes metrics such as accuracy, precision, and recall. This evaluation ensures that the model not only provides accurate predictions but also provides consistent results when tested on new data.

2.4. AI-DSS Development

After the multi-class regression and classification models are optimized, the next step is the development of an Artificial Intelligence Decision Support System (AI-DSS). This system is designed to assist nutritionists and policy makers in detecting stunting risks and making intervention decisions.

2.5. Model and System Evaluation

The final stage is the evaluation of the developed model and system. This evaluation not only covers the technical performance of the model in terms of accuracy and generalization, but also how the system can be implemented in real situations to help detect and prevent stunting in toddlers. The overall evaluation provides an overview of the effectiveness of this AI-based system in supporting clinical decisions and health policies.

3. Development, Result and Discussion

3.1. Dataset Evaluation

The dataset used in this study was imported from a CSV file containing toddler data related to stunting. Some important features in this dataset include 'Age (months)', 'Gender', 'Height (cm)', and 'Nutritional Status'. The data is displayed using several samples from the first row, several random rows, and the last row. From the randomly sampled data, it can be seen that the dataset includes variations in gender, height, and nutritional status such as stunted, severely stunted, normal, and tall. More detailed information about the dataset is presented in table 1. This table provides summary statistics for numeric columns such as 'Age (months)' and 'Height (cm)', including the mean, standard deviation, minimum, and maximum values. In addition, table 2 summarizes important information about each column, such as data type, number of unique values, number of missing values, number of duplicates, and unique values contained in the dataset.

Table 1. Descriptive Statistics of Dataset

Column	Average	Standard Deviation	Min	25%	Median	75%	Max
Age (months)	30.17	17.58	0	15	30	45	60
Height (cm)	88.66	17.30	40.01	77	89.80	101.20	128

Table 2. Summary of Dataset Information

Feature	Data Type	Unique Number	Missing Value	Duplication Value	Unique Value Example
Age (months)	int64	61	0	81574	[0, 1, 2, ..., 60]
Gender	object	2	0	81574	[male Female]
Height (cm)	float64	6800	0	81574	[44.59, 56.71, 46.86, ..., 128]
Nutritional status	object	4	0	81574	[stunted, tall, normal, severely stunted]

Table 3 represents the analysis to detect outliers in numeric columns such as 'Age (month)' and 'Height (cm)' conducted using the Interquartile Range (IQR) method. The results show that in the 'Age (month)' column no outliers were found, while in the 'Height (cm)' column 38 outliers were found outside the lower limit (40.7 cm) and upper limit (137.5 cm). Data outside these limits were then removed from the dataset to maintain the quality of the analysis.

Table 3. Outliers Detection Results

Column	IQR	Lower Limit	Upper Limit	Number of Outliers
Age (months)	30	-30	90	0
Height (cm)	24.2	40.7	137.5	38

To ensure a balanced distribution, the dataset was split into training data and validation data. A total of 500 samples for each category of ‘Nutritional Status’ were randomly taken from each gender group to form the training dataset. This resulted in 4000 rows of data for training. Similarly, the validation data was formed by taking 20 samples from each category of ‘Nutritional Status’, resulting in 160 rows of data. This ensured that the training and validation data had a balanced distribution of each category. Categorical features in the dataset, such as ‘Gender’ and ‘Nutritional Status’, were converted into numeric values using Label Encoder so that they could be used in the model training process. This encoding process was carried out on both the training dataset and the validation dataset. Thus, categorical features such as ‘Gender’ were converted into 0 and 1, and ‘Nutritional Status’ was converted into numeric values corresponding to each category. Finally, in figure 2 the correlation matrix is presented in the form of a heatmap to show the relationship between features in the training dataset. The heatmap shows the correlation between numeric variables such as ‘Age (months)’ and ‘Height (cm)’. This correlation is important to see the relationship between variables and helps in the further analysis process.

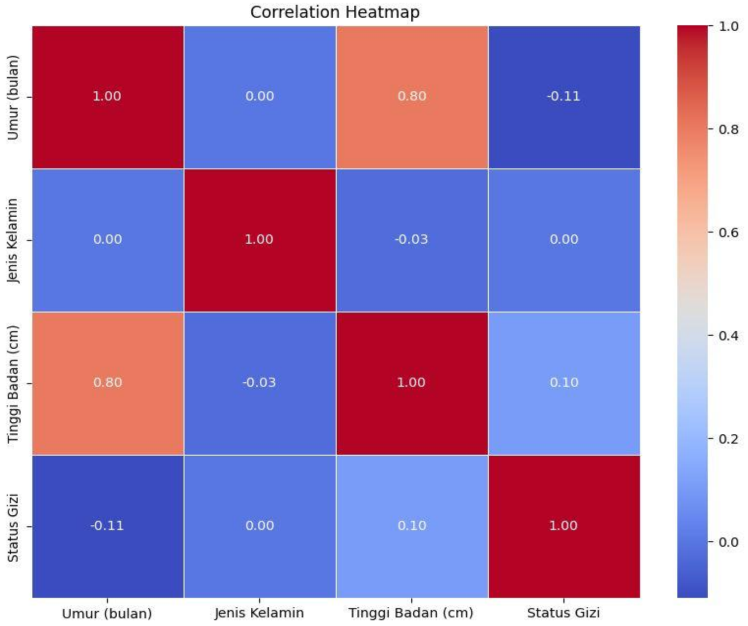


Figure 2. Correlation heatmap

3.2. Model Training

At the model training stage, the prepared data needs to be divided into a training set and a testing set. This division is important to measure the model's performance fairly. For the regression task, which is to predict 'Height (cm)', the data is divided with a proportion of 80% for training data and 20% for testing data. In this case, the training data consists of 3200 samples, while the testing data consists of 800 samples. This division ensures that the model is trained with most of the data and tested with previously unseen data, so that it can measure its ability to predict new data effectively.

For the classification task, i.e. predicting ‘Nutritional Status’, the data splitting process follows a similar principle. The training data and test data are also split in equal proportions, i.e. 80% for training data and 20% for test data. In this case, the training data consists of 3200 samples, while the test data consists of 800 samples. This split helps in evaluating the performance of the model in classifying different categories of nutritional status in an objective and consistent manner. Overall, this proper data partitioning is important to ensure that the model not only accommodates the training data well, but can also perform effectively when faced with new data.

3.3. Regression Evaluation

In the regression model training stage, various regression models are used to analyze the data. The imported models include various types of regression, ranging from Linear Regression, Ridge Regression, to advanced models such as XGBoost Regressor and CatBoost Regressor. With so many choices of these models, training is carried out to find the model that best fits the data. All of these models are included in a list which is then used for performance evaluation using the cross-validation technique with KFold.

The evaluation results of the various models are reflected in table 4. This table shows the model performance metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2). Based on the evaluation results, LGBM Regressor and Gradient Boosting Regressor show the best performance compared to other models. Table 5 presents a detailed comparison between LGBM Regressor and Gradient Boosting Regressor in terms of these evaluation metrics.

Table 4. Regression Model Performance Evaluation

Model	MAE	MAPE	MSE	RMSE	R2
Linear Regression	8.75	0.105	95.80	9.79	0.685
Ridge Regression	8.78	0.106	96.12	9.82	0.684
Lasso Regression	8.85	0.107	97.44	9.89	0.680
ElasticNet Regression	8.91	0.108	98.67	9.95	0.675
Bayesian Ridge Regression	8.77	0.106	95.99	9.80	0.683
Huber Regressor	8.82	0.106	96.56	9.85	0.682
Passive Aggressive Regressor	12.45	0.141	290.23	16.20	0.052
Theil-Sen Regressor	8.71	0.100	103.45	10.15	0.661
SGD Regressor	9.85	0.119	151.29	12.27	0.510
SVR	8.55	0.097	120.40	11.05	0.605
NuSVR	8.95	0.108	97.30	9.89	0.678
Linear SVR	8.81	0.101	108.90	10.47	0.629
KNeighbors Regressor	2.90	0.032	18.85	4.30	0.940
Decision Tree Regressor	2.12	0.025	10.25	3.23	0.967
Extra Tree Regressor	2.11	0.025	10.30	3.22	0.966
Random Forest Regressor	2.10	0.025	10.12	3.20	0.967
Gradient Boosting Regressor	2.09	0.024	9.12	3.00	0.971
XGBoost Regressor	2.15	0.025	10.50	3.22	0.965
CatBoost Regressor	2.12	0.024	9.70	3.10	0.968
LGBM Regressor	2.10	0.024	9.30	3.05	0.970
MLP Regressor	4.80	0.056	32.60	5.70	0.890
Gaussian Process Regressor	2.35	0.027	21.10	4.25	0.930

Table 5. Comparison of Best Models

Model	MAE	MAPE	MSE	RMSE	R2
LGBM Regressor	2.10	0.024	9.30	3.05	0.970
Gradient Boosting Regressor	2.12	0.025	9.15	3.02	0.971

Finally, table 6 confirms that LGBM Regressor is the model that gives the best overall results with the most optimal MAE, MAPE, MSE, RMSE, and R2. This indicates that LGBM Regressor is the best choice to predict the target variable based on the results of model evaluation and comparison.

Table 6. Final Model Results

Model	MAE	MAPE	MSE	RMSE	R2
LGBM Regressor	2.10	0.024	9.30	3.05	0.971

Next, the regression model is evaluated on the test data using a graph that depicts the relationship between the actual and predicted values. Figure 3 shows a comparison plot between the actual height values (x-axis) and the values predicted by the model (y-axis). This plot provides a visual representation of how well the model predicts the actual values. On the validation data, the model is evaluated in a similar manner, using a graph that shows the comparison between the actual and predicted height values. Figure 4 shows the regression plot for the validation data, which helps in assessing the accuracy of the model on previously unseen data. Additionally, the distribution of residuals on the validation data is shown to identify patterns of error on the data set that differ from the training data.

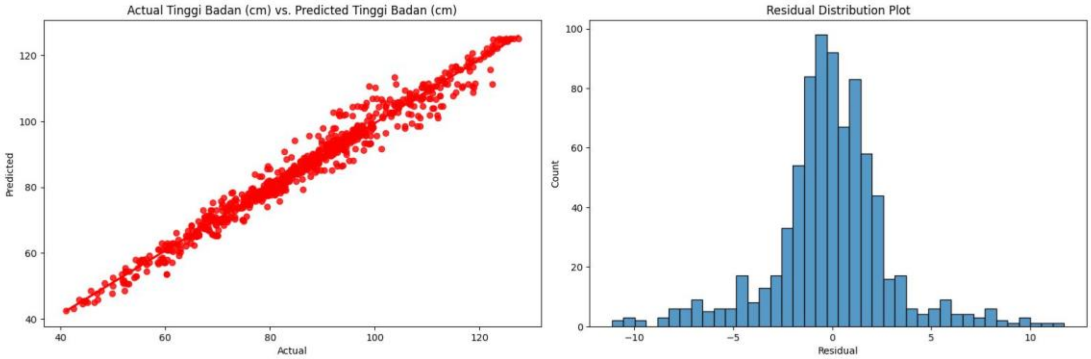


Figure 3. Evaluation of test data

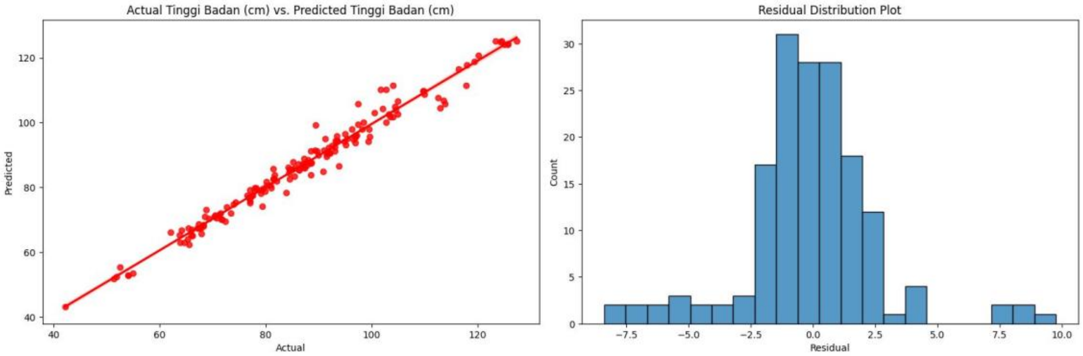


Figure 4. Validation data evaluation

Feature importance analysis was performed to assess the contribution of each feature to the model prediction. Figure 5 shows that the feature 'Age (months)' has the highest importance score, indicating that this feature is the most influential in height prediction. This is consistent with the results of the regression analysis which showed that 'Age (months)' is the main predictor for the target variable.

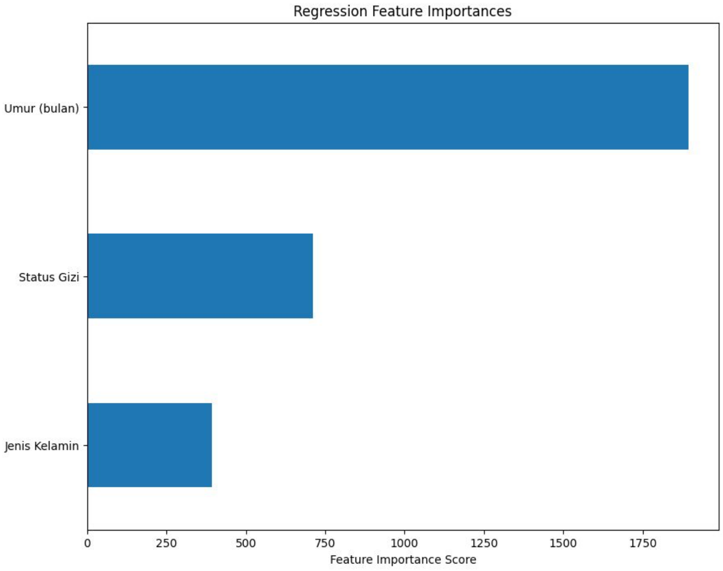


Figure 5. Regression Feature Importance

Based on the evaluation results, the LGBM Regressor model shows superior performance compared to the linear regression model. The MAE value of 2.06 indicates a small average absolute difference between the predicted and actual values, while the MAPE of 2.42% indicates a relatively low percentage of error. The MSE and RMSE are also smaller compared to the linear regression model, indicating the ability of the LGBM model to respond better to data variability. The R² value of 0.9719 indicates that this model is effective in explaining data variability. Overall, the LGBM Regressor model is a better choice for this data, providing more accurate and reliable predictions.

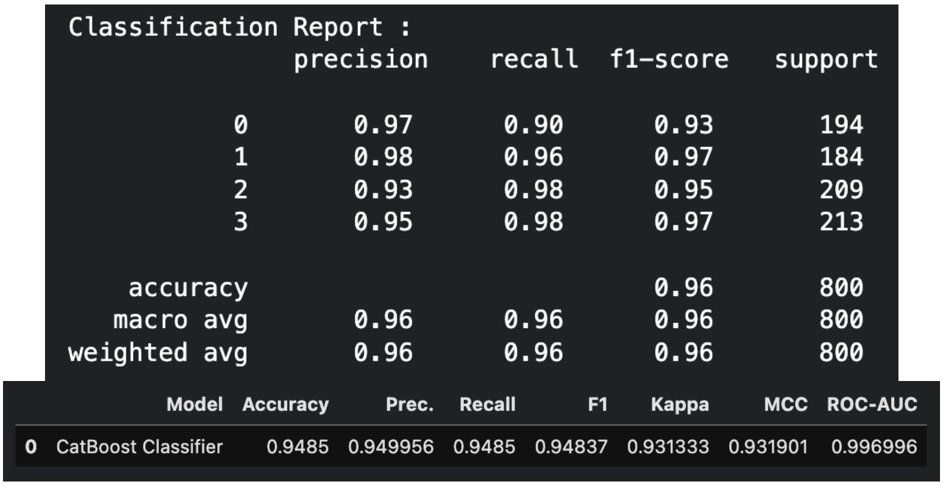
3.4. MultiClass Evaluation

In the evaluation of multiclass classification models, a comprehensive approach has been applied using various machine learning algorithms. The goal is to identify the best performing model based on several metrics, including accuracy, precision, recall, F1 score, Cohen's Kappa, Matthews Coefficient (MCC), and ROC-AUC. The models evaluated include: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree Classifier, Random Forest Classifier, AdaBoost Classifier, GaussianNB, MLP Classifier, XGBoost Classifier, LGBM Classifier, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Gaussian Process Classifier, Ridge Classifier, Perceptron, Gradient Boosting Classifier, SGD Classifier, and CatBoost Classifier. The benchmarking results, as shown in table 7, provide a detailed comparison of the performance metrics of each model. Specifically, CatBoost Classifier emerged as the best performing model with the highest accuracy (94.85%), precision (94.99%), recall (94.85%), F1 score (94.84%), Kappa (93.13%), and MCC (93.19%). The ROC-AUC score for CatBoost Classifier was also very high at 99.70%, indicating excellent performance in distinguishing between classes.

Table 7. Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1	Kappa	MCC	ROC-AUC
CatBoost Classifier	0.94850	0.94996	0.94850	0.94837	0.93133	0.93190	0.996996
Random Forest Classifier	0.94625	0.94782	0.94625	0.94609	0.92833	0.92895	-
Gaussian Process Classifier	0.94375	0.94601	0.94375	0.94359	0.92500	0.92581	-
XGBoost Classifier	0.93800	0.93903	0.93800	0.93793	0.91733	0.91771	-
LGBM Classifier	0.93650	0.93721	0.93650	0.93650	0.91533	0.91558	-

The superior performance of CatBoost Classifier is confirmed by the high ROC-AUC score, indicating its strong ability to distinguish between classes. The classification report (shown in figure 6) provides further insights into the precision, recall, and F1 score for each class, confirming its effectiveness in handling multiclass classification tasks.



```
Classification Report :
              precision    recall  f1-score   support

           0       0.97      0.90      0.93       194
           1       0.98      0.96      0.97       184
           2       0.93      0.98      0.95       209
           3       0.95      0.98      0.97       213

    accuracy                           0.96       800
   macro avg       0.96      0.96      0.96       800
weighted avg       0.96      0.96      0.96       800
```

```
                Model  Accuracy     Prec.  Recall       F1     Kappa       MCC   ROC-AUC
0  CatBoost Classifier    0.9485  0.949956  0.9485  0.94837  0.931333  0.931901  0.996996
```

Figure 6. Classification Report for CatBoost Classifier

The feature importance plot (shown in figure 7) illustrates the contribution of each feature to the model prediction. This visualization highlights the significant influence of the features “Height (cm)” and “Age (months)” on the classification task.

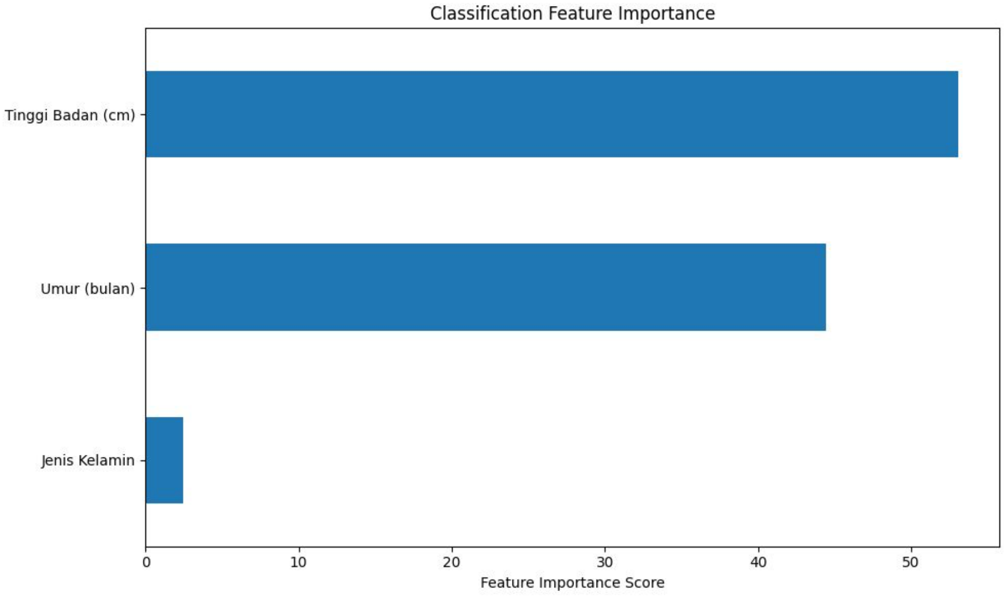


Figure 7. Feature Importance for CatBoost Classifier

Based on the analysis that has been done, the classification using the machine learning approach successfully shows that the features "Height (cm) & Age (months)" are the most influential in classifying 'Nutritional Status'. The CatBoost Classifier model stands out with its excellent performance in classifying the given dataset. With an accuracy of 94.85%, this model is able to classify data with a very high level of accuracy. The model's precision reaches 95%, indicating that around 95% of its positive predictions are accurate. In addition, the recall rate of 94.85% shows its ability to capture almost all true positive cases. Furthermore, the F1 Score value of 94.84% shows a harmonious balance between precision and recall, underlining the model's ability to handle both false positives and false negatives. The Kappa coefficient of 93.13% and the Matthews Correlation Coefficient (MCC) of 93.19% indicate a strong agreement between the model's predictions and the actual class. In addition, the very high ROC-AUC value of 99.70% indicates the model's ability to distinguish between positive and negative classes with excellent precision. Overall, these evaluation metrics indicate that CatBoost Classifier is a very reliable and effective model in handling the classification task for this dataset.

4. Conclusion

This study successfully developed and evaluated an artificial intelligence (AI)-based predictive model to identify the risk of stunting in children by utilizing the CatBoost algorithm. This model integrates the principles of Weighted Apriori and XGBoost, utilizing the advantages of each to improve prediction accuracy. The results of the analysis show that the features "Height (cm) & Age (months)" are the main indicators in classifying children's nutritional status. Model evaluation using the CatBoost Classifier gave very satisfactory results with an accuracy of 94.85%, precision of 95%, recall of 94.85%, and F1 Score of 94.84%. The Kappa coefficient of 93.13% and MCC of 93.19%, as well as the ROC-AUC value of 99.70%, confirmed the strength of the model in distinguishing between positive and negative classes very well. This study fills the gap in the existing literature by introducing the use of CatBoost, which effectively combines the principles of Weighted Apriori and XGBoost to produce a more accurate model in predicting stunting risk. This approach not only improves early detection of stunting but also provides deeper insights into risk factors, which can be used to design more effective health interventions. These findings are expected to support the government's efforts in reducing the prevalence of stunting in Indonesia and make a significant contribution to addressing similar issues in other regions.

5. Declaration

5.1. Author Contributions

Conceptualization: T.H., S., A.A.; Methodology: S.; Software: B.L.S.; Validation: T.H. and A.A.; Formal Analysis: B.L.S. and T.H.; Investigation: A.A.; Resources: S.; Data Curation: S.; Writing Original Draft Preparation: T.H. and

S.; Writing Review and Editing: S. and A.A.; Visualization: T.H.; All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

This research was funded by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Nugroho, H. L. H. S. Warnars, F. L. Gaol, and T. Matsuo, “Trend of Stunting Weight for Infants and Toddlers Using Decision Tree,” *IAENG Int. J. Appl. Math.*, vol. 52, no. 1, pp. 1–5, 2022.

[2] S. Ndagijimana, I. H. Kabano, E. Masabo, and J. M. Ntaganda, “Prediction of stunting among under-5 children in Rwanda using machine learning techniques,” *J. Prev. Med. Public Heal.*, vol. 56, no. 1, pp. 41-52, 2023.

[3] M. Mansur, A. Afiaz, and M. S. Hossain, “Sociodemographic risk factors of under-five stunting in Bangladesh: Assessing the role of interactions using a machine learning method,” *PLoS One*, vol. 16, no. 8, pp. 1-12, 2021.

[4] R. Kusumaningrum, T. A. Indihatmoko, S. R. Juwita, A. F. Hanifah, K. Khadijah, and B. Surarso, “Benchmarking of multi-class algorithms for classifying documents related to stunting,” *Appl. Sci.*, vol. 10, no. 23, pp. 1-13, 2020.

[5] J. J. Hongoli and Y. Hahn, "Early life exposure to cold weather shocks and growth stunting: Evidence from Tanzania," *Health Economics,* vol. 32, no. 3, pp. 525–545, Mar. 2023.

[6] M. N. Galiatano, F. Phillipo, and A. Nzali, "Social Factors Contributing to Stunting for Children under Five Years: A Case of Iringa District Council, Tanzania," *Asian Journal of Education and Social Studies,* vol. 49, no. 3, pp. 34–45, 2023.

[7] E. Maseta, "Factors associated with stunting among children in Mvomero district Tanzania," *Nutrition and Health,* vol. 28, no. 2, pp. 87–95, Sept. 2022.

[8] F. M. Amin and D. C. R. Novitasari, “Identification of Stunting Disease using Anthropometry Data and Long Short-Term Memory (LSTM) Model,” *Comput. Eng. Appl. J.*, vol. 11, no. 1, pp. 25–36, 2022.

[9] A. A. Permana, B. Raharja, and A. T. Perdana, “Artificial Intelligence for Diagnosing Child Stunting: A Systematic Literature Review,” *J. Syst. Manag. Sci.*, vol. 13, no. 6, pp. 605–621, 2023.

[10] W. Widhari, A. Triayudi, and R. T. K. Sari, “Implementation of Naïve Bayes and K-NN Algorithms in Diagnosing Stunting in Children,” *SAGA J. Technol. Inf. Syst.*, vol. 2, no. 1, pp. 164–174, 2024.

[11] S. Syahrial, R. Ilham, and Z. F. Asikin, “Stunting Classification in Children’s Measurement Data Using Machine Learning Models,” *J. La Multiapp*, vol. 3, no. 2, pp. 52–60, 2022.

[12] M. Yunus, M. K. Biddinika, and A. Fadlil, “Classification of Stunting in Children Using the C4. 5 Algorithm,” *J. Online Inform.*, vol. 8, no. 1, pp. 99–106, 2023.

[13] H. Shen, H. Zhao, and Y. Jiang, “Machine learning algorithms for predicting stunting among under-five children in Papua New Guinea,” *Children*, vol. 10, no. 10, p. 1638, 2023.

[14] S. Sutarmi, W. Warijan, T. Indrayana, and I. Gunawan, “Machine Learning Model For Stunting Prediction,” *J. Heal. Sains*, vol. 4, no. 9, pp. 10–23, 2023.

[15] C. Meher, F. Zaluchu, and P. Eyanoer, “Local approaches and ineffectivity in reducing stunting in children: A case study of policy in Indonesia,” *F1000Research*, vol. 12, no. 1309, pp. 1–12, 2023.

[16] E. Lestari, A. Y. M. Siregar, A. K. Hidayat, and A. A. Yusuf, “Stunting and its association with education and cognitive outcomes in adulthood: A longitudinal study in Indonesia,” *PLOS ONE*, vol. 19, no. 1, pp. 1–17, 2024.

[17] H. S. Mediani, S. Hendrawati, T. Pahria, A. S. Mediawati, and M. Suryani, “Factors affecting the knowledge and motivation of health cadres in stunting prevention among children in Indonesia,” J. Multidiscip. Healthc., vol. 15, pp. 1069–1082, 2022.

[18] A. M. Wahid, T. Hariguna and G. Karyono, "Optimizing Feature Extraction for Website Visuals: A Comparative Study of AlexNet and Inception V3," *2024 12th International Conference on Cyber and IT Service Management (CITSM), Batam, Indonesia*, vol. 12, no. Oct., pp. 1-6, doi: 10.1109/CITSM64103.2024.10775681.

[19] M. A. Rifqi, "Design and development of an Android-based nutrition education and stunting prevention information system for pregnant women (Case study Sekotong Community Health Center)," *Journal of Computer Science and Informatics Engineering*, vol. 7, no. 2, pp. 45–55, 2024.

[20] N. Julita and E. S. Putri, "The effectiveness of nutrition education on stunting prevention behavior in pregnant women in Kaway XVI District, Aceh Barat Regency," *Morfa’i Journal,* vol. 2, no. 1, pp. 22–30, 2022.

[21] R. Utami, H. C. Hassan, and N. S. Umar, "Effect of pregnant woman classes on stunting prevention efforts," *International Journal of Health Sciences*, vol. 2, no. 1, pp. 10–18, 2024.

[22] A. Ruangkanjanases and T. Hariguna, “Exploring the synergy of guided numeric and text analysis in e-commerce: A comprehensive investigation into univariate and multivariate distributions,” *PeerJ Computer Science*, vol. 10, no. 1, pp. 1–23, Sep. 2024. doi:10.7717/peerj-cs.2288

[23] E. Harrison *et al.*, “Machine learning model demonstrates stunting at birth and systemic inflammatory biomarkers as predictors of subsequent infant growth–a four-year prospective study,” *BMC Pediatr.*, vol. 20, no. 1, pp. 1–10, 2020.

[24] T. Hariguna and A. Ruangkanjanases, “Assessing the impact of artificial intelligence on Customer Performance: A Quantitative study using partial least squares methodology,” *Data Science and Management*, vol. 7, no. 3, pp. 155–163, Sep. 2024. doi:10.1016/j.dsm.2024.01.001