ISSN 2462-2052 | eISSN 2600-8718
DOI: https://doi.org/10.37134/jsml.vol13.1.12.2025

Research Article

Essentials of Preprocessing Data in Improving Logistic Regression Performance Based on Rough Sets Theory: A Case Study of Stunting in West Sumatra, Indonesia

Izzati Rahmi[1,2], Riswan Efendi[1,3*], Nor Azah Samat[1], Ferra Yanuar[2], Hazmira Yozza[2], Muhammad Wahyudi[4], S.M.A. Burney[5] and Erol Eğrioğlu[6]

[1]Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, 35900 Tanjung Malim, Perak, Malaysia
[2]Department of Mathematics and Data Science, Faculty of Mathematics and Natural Science, Andalas University, 25163 Padang, Indonesia
[3]Department of Mathematics, Faculty of Science and Technology, UIN Sultan Syarif Kasim Riau, 28293, Pekanbaru, Indonesia
[4]Department of Information Technology Politeknik Caltex Riau, 28265 Pekanbaru, Indonesia
[5]Department of Computer Science, Faculty of Science, University of Karachi, 75190 Karachi, Pakistan
[6]Department of Statistics, Faculty of Arts and Sciences, Giresun University, 28610 Giresun, Turkey
*Corresponding author: riswanefendi@fsmt.upsi.edu.my

Received: 5 February 2025; **Accepted:** 8 June 2025; **Published:** 12 June 2025

ABSTRACT

Various studies have considered a logistic regression model for investigating stunting and its determinants. However, some models above fall short of the researcher's expectations, such as the fact that there is no significant variable, making them very challenging to interpret. In this paper, we are interested in handling the problem above by applying data reduction strategies using rough set theory in the preprocessing phase and implementing them for stunting data sets in Solok Regency, West Sumatra Province, Indonesia. There are three types of data reduction: removing inconsistent samples, irrelevant attributes, or both, so three kinds of modified models will be used in this paper, namely Logistic Regression Reduction Rough Set (LR3S) type I, II, and III. The classic and modified models were compared using performance model criteria, which include precision, recall, F1-score, accuracy, ROC curves, and AUC values. It was determined that the stunting data set was unsuitable for the classical logistic model, and the best model was built by removing inconsistent observations (LR3S type I). At a 5% significance level, the best model indicates the factors significantly influencing stunting incidence: exclusive breastfeeding, birth weight, smoking family, immunization, and gender. Stunting classes are not considerably differentiated by factors such as worms, clean water, comorbidities, hygienic latrines, and health assurance in Solok Regency. A priority program and policies for stunting prevention in this regency can be developed by considering the extent of these factors' major influence.

Keywords: logistic regression; the risk factor of stunting; rough set theory; data reduction; inconsistent samples; irrelevant attributes

1. INTRODUCTION

Stunting is a chronic malnutrition issue caused by insufficient nutritional intake over a prolonged period due to the availability of food that does not satisfy dietary needs. Stunting can occur from the fetal stage and only becomes apparent when the child is two years old. This condition is presented with a height-for-age z-score less than -2 standard deviations (SD) based on the WHO growth standards (WHO, 2010). According to the WHO guidelines, stunting prevalence in each area must be reduced to less than 20% (Kemenkes, 2022). Reducing stunting prevalence became the objective of the Sustainable Development Goal's key indicator number 2.2 (United Nations, 2022) and the Global Nutrition Target for 2025 (WHO, 2024). According to Presidential Decree No. 72/2021 on the acceleration of stunting reduction, the Indonesian government pledged to cut the prevalence of stunting by 14% by 2024. Meanwhile, the Ministry of Health of the Republic of Indonesia performed the Indonesian Nutrition Status Survey (SSGI) and discovered that 21.6% of children in Indonesia suffered from stunting in 2022 (Kemenkes, 2024). Thus, the acceleration of reducing stunting prevalence has become the Indonesian government's priority program as outlined in the 2000-2024 development.

One strategy to reduce the stunting prevalence is to investigate the risk factors. Some previous studies in Indonesia reported that stunting incidence in toddlers was associated with factors such as birth weight (Katoch, 2022; Wijayanti & Nurseskasatkmata, 2022; Podungge et al., 2021; Putri et al., 2021); gender (Supadmi et al., 2024; Sihotang et al., 2023); exclusive breastfeeding (Andika et al., 2024; Sihotang et al., 2023; Podungge et al., 2021; Resiyanthi & Yanti, 2021); sanitation and hygiene environmental factors (Ernawati et al., 2024; Adi et al., 2023; Badriyah & Syafiq, 2017); worms (Adrizain et al., 2024; Benvenuto et al., 2022; Paun et al., 2021); immunization (Permatasari et al., 2023; Brahima et al., 2020), and smoke exposure (Saenong et al., 2024; Muchlis et al., 2023; Panggabean et al., 2023). Research on stunting risk factors, particularly in toddlers, can help the government create policies and programs to accelerate the decrease in stunting.

One way to analyze the factors affecting the occurrence of stunting is through machine learning modeling. Any model represents a few characteristics of a real issue in a simplified manner (Aziz, 2010). However, using ML models sometimes does not yield the expected results, making them difficult to interpret. One of the common issues in ML modeling is that the model has a reasonably high accuracy, but only a few or none of the variables are significant. Some research studies in ML models found that the original dataset problems cause issues in various models. Thus, a preprocessing stage is required before performing the ML model (Osisanwo et al., 2017). It has been proven that preprocessing the data can improve the ML model's performance (Rahhali et al., 2024; Sami et al., 2021; Iliou et al., 2015). Many approaches can be used in the data preprocessing process, one of which is by using rough set theories (RST) strategies to perform data reduction before further modeling. RST data reduction in ML modeling has been demonstrated in several prior research studies to enhance model performance through the variable selection process (Bhukya & Manchala, 2023; Kaka-Khan et al., 2022) and observation selection (Efendi et al., 2019, 2024; Rahmi et al., 2024a, 2024b; Rasyidah et al., 2023).

Literature studies reveal that some researchers applied RST data reduction before constructing the logistic regression (LR) model. Most of the previous research related to the integration of RST and logistic regression focused on applying attribute selection to reduce data because it is assumed that some factors are irrelevant or not important to the dependent (Li, 2014; Liu et al., 2014; Kim et al., 2008). However, some studies in LR use variables that have been supervised. In other words, they have followed the existing mechanisms in the system, so there is no reason to eliminate these variables in the model. As a result, alternatives are required for data reduction, such as removing inconsistent observations (Rahmi et al., 2024a, 2024b).

According to some research about the examination of stunting risk factors, the LR model is one of the machine learning methods that researchers have utilized extensively (Permatasari et al., 2023; Bhowmik & Das, 2019; Rakhmahayu et al., 2019; Ohyfer et al., 2017). The extent to which the observed variables impact the risk of stunting can be ascertained by LR modeling. This paper applied the RST approach in data preprocessing before building the LR model for the stunting data. Three types of data reduction were used. The first type is reducing data by eliminating inconsistent samples initiated by Efendi et al. (2019; 2024) and modified by Rahmi et al. (2024a, 2024b). The second one is the reduction variable, which is a common RST strategy in reduction data. The last type is data reduction, which removes irrelevant variables and inconsistent samples. The LR model results for three data reduction types were compared to the ordinary LR. The primary objective of this study is to apply the new RST strategy for data reduction in preprocessing data to enhance the LR model's performance in analyzing the stunting risk factor. The best model was then analyzed to recommend a policy to lower the prevalence of stunting.

2. MATERIALS AND METHODS

2.1. Data

Solok Regency is one of the districts in West Sumatra Province with a high stunting prevalence rate of 24.2% in 2023. The secondary data was collected from the Solok Regency Health Service in 2023. In this study, there are 10 conditional attributes (independent variables) as presented in Table 1.

Table 1. List of conditional attributes (independent variables) in stunting data

Conditional Attribute (Independent Variable)	Categories
Sex (X_1)	Female (0); Male (1)
Birth Weight (X_2)	Low (0); Normal (1)
Exclusive Breastfeeding (X_3)	No (0); Yes (1)
Healthy Latrines (X_4)	No (0); Yes (1)
Clean Water (X_5)	No (0); Yes (1)
Health Insurance (X_6)	No (0); Yes (1)
Worms (X_7)	No (0); Yes (1)
Immunization (X_8)	No (0); Yes (1)
Smoking Family (X_9)	No (0); Yes (1)
Comorbidities (X_{10})	No (0); Yes (1)

A decision attribute (dependent variable) is the stunting status of children under five with two categories, namely "stunting" and "severe stunting". This category refers to the WHO growth standard deviation table. A child is considered stunted if the ($-3 < z$-score $\leqslant -2$) and severely stunted if (z-score <-3).

2.2. The Proposed Modified Data Reduction Using RST

In this study, three types of modified data reduction are proposed using RST before modelling the LR model. Overall, the procedure of data reduction using RST is provided in Figure 1.

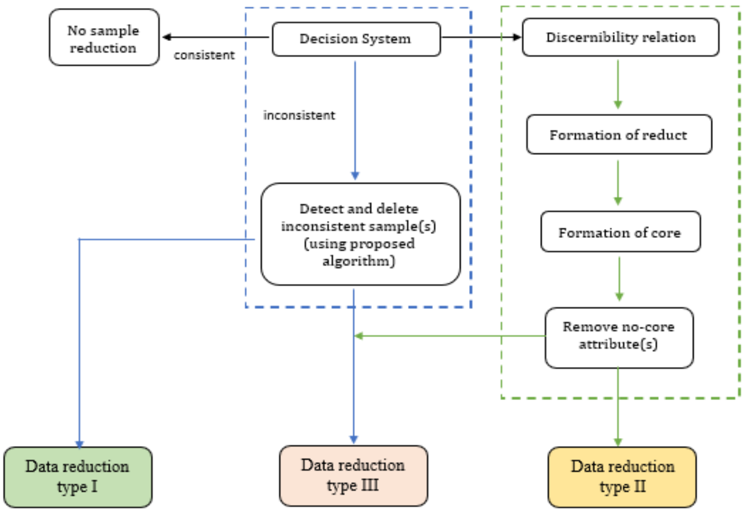

Figure 1. Three types of data reduction: type I (removing inconsistent samples), type II (removing irrelevant attributes), and type III (removing inconsistent samples and irrelevant attributes)

The RST Procedure for Data Reduction Type I - As initiated by Rahmi et al. (2024a, 2024b), the proposed algorithm is detailed step by step in this paper. This procedure has been modified based on the previous works by Efendi et al. (2024; 2019) and Rasyidah et al. (2023) by adding the probability approach. The following is the detailed procedure of this algorithm.

Suppose D is a binary decision attribute, $C_1, C_2, \ldots, C_k$ are k condition attributes, and n samples are $P_1, P_2, \ldots, P_n$, the decision system (DS) can be written as:

$$DS = (U, F) = \left(U, F = C \cup D, \{V_f | f \in F\}, I_f | f \in F\right) \qquad \text{Eq. 1}$$

From Eq. (1), $U = (P_1, P_2, \ldots P_n)$, $C = (C_1 \ldots, C_k)$, D = binary dependent attribute (Yes/No), V_f is a set of values $f \in F$, and $I_f : U \rightarrow V_f$ is a function that maps samples from U to a single value in V_f.

The proposed algorithm related to sample reduction is as follows:

a. Verify the decision table is inconsistent.
   To verify the decision table is inconsistent, for $i \neq j$, show that:
$$\exists (P_i, P_j) \in UxU \left[\forall c \in C \middle| I_c(P_i) = I_c(P_j) \wedge I_d(P_j) \neq I_d(P_j)\right] \qquad \text{Eq. 2}$$
b. Define the equivalence class (EC) of DS based on condition attributes.
$$EC = \{(P_i, P_j) \in U \times U | \forall c \in C, I_c(P_i) = I_c(P_j)\} \qquad \text{Eq. 3}$$
c. For every equivalence class, according to condition attributes, define the approximation of a set using the discernibility relation.
   Let DS be a decision system, $X \subseteq U$. The estimate of a set X is verifiable with the available information in the set of condition attributes C by forming the *C-lower* and *C-upper approximations of X*, which are signified by $\underline{C}$X and $\overline{C}$X, respectively, in which:
$$\underline{C}X = \{x \in U \mid [x]_C \subseteq X\}, \qquad \text{Eq. 4}$$
$$\overline{C}X = \{x \in U \mid [x]_C \cap X \neq \phi\}. \qquad \text{Eq. 5}$$
   C-boundary region of X: $BR_C(X) = \overline{C}X - \underline{C}X$ Eq. 6
d. Divide DS into the following subsets:
$$DS = (S_C, S_{IC(1)}, S_{IC(2)}, \ldots, S_{IC(j)});$$
   where $S_C = \cup \{\{x \in EC_i\} \mid BR_{EC(i)}(X) = \phi\}$, *and*
$$S_{IC(j)} = \{\{x \in EC_j\} \mid BR_{EC(j)}(X) \neq \phi\}.$$

e. For every $x \in S_{IC(j)}$, count $P_{1j} = \frac{n(Y)}{n\,(S_{IC(j)})}$ dan $P_{2j} = \frac{n(Y')}{n\,(S_{IC(j)})}$;
   where $Y = \{x \in S_{IC(j)} \mid I_d(x) = \text{Yes}\}$ and $Y' = \{x \in S_{IC(j)} \mid I_d(x) = \text{No}\}$.
f. Based on step (e):
   If $P_{1j} = P_{2j}$ then remove $\forall x \in S_{IC(j)}$.
   If $P_{1j} < P_{2j}$, remove $\forall x \in Y$.
   If $P_{1j} > P_{2j}$, remove $\forall x \in Y'$.
g. Based on step (f), a consistent decision system is produced after removing specific objects.

The Proposed Data Reduction Type II - In this study, attribute reduction was carried out by applying the concepts of reduce and core, as explained in some literature (Cao, 2021; Pessoa & Stephany, 2014; Pawlak, 1998). Many researchers have applied this procedure (Li, 2014; Liu et al., 2014; Kim et al., 2008; Yaile et al., 2007). After the reductions are obtained in the RST procedure, the proposed algorithm related to attribute reduction is outlined as follows:

a. Determine the core for each class equivalent based on the reduction values obtained.
b. Variables that appear as core are important in the model and are maintained.
c. Remove variables that are not the core of the data set.
d. Data that has been reduced will become a new data set, hereinafter referred to as type II reduced data.

The RST Procedure for Data Reduction Type III - Data reduction type III is obtained by combining the process of sample reduction and attribute reduction. In the first step, the data reduction type I procedure is applied to detect and remove inconsistent samples. After that, the data reduction type II procedure is used to select the relevant attributes.

2.3. Research Method

This paper proposes a novel integrated classification method combining RST and binary logistic regression analysis. It emphasizes how crucial it is for researchers to think about data quality before conducting any additional analysis. The first step is preprocessing data by reducing the data using the RST technique. As mentioned in Section 2.2, data reduction can be divided into three types. The first type of data reduction is obtained by removing samples that are detected as inconsistent. The second one is obtained through attribute selection, where the rough set eliminates irrelevant attributes for the decision variable. Furthermore, removing inconsistent samples and unnecessary attributes yields the third form of data reduction. The next stage is to build several LR models on the original data (without reduction) and the reduced data. The models tested in this study are presented in Table 2.

Table 2. Various model proposed and their characteristics of data reduction

Model Logistic Regression (LR)	Data Reduction
Original LR (OLR)	No data reduction
LR3S type I	Data reduction type I
LR3S type II	Data reduction type II
LR3S type III	Data reduction type III

These models are used on the stunting data set and contrasted with several performance models, including precision, recall, F1-score, accuracy, ROC curves, and AUC values. The overall work in this research is presented as a framework, as seen in Figure 2.

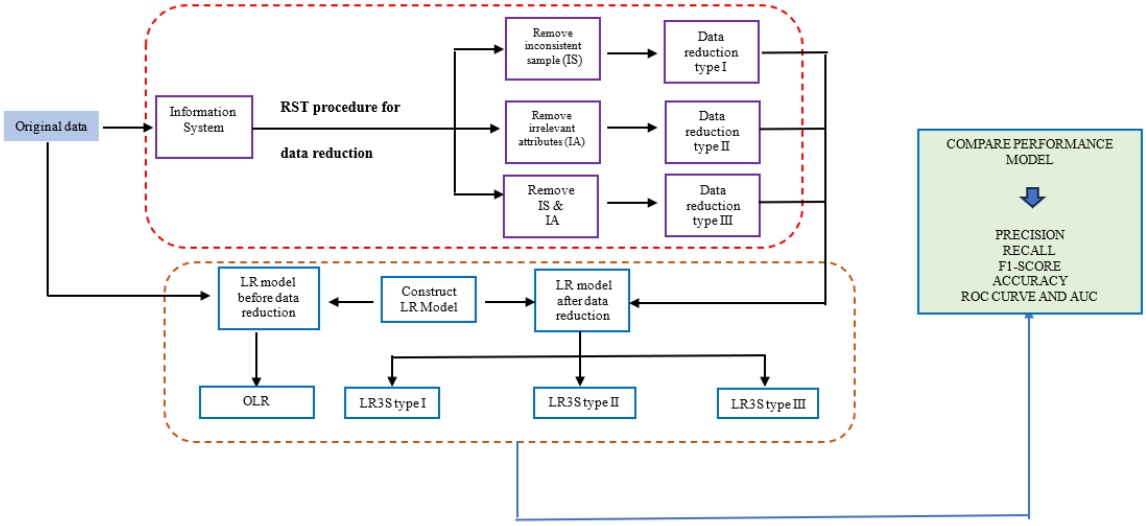

Figure 2. Proposed Research Framework

3. RESULTS AND DISCUSSION

This study implements the classic model LR and three modified models to determine the risk factor of stunting in Solok District, Province of West Sumatra, Indonesia. The first step in the data analysis method is to convert the original data into a decision system with ten condition attributes. The stunting class, which is further subdivided into stunting and severe stunting, is the decision attribute. This study observed 362 toddlers: 256 experienced stunting, and 106 experienced severe stunting. As shown in Table 3, the decision system for the stunting data set has 362 rows and 11 columns.

Table 3. Decision system of data stunting

Toddler	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
T_1	0	1	1	1	1	0	0	1	1	1	0
T_2	1	1	1	1	1	0	0	1	1	0	0
T_3	0	1	1	1	1	0	0	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
T_{362}	1	1	1	0	1	0	0	1	0	0	1

Another way to write the decision system in Table 6 is as follows:

$$DS = (U, F) = \left(U, F = C \cup Y, \{V_f | f \in F\}, I_f | f \in F\right), \qquad \text{Eq. 7}$$

$U = (T_1, T_2, \ldots, T_{362})$
$C = (X_1, X_2, \ldots, X_{10})$
Y = Stunting class

Based on a preliminary diagnostic analysis, it is informed that the decision table of the stunting data set is inconsistent, so all three types of data reduction using RST are applicable.

3.1. OLR Model

Initially, the OLR model, which uses original data (before the reduction process), was employed. There are ten independent variables and 362 observations in the modeled data. Table

 Journal of Science and Mathematics Letters
Volume 13, Issue 1, 124-139, 2025

4 displays the estimated findings of the OLR model coefficients. It is evident from Table 4 that absolutely none of the ten independent variables in the model are significant. In addition, as illustrated in Figure 3(A), the performance model values, including precision, recall, F1-score, and accuracy, are reasonably good. Figure 3(B) shows the ROC curve with an AUC value of 0.58, which falls into the fail category according to the AUC value criteria (Çorbacıo˘glu, 2023). Therefore, it is not appropriate to apply this model.

Table 4. The estimated parameter of the OLR model

Variable	Coefficient (B)	Significance	Exp(B)
Gender (X_1)	0.195	0.418	1.215
Birth Weight (X_2)	-0.505	0.169	0.604
Exclusive Breastfeeding (X_3)	-0.307	0.231	0.735
Healthy Latrines (X_4)	-0.103	0.690	0.902
Clean Water (X_5)	0.342	0.346	1.408
Health Insurance (X_6)	-0.393	0.164	0.675
Worms (X_7)	1.161	0.423	3.192
Immunization (X_8)	-0.059	0.828	0.943
Smoking Family (X_9)	-0.213	0.553	0.808
Comorbidities (X_{10})	0.232	0.716	1.261
Constant	-0.095	0.953	0.910

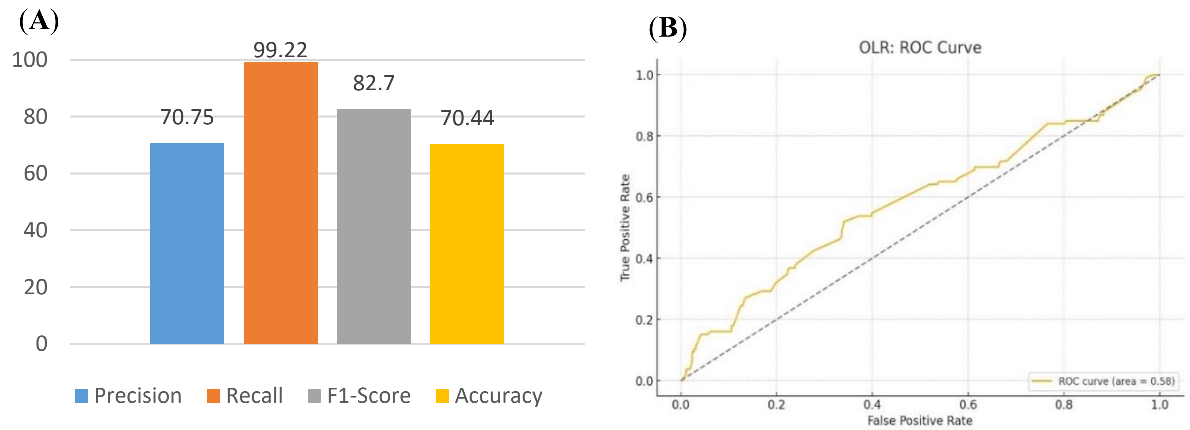

Figure 3. The performance of the OLR model. (**A**) precision, recall, F1-score, and accuracy, (**B**) ROC Curve and AUC

3.2. LR3S Model Type I

The LR3S model type I is constructed for data reduction type I. Firstly, 144 observations were detected as inconsistent samples. Next, by applying the algorithm of data reduction type I, 89 observations were discarded from the data set. Thus, out of the 273 observations in the LR3S type I model, 231 had stunting, and 42 had severe stunting. The results of the LR3S type I model are presented in Table 5. Unlike the OLR model, most of the variables in the LR3S type I model have a significant effect at the 5% test level. As illustrated in Figure 4(A), the precision, recall, F1-score, and accuracy values are also reasonably good. Additionally, the resulting model's AUC value of 0.82 indicates that it is appropriate for use, as shown in Figure 4(B).

Table 5. The estimation of the LR3S type I model coefficients

Variable	Coefficient (B)	Significance	Exp(B)
Gender (X_1)	0.844	0.041	2.328
Birth Weight (X_2)	-1.357	0.005	0.257
Exclusive Breastfeeding (X_3)	-1.564	0.000	0.209
Healthy Latrines (X_4)	-0.747	0.067	0.474
Clean Water (X_5)	0.832	0.175	2.299
Health Insurance (X_6)	0.100	0.816	1.105
Worms (X_7)	2.558	0.089	12.908
Immunization (X_8)	-1.061	0.009	0.346
Smoking Family (X_9)	-1.355	0.006	0.258
Comorbidities (X_{10})	1.277	0.087	3.588
Constant	1.214	0.662	3.367

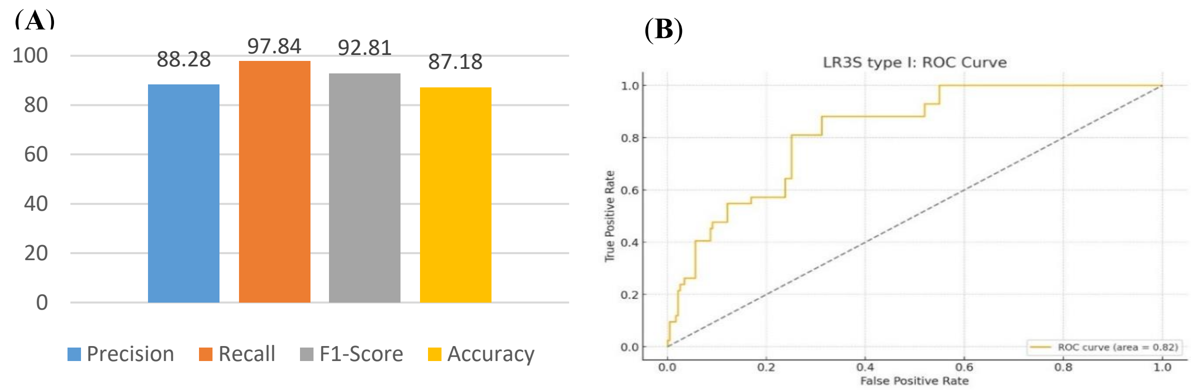

Figure 4. The performance of the LR3S type I model. (**A**) precision, recall, F1-score, and accuracy, (**B**) ROC Curve and AUC

3.3. LR3S Model Type II

The LR3S type II model was developed using a type II reduction technique. Based on the procedure for data reduction type II, it is concluded that two variables have to be removed because they are not included in the decision rule. Consequently, the outcomes of the LR3S Type II consist of 362 observations and eight variables. The results of the LR3S type II model are presented in Table 6. None of the eight independent variables in the model is significant at all, as Table 6 makes clear. Furthermore, as shown in Figure 5(A), the performance model values, including precision, recall, F1-score, and accuracy, are comparatively reasonable, just like the OLR model. However, with an AUC value of 0.58, the ROC curve in Figure 5(B) is classified as failing based on the AUC value criteria. Consequently, this model is not appropriate for applying to Stunting Data.

Table 6. The estimation of the LR3S type I model coefficients

Variable	Coefficient (B)	Significance	Exp(B)
Gender (X_1)	-0.186	0.438	0.830
Birth Weight (X_2)	0.499	0.172	1.648
Exclusive Breastfeeding (X_3)	0.293	0.252	1.341
Healthy Latrines (X_4)	0.123	0.632	1.130
Clean Water (X_5)	-0.328	0.364	0.720
Health Insurance (X_6)	0.372	0.185	1.451
Immunization (X_8)	0.054	0.842	1.055
Smoking Family (X_9)	0.242	0.497	1.273
Constant	-1.276	0.000	0.279

ISSN 2462-2052 | eISSN 2600-8718
DOI: https://doi.org/10.37134/jsml.vol13.1.12.2025

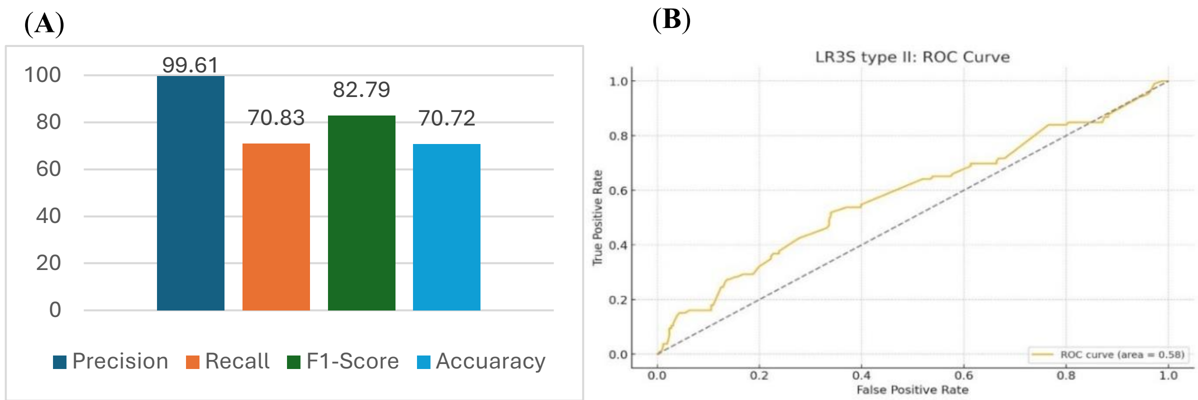

Figure 5. The performance of the LR3S type II model. (**A**) precision, recall, F1-score, and accuracy, (**B**) ROC Curve and AUC

3.4. LR3S Type III Model

The LR3S type III model is a logistic regression model developed using data already obtained through inconsistent sample reduction followed by variable selection. These elimination procedures yield 273 observations with eight independent variables for the LR3S type 3 model. Therefore, two variables, X_7 (worms) and X_{10} (comorbidities), seem useless because they are not involved in the RST decision-making process. The results of the estimation and test of the model's coefficients are shown in Table 7. Table 7 indicates that several variables significantly impact the modeling of the stunting class. Furthermore, as shown in Figure 6(A), the model performs well based on precision, recall, F1-score, and accuracy metrics. Likewise, the final model's usability is shown by its AUC value of 0.81, as presented in Figure 6(B).

Table 7. The estimation of the LR3S type III model coefficients

Variable	Coefficient (B)	Significance	Exp(B)
Gender (X_1)	0.752	0.060	2.121
Birth Weight (X_2)	-1.333	0.005	0.264
Exclusive Breastfeeding (X_3)	-1.451	0.000	0.234
Healthy Latrines (X_4)	-0.799	0.045	0.450
Clean Water (X_5)	0.785	0.192	2.192
Health Insurance (X_6)	0.203	0.627	1.226
Immunization (X_8)	-1.069	0.007	0.344
Smoking Family (X_9)	-1.419	0.003	0.242
Constant	1.404	0.000	4.073

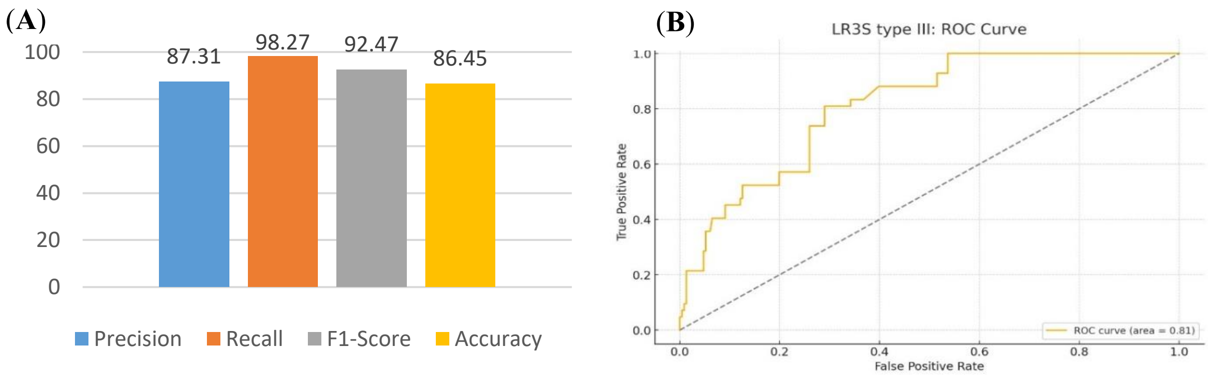

Figure 6. The performance of the LR3S type III model. (**A**) precision, recall, F1-score, and accuracy. (**B**) ROC Curve and AUC

3.5. The Determination of the Best LR Model

This study employs classical and modified logistic regression models to identify the variables impacting the incidence of stunting and severe stunting in toddlers. Four models are used in this study, namely OLR and LR3S types I, II, and III. Modifying the LR model can be a consideration for future researchers, especially in cases where the resulting model is unsatisfactory. This model may be poor due to issues in the data, such as inconsistent samples, irrelevant variables, or a combination of both. Therefore, researchers may perform data preprocessing by eliminating problematic data sections. The modified model is invaluable for studies that collect data through surveys, as it allows researchers to improve the model by preprocessing the data whenever they find a less satisfactory model, saving the need to conduct the survey again.

The performances of the four models used are compared in this section to identify the best model, which will serve as the foundation for investigating stunting risk in the Solok district. Table 8 and Figure 7 compare how well the four LR models performed. Figure 7(A) shows that all four models perform reasonably well regarding precision, recall, F1-score, and accuracy. Nevertheless, none of the OLR and LR3S type II model variables are significant. Furthermore, the ROC curve and the AUC values generated for the OLR and LR3S type II indicate that these models are unsuitable for application, as seen in Figure 7(B). Thus, the plausible candidate models are the LR3S type I and LR3S type III models. Except for the recall value, which is marginally lower than that of the LR3S model, the LR3S type I model outperforms the LR3S type III model according to all performance metrics. As a result, the LR3S type I model is the best of the four models used and will be employed to examine stunting risk in Solok Regency.

Table 8. Performances of LR Models on Stunting Data

Model	Obs.	Number of Sig. Var ($\alpha = 5\%$)	Recall	Precision	F1-score	Accuracy
OLR	362	0	70.75	99.22	82.70	70.44
LR3S type I	273	5	88.28	97.84	92.81	87.18
LR3S type II	362	0	70.83	99.60	92.31	70.70
LR3S type III	273	5	87.31	98.27	92.47	86.45

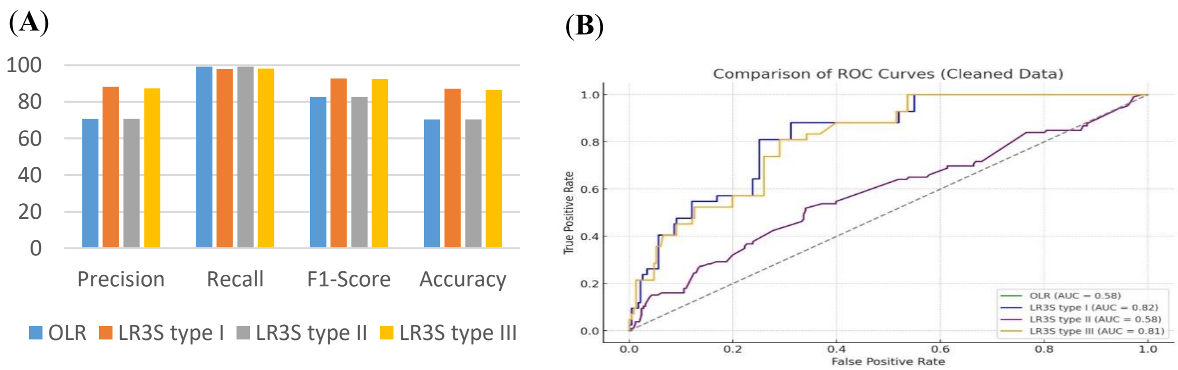

Figure 7. Comparison of LR models' performances. (**A**) precision, recall, F1-size, and accuracy. (**B**) ROC Curve and AUC

This result is also pertinent to the data used, as the variables used are those that Solok Regency policymakers have determined as important in reducing stunting. As a result, all variables remain in the model (no variable reduction), and the variables' degree of significance and impact will be considered when creating priority stunting prevention programs. Based on the LR3S type I model, the variables' positions according to their significance levels are displayed in Table 8. Variables with the smallest p-value are considered the most crucial factor in modeling, and vice versa. Table 8 shows that at a 5% significance level, five variables significantly differentiate stunting classes, namely exclusive breastfeeding (X_3), birth weight (X_2), smoking family (X_9), immunization (X_8), and gender (X_1). Meanwhile, the variables healthy latrines (X_4), comorbidities (X_{10}), worms (X_7), clean water (X_5), and healthy assurance (X_6) do not significantly differentiate stunting classes in Solok Regency.

Next, the significant risk factors differentiating the occurrence of stunting and severe stunting will be interpreted. The findings found that the risk of toddlers experiencing severe stunting is lower in those who are exclusively breastfed (OR = 0.2), those who are born with normal birth weight (OR = 0.3), and those who receive complete immunization (OR = 0.4). Interestingly, toddlers with family members who smoke also have a lower risk of experiencing severe stunting (OR = 0.3). Additionally, the findings indicate that boys are more likely than girls to suffer from severe stunting based on gender.

Table 8. List of independent variables based on significance level in the LR model

Order	Variable	Significance	Odd Ratio (OR)
1.	Exclusive Breastfeeding (X_3)	0.000	0.2
2.	Birth Weight (X_2)	0.005	0.3
3.	Smoking Family (X_9)	0.006	0.3
4.	Immunization (X_8)	0.009	0.4
5.	Gender (X_1)	0.041	2.3
6.	Healthy Latrines (X_4)	0.067	0.5
7.	Comorbidities (X_{10})	0.087	3.6
8.	Worms (X_7)	0.089	12.9
9.	Clean Water (X_5)	0.175	2.3
10.	Healthy Assurance (X_6)	0.816	1.1

Considering the study results, breastfeeding is the primary factor that differentiates between the incidence of stunting and severe stunting among the risk variables affecting these conditions. According to the findings, toddlers who are exclusively breastfed have a decreased chance of experiencing severe stunting. This result is in line with earlier studies that demonstrate that one of the primary risk factors for the development of stunting or severe stunting in babies is not receiving exclusive breastfeeding (Agussalim et al., 2024; Andika et al., 2024; Panggabean et al., 2023; Sihotang et al., 2023; Podungge et al., 2021; Resiyanthi & Yanti, 2021). With exclusive breastfeeding, the immune system is well developed and shielded from infectious illnesses that could compromise nutritional conditions (Fatimah et al., 2021). However, Badriyah's research (2017) found that toddlers who were not exclusively breastfed were less prone to experience stunting. This condition suggests that numerous other factors influence the likelihood of children having stunting or severe stunting and that exclusive breastfeeding is not the only factor determining the occurrence of stunting or severe stunting. For example, even if toddlers receive exclusive breastfeeding for 6 months but are not supported by sufficient nutritious food intake, it will undoubtedly increase the chances of the children experiencing stunting or severe stunting. Thus, in addition to ensuring that the baby receives exclusive breastfeeding, it is also important to consider other variables that may affect the occurrence of stunting or severe stunting.

Furthermore, studies indicate that infants with normal birth weight are less likely to suffer from severe stunting. The results align with research results (Agussalim et al., 2024; Suwarni et al., 2023; Siswati, 2019). Babies with low birth weight have skinny bodies and look very small, which is different from babies with normal birth weight. If the baby fails to grow at an early age, within 2 months, the risk of failure to grow in the following period will be even greater. Therefore, birth weight is an essential measurement of newborn babies and is the best indicator for measuring nutritional status, child growth, and development.

The next significant variable in the Solok Regency stunting model is smoking families. Some previous studies have demonstrated that smoking in a family can increase the risk of stunting. It can happen due to second-hand exposure to cigarette smoke and reduced access to nutritional food (Muchlis et al., 2023; Cao et al., 2022; Shinsugi & Mizumoto, 2022). Interestingly, it is found in this research that smoking family members is one of the significant factors of stunting risk. Still, contrary to the previous studies, the result indicates that the risk of severe stunting is lower for children with smoking families. It could have occurred because a family member was smoking outside the house to protect the kids from cigarette exposure.

The other significant factor associated with stunting in this study is immunization records. This study found that the under-fives who received complete basic immunization were at 0.4 times lower risk of experiencing stunting than the under-fives who did not receive complete basic immunization. This result supports the findings of earlier studies (Permatasari et al., 2023; Shinsugi & Mizumoto, 2022). One particular objective of childhood immunization is to reduce the risk of morbidity and mortality in children from diseases that vaccines can prevent. Children who receive vaccinations are more likely to experience lower rates of morbidity and mortality from diseases that can be avoided by vaccination. Toddlers who do not receive all of their recommended essential vaccinations will develop infections and lose their appetite, resulting in inadequate nutritional intake and stunting (Arsyad et al., 2023).

Moreover, the findings reveal that a child's gender is a significant predictor of stunting and severe stunting, with boys being more likely than girls to have severe stunting. The findings align with previous research by Jiang et al. (2014), Thurstans et al. (2020), and Supadmi et al. (2024). Boys are more susceptible to infection and malnutrition due to sex differences in growth trajectories and immune function that start during pregnancy. However, parents and the home environment are shaped by social and cultural norms, which interpret these biological differences and affect feeding and care practices (Thompson, 2021).

Considering the results of this investigation, the importance of exclusive breastfeeding, normal birth weight, and complete immunization to prevent accidents of stunting or severe stunting must be campaigned intensively. Furthermore, initiatives to discourage family members from smoking at home can greatly lower the amount of second-hand smoke indoors, improving air quality and creating a healthier living space. A healthier living environment and better air quality can result from initiatives to reduce family smoking at home, drastically lowering the amount of second-hand smoke indoors. Of course, it is better if all of the smoking family members can stop their habit. This decision may significantly impact pregnant women, small children, and other vulnerable groups. A healthier living environment and better air quality can result from campaigns to reduce family smoking at home, drastically lowering the amount of second-hand smoke indoors.

Furthermore, as evident in the results, boys may be more susceptible to stunting because of sex differences in immune function and growth trajectories that begin in utero. Social and cultural norms around inappropriate newborn care methods within the family may contribute to this situation by increasing the likelihood of disease. Infants' and children's growth and immune development can be affected by stunting, which can be influenced by how they are fed and cared for. However, stunting or severe stunting can be influenced by a wide range of factors, each of which has varying degrees of importance. Therefore, policies and programs aimed at lowering the prevalence of stunting must be implemented comprehensively while also considering program priorities for the factors that have the most significant impact on stunting incidence.

4. CONCLUSION

This study examines the risk factors of severe stunting in high-risk areas of stunting in Solok Regency, West Sumatra Province, Indonesia, by using ordinary logistic regression (LR) and some modified LR models by applying rough set theory (RST) in data reduction. Data can be reduced in three ways: by removing inconsistent samples, irrelevant attributes, or both. Three types of modified models, Logistic Regression Reduction Rough Set (LR3S) type I, II, and III, have been employed in this study according to each type of data reduction. Based on some criteria, this research finds the classical model (OLR) inappropriate for this stunting data set. Applying the data reduction procedure in data processing using RST found that the modified model, which removed inconsistent samples (LR3S type I), is the best model among the four models used. Based on the best model, it is found that at a 5% significance level, five variables significantly differentiate stunting classes, namely exclusive breastfeeding, birth weight, smoking family, immunization, and gender. Meanwhile, the variables healthy latrines, comorbidities, worms, clean water, and healthy assurance do not significantly differentiate stunting classes in Solok Regency. Priority programs and strategies for preventing stunting in Solok Regency can be developed by considering the degree of significance of these factors' influence.

The proposed model is expected to be useful for researchers in various fields, especially in the field of study that uses surveys in data collection. It is challenging to prevent human and material errors that cause problems in the dataset. The existence of inconsistent samples or irrelevant variables can impact the performance of the logistic regression model. Researchers have alternative models that can be used when there are inconsistent samples or irrelevant variables in the survey data. If they find the model unsatisfactory, they can avoid conducting new surveys or collecting more data, which would require a significant investment of time, money, and energy.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contribution Statement

Izzati Rahmi: Conceptualization, Methodology & Writing original draft. Riswan Efendi: Writing, Review & Editing. Nor Azah Samat: Writing, Review & Editing. Ferra Yanuar: Investigation, Writing & Review. S.M.A. Burney: Writing, Review & Editing. Hazmira Yozza: Investigation, Writing & Review. Muhammad Wahyudi: Software, Visualization & Validation. Erol Egrioglu: Writing, Review & Editing.

Data Availability Statement

The authors confirm that the data supporting the findings of this study are available in the article.

Acknowledgment

The authors are grateful to Andalas University, Padang, Indonesia for supporting this study.

REFERENCES

Adi S, Krisnana I, Rahmawati P D, Magfiroh U. (2023). Environmental factors that affect the incidence of stunting in underfive children: A literature review. *Pediomaternal Nursing Journal*, 9(1), 42-44.

Adrizain R, Faridah L, Fauziah N, Berbudi A, Afifah DN, Setiabudi D, Setiabudiawan B. (2024). Factors influencing stunted growth in children: A study in Bandung Regency focusing on a deworming program. *Parasite Epidemiology and Control*, 26, 1-9.

Agussalim, Zulkifli A, Noor NN, Ansariadi, Stang, Riskiyani S. (2024). Risk factor analysis of stunting in children aged 6-23 months in Tanralili district, Maros regency, Indonesia. *National Journal of Community Medicine*, 15(7), 559-565.

Andika F, Marniati, Rahmi N, Anwar C, Husna A, Safitri F. (2021). Analysis of stunting incidence factors in toddlers aged 23-59 months in the work area of the Padang Tiji Community Health Center, Pidie Regency, 2020. *International Journal of Science, Technology & Management*, 2(3), 642-649.

Arsyad R, Sutarto, Carolia N. (2023). Relationship of basic immunization history and history of infection with stunting incidence in toddlers: A literature review. *Medical Profession Journal of Lampung*, 13(2), 179-181.

Aziz ZA. (2010). An overview of the importance of industrial mathematics. *Journal of Science and Mathematics Letters*, 2(2), 9-17.

Badriyah L, Syafiq A. (2017). The association between sanitation, hygiene, and stunting in children under two years (An analysis of Indonesia's basic health research, 2013). *Makara Journal of Health Research*, 21(2), 35-41.

Benvenuto AF, Adnyana IGA, Samodra VM, Azmi F. (2022). The relationship between worm infection and stunting in children in central Lombok regency, West Nusa Tenggara province. *Jurnal Aisyah: Jurnal Ilmu Kesehatan*, 7(2), 663-668.

Bhowmik KR, Das S. (2019). On selection of an appropriate logistic model to determine the risk factors of childhood stunting in Bangladesh. *Maternal & Child Nutrition*, 15(1), 1-10.

Bhukya H, Manchala S. (2023). Rough set-based feature selection for prediction of breast cancer. *Wireless Personal Communications*, 130(3), 2197-2214.

Brahima JJ, Noor NN, Jafar N. (2020). Immunization and distance relationship status on the birth events 1000 HPK stunting work in bone health district Barebbo. *Enfermeria Clinica*, 30(4), 318-322.

Caballero Y, Alvarez D, Bello R, Garcia M. (2007). Feature selection algorithms using rough set theory, Proceedings of 7th *International Conference on Intelligent Systems Design and Applications*. Rio de Janeiro, Brazil

Cao H. (2021). The utilization of rough set theory and data reduction based on artificial intelligence in recommendation system. *Soft Computing*, 25(3), 2153-2164.

Cao S, Xie M, Jia C, Zhang Y, Gong J, Wang B, Qin N, Zhao L, Yu D, Duan X. (2022). Household second-hand smoke exposure and stunted growth among Chinese school-age children. *Environmental Technology & Innovation*, 27, 1-14.

Çorbacıoglu SK, Aksel G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine*, 23(4), 195-198.

Cramer JS. (2002). *The Origins of Logistic Regression*. Tinbergen Institute Amsterdam.

Efendi R, Mu'at S, Arisandi N, Samsudin NA. (2019). Removing unclassified elements in investigating of financial wellbeing attributes using rough-regression model. Proceedings of the ICSCA '19, February 19-21, 2019, Penang, Malaysia.

Efendi R, Rejab MM, Arbaiy N, Yofi WT, Widyawati SR, Rahmi I, Yozza H. (2024). Improved Rough-Multiple Regression for Unemployment Rate Model in Indonesia. *Recent Advances on Soft Computing and Data Mining*, 94-104.

Ernawati R, Nurjanah M, Wahyuni T. (2024). The correlation of environmental sanitation with stunting incidents in school-age children. *Indonesian Journal of Global Health Research*, 6(2), 553-564.

Fatimah, Massi MN, Febriani ADB, Hatta M, Permatasari TAE, Hidayati E, Hamidah, Khumaidi MA, Akaputra R, Turrahmi H, Anggriani RP (2021). Effect of breastfeeding on children's health and its relationship to NRAMP1 expression: a cross-sectional study. *Annals of Medicine and Surgery*, 71, 1-7.

Hilbe JM. (2016). *Practical Guide to Logistic Regression*. CRC Press, USA.

Hosmer DW, Lemeshow S. (2000). Applied logistic regression, second edition. John Willey and Sons, USE.

Iliou T, Nikolaos C, Nerantzaki M, Anastassopoulos G. (2015). A novel machine learning data preprocessing method for enhancing classification algorithms performance. Proceeding of 16th International Conference on Engineering Applications of Neural Networks (INNS), September 25-28, 2015, Rhodes Island, Greece.

Jiang Y, Su X, Wang C, Zhang L, Zhang X, Wang L, Cui Y. (2014). Prevalence and risk factors for stunting and severe stunting among children under three years old in mid-western rural areas of China. *Child: Care, Health, and Development*, 41(1), 45-51.

Kaka-Khan KM, Mahmud H, Ali AA. (2022). Rough set-based feature selection for predicting diabetes using logistic regression with stochastic gradient descent algorithm. *UHD Journal of Science and Technology*, 6(2), 85-93.

Katoch OR. (2022). Determinants of malnutrition among children: A systematic review. *Nutrition*, 96, 1-8.

Kementrian Kesehatan Republik Indonesia. (2022). *Keputusan Menteri Kesehatan Nomor HK.01.07/MENKES/1928/2022 tentang Pedoman Nasional Pelayanan Kedokteran Tata Laksana Stunting*. Kementrian Kesehatan Republik Indonesia, Jakarta.

Kementrian Kesehatan Republik Indonesia. (2024). *Panduan Hari Gizi Nasional ke 64 Tahun 2024*. Kementrian Kesehatan Republik Indonesia, Jakarta.

Kim K, Pant P, Yamashita EY. (2008). Hit-and-run crashes: Use of rough set analysis with logistic regression to capture critical attributes and determinants. *Transportation Research Record Journal of the Transportation Research Board*, 2083(1), 114-121.

Li X. (2014). Attribute selection methods in rough set theory. Master's Thesis, San Jose State University, California.

Liu D, Li T, Liang D. (2014). Incorporating logistic regression to decision-theoretic rough sets for classifications, *International Journal of Approximate Reasoning*, 55(1), 197-210.

Muchlis N, Yusuf RA, Rusydi AR, Mahmud NU, Hikmah N, Qanitha A, Ahsan A. (2023). Cigarette smoke exposure and stunting among under-five children in rural and poor families in Indonesia. *Environmental Health Insights*, 17, 1-7.

Ohyfer M, Moniaga JV, Yunidwi KR, Setiawan MI. (2017). Logistic regression and growth charts to determine children nutritional and stunting status: A Review. *Procedia Computer Science*, 116, 232-241.

Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology*, 48(3), 128-138.

Panggabean ER, Lustiyati ED, Yuningrum H, Trisnowati H. (2023). Family smoking behavior and stunting among children in rural areas of Sleman, Yogyakarta: A case-control study. *Public Health and Preventive Medicine Archive*, 11(2), 222-232.

Paun R, Bia MB, Shagti I, Gunawan YES, Krisyudhanti E, Dafroyati Y, Mau F. (2021). The relationship between intestinal worm infection and stunting in elementary school children in South Central Timor Regency, East Nusa Tenggara. Proceeding of the 8th International Conference on Public Health, November 17-18, 2021, Solo, Indonesia.

Pawlak Z. (1991). *Rough Sets, Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, Amsterdam.

Pawlak Z. (1998). Rough set theory and its applications to data analysis. *Cybernetics and Systems*, 29(7), 661-688.

Permatasari T, Chairunnisa, Djarir H, Herlina L, Fauziah M, Andriyani, Chadirin Y. (2023). The determinants of stunting in the under-five in three municipalities in the special capital region of Jakarta. *National Public Health Journal*, 18(1), 32-40.

Pessoa ASA, Stephany S. (2014). An innovative approach for attribute reduction in rough set theory. *Intelligent Information Management*, 6(5), 223-239.

Podungge Y, Yulianingsih E, Porouw HS, Saraswati E, Tompunuh MM, Claudia JG, Zakaria R, Labatjo R. (2021). Determinant factors of stunting in under-five children. *Open Access Macedonian Journal of Medical Sciences,* 9(B), 1717-1726.

Putri TA, Salsabilla DA, Saputra RK. (2021). The effect of low birth weight on stunting in children under five: a meta-analysis. *Journal of Maternal and Child Health*, 6(4), 496-506.

Rahhali M, Oughdir L, Lahmadi Y, Khattabi MZE. (2024). The impact of data preprocessing on the quality and effectiveness of e-learning. *International Journal of Intelligent Systems and Applications in Engineering*, 12 (16), 59-65.

Rahmi I, Efendi R, Samat NA, Yozza H, Syafwan M. (2024). Examining risk factors of anemia in pregnancy using hybrid logistic regression model and rough set theory. *Barekeng: Journal of Mathematics and Its Application*, 18(1), 537-552.

Rahmi I, Efendi R, Samat NA, Yozza H, Wahyudi M. (2024). *The Effects of Data Reduction Using Rough Set Theory on Logistic Regression Model*. In: Ghazali, R., Nawi, N.M., Deris, M.M., Abawajy, J.H., Arbaiy, N. (eds) Recent Advances on Soft Computing and Data Mining. SCDM 2024. Lecture Notes in Networks and Systems, Springer.

Rakhmahayu A, Dewi YLR, Murti B. (2019). Logistic regression analysis on the determinants of stunting among children aged 6-24 months in Purworejo regency, Central Java. *Journal of Maternal and Child Health*, 4(3), 158-169.

Rasyidah N, Efendi R, Nawi NM, Deris MM, Burney SMA. (2023). Cleansing of inconsistent sample in linear regression model based on rough sets theory. *Systems and Soft Computing*, 5, 1-14.

Resiyanthi NK, Yanti NL. (2021). Relationship between exclusive breastfeeding and basic immunization status related with stunting in toddlers aged 12-24 months. *Indonesia Journal of Global Health Research*, 3(4), 563-570.

Rissino S, Torres GL. (2009). Rough set theory - fundamental concepts, principals, data extraction, and applications. *In Data Mining and Knowledge Discovery in Real Life Applications*. Intech Open.

Saenong MT, Sulaeman, Bakhtiar, Purnama J, Kenre I. (2024). The relationship between cigarette smoke exposure and stunting among children in the working area of the Pangkajene Health Center, Sidrap Regency in 2023. *Journal Of Nursing Practice*, 7(2), 325-334.

Sami O, Elsheikh Y, Almasalha F. (2021). The role of data preprocessing techniques in improving machine learning accuracy for predicting coronary heart disease. *International Journal of Advanced Computer Science and Applications*, 12(6), 812-820.

Shinsugi C, Mizumoto A. (2022). Associations of nutritional status with full immunization coverage and safe hygiene practices among Thai children aged 12-59 months. *Nutrients*, 14(1), 1-8.

Sihotang WY, Hulu VT, Samosir FJ, Pane, Hartono, Manalu P, Siagian M, Panjaitan HIL. (2023). Determinants of stunting in children under five: a scoping review. *The Indonesian Journal of Nutrition*, 12(1), 9-20.

Siswati T. (2019). Risk factors for stunting and severe stunting among under five years children in rural areas in Indonesia. *International Journal of Science and Research*, 8(11), 1635-1640.

Supadmi S, Laksono AD, Kusumawardani HD, Ashar H, Nursafingi A, Kusrini I, Musoddaq MA. (2024). Factor related to stunting of children under two years with working mothers in Indonesia. *Clinical Epidemiology and Global Health*, 26, 1-6.

Suwarni L, Selviana, Vidyastuti, Abdullah A, Adi P. (2023). Risk factors for severe stunded among Children aged 2-5 years with stunting in Pontianak City, Indonesia. *Global Health Management Journal*, 6(2), 81-89.

Thompson AL. (2021). Greater male vulnerability to stunting? Evaluating sex differences in growth, pathways and biocultural mechanisms. *Ann Hum Biol,* 48(6), 466-473.

Thurstans S, Opondo C, Seal A, Wells J, Khara T, Dolan C, Briend A, Myatt M, Garenne M, Sear R, Kerac M. (2020). Boys are more likely to be undernourished than girls: a systematic review and meta-analysis of sex differences in undernutrition. *BMJ Global Health*, 5(12), 1-17.

TPPS Provinsi Sumatera Barat. (2024). Penyelenggaraan Percepatan Penurunan Stunting 2023. Pemerintah Daerah Provinsi Sumatera Barat, Padang.

United Nations. (2022). *Goal 2:* End hunger, achieve food security and improved nutrition and promote sustainable agriculture. United Nations; [Accessed November 11, 2024]. https://sdgs.un.org/goals/goal2

WHO. (2010). Nutrition Landscape Information System (NLIS) Country Profile Indicators:Interpretation Guide. Switzerland, WHO Press.

WHO. (2024). Nutrition Targets 2025: Policy Brief Series. WHO; [Accessed: November 11, 2024]. https://www.who.int/publications /i/item/WHO-NMH-NHD-14.2,

Wijayanti LA, Nurseskasatmata SE. (2022). The relationship between history of birth weight and stunting event in children of 24 - 60 months. *International Journal of Health Sciences*, 6(6), 9732-9741.

Yaile C, Delia A, Rafael B, Maria G. (2007). Feature Selection Algorithms Using Rough Set Theory. *Intelligent Systems Design and Applications Conference*, 1-5.