BAREKENG: Journal of Mathematics and Its Applications
March 2025 Volume 19 Issue 1 Page 0001–0012
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss1pp0001-0012

ENSEMBLE BAGGING WITH ORDINAL LOGISTIC REGRESSION TO CLASSIFY TODDLER NUTRITIONAL STATUS

Luthfia Hanun Yuli Arini [1*], Solimun[2], Achmad Efendi[3], Adji Achmad Rinaldo Fernandes[4]

[1,2,3,4]Department of Statistics, Faculty of Mathematics and Natural Sciences, Brawijaya University
Jl. Veteran No.10-11, Ketawanggede, Lowokwaru, Malang, 65145, Indonesia

*Corresponding author's e-mail: * **luthfiaarini@student.ub.ac.id***

ABSTRACT

Article History:

Received: 22nd November 2023
Revised: 22nd June 2024
Accepted: 5th November 2024
Published: 13th January 2025

Keywords:

Bagging;
Ordinal logistic regression;
Stunting

One problem in classifying stunting data is that the data used does not have a balanced proportion. This study aims to apply the logistic regression classification method with ordinal scale response variables to overcome class imbalance through the ensemble bagging approach. The data used is secondary data in the form of final research reports that have been tested for validity and reliability. The predictor variables used are economic conditions, health services and the environment with categorical response variables, namely the nutritional status of toddlers in the categories of stunting, normal and high. The methods used are ordinal logistic regression and ensemble bagging on ordinal logistic regression with bootstraps of 100, 500, and 1000. The variables that influence the nutritional status of toddlers are Economic Conditions, Health Services, and the Environment. The results of the study showed that the accuracy, sensitivity, specificity, and F1-Score for ordinal logistic regression were smaller than ensemble bagging in ordinal logistic regression. The best classification method obtained was bagging logistic regression with a bootstrap number of 500 and obtained an accuracy value of 85%, sensitivity of 87.2%, specificity of 72.6%, and F1-Score of 79.3%.

How to cite this article:

L. H. Y. Arini, Solimun, A. Efendi and A. A. R. Fernandes., "ENSEMBLE BAGGING WITH ORDINAL LOGISTIC REGRESSION TO CLASSIFY TODDLER NUTRITIONAL STATUS", *BAREKENG: J. Math. & App.*, vol. 19, iss. 1, pp. 0001-0012, March, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

Classification is a method in statistics that deals with the separation and grouping of objects into certain categories. In machine learning, classification is the process of grouping objects that have the same traits or characteristics into several classes. The current classification is not only applied to binary category response variables but has developed into nominal or ordinal with more than two categories. The classification method commonly used is logistic regression. Logistic regression can be used to describe and test hypotheses about the relationship between a categorical response variable and one or more shock or continuous predictor variables [1].

The results of the 2022 Indonesian Toddler Nutritional Status Survey show that the prevalence of stunting in Indonesia is 21.6%, which means that in 2022, 21.6% of toddlers in Indonesia were detected as stunted. This figure is still noticeably higher than the maximum tolerance for stunting of 20% set by WHO. Based on Regulation Number 2 of the Minister of Health of the Republic of Indonesia on Anthropometric Standards for Indonesian Children, the Gini status category based on Height/Age is divided into five categories, namely very stunted, stunted, normal, tall, and very tall. However, based on the data acquired in the field, no categories of very stunted toddlers and very tall toddlers were found, so that the response variable categories used are stunted, normal and tall toddlers with an unbalanced proportion where toddlers with normal height are more numerous than stunted and tall toddlers.

Class imbalance in the classification process can cause the classification results for minor data to be covered by predictions for major data, in other words, the classification results for minor data will be incorrect [2]. One way to overcome the problem of data imbalance is to use an ensemble algorithm. Ensemble learning can be used to improve model accuracy by combining several different learning models or algorithms. Ensemble learning can reduce the spread (dispersion) of predictions and model performance. The ensemble method that is commonly used is the ensemble bagging. Bagging is an ensemble method for training data on a subset of random samples from the original dataset [3]. Ensemble Bagging Logistic Regression is a technique used to improve the performance of Logistic Regression. It involves combining multiple models to improve the accuracy of predictions. One study proposed an improvement to the Logistic Regression method by using the Ensemble technique with the Newton Raphson parameter estimation method [4]. Another study implemented the Bagging technique to improve the performance of J48 and Logistic Regression in predicting online purchasing interest [5]. The study used the Bagging technique, which is one of the Ensemble Learning methods, to improve the accuracy of the model. The Bagging technique uses only one type of base model and trains it independently in parallel, then combines the results to obtain the best outcome [6]. The results of this study indicate that the ensemble bagging approach method is better than the classical classification method for data with proportion imbalance problems.

Based on the description above, previous research used ensemble methods on binary response data. In this study, it was applied to three categories of responses. This research develops a logistic regression classification method with ordinal scale response variables to overcome class imbalance through an ensemble bagging approach. The bagging method used is bootstrap aggregation. Evaluation of the classification performance based on accuracy, sensitivity, specificity, and F1-Score values to determine the best classification method for classifying the nutritional status of toddlers.

2. RESEARCH METHODS

2.1 Linear Probability Model

Linear probability model is used when the response variable is qualitative (categorical) [7]. The linear probability model with a binary response variable can be expressed by Equation (1).

$$Y_i = \beta_0 + \beta_{1\mathrm{i}} X_{1i} + \cdots + \beta_{\mathrm{ji}} X_{ji} + \varepsilon_i ; i = 1,2, \dots , n ; j = 1,2, \dots p \tag{1}$$

with

Y_i = 1, if the response falls into the first category
Y_i = 0, if the response falls into the second category
X_{ji} = the i-th value of the j-th predictor variable
ε_i = error that is assumed to be binomial distribution
n = the number of samples

p = the number of predictor variables

Y_i is a binary response variable, so Y_i has a Bernoulli distribution with a probability distribution as shown in Equation (2) and Equation (3).

$$P(Y_i = 1) = \pi_i \tag{2}$$

$$P(Y_i = 0) = 1 - \pi_i \tag{3}$$

The expected value or average response ($E[Y_i]$) is presented in Equation (4).

$$E[Y_i] = \beta_0 + \beta_1 X_i = 1(\pi_i) + 0(1 - \pi_i) = \pi_i \tag{4}$$

There are three main problems if the response variable values are divided into binary categories, namely not normally distributed, inconsistent error ranges, and limitations on the response. [7]. These three problems can be overcome by transforming $E[Y_i] = \beta_0 + \beta_1 X_i$ so that the values range between 0 and 1. The transformation function that is popularly used is the logistic cumulative distribution function in Equation (5).

$$F(z) = \frac{\exp(z)}{1 + \exp(z)} \tag{5}$$

2.2 Ordinal Logistic Regression Model

The logistic regression model obtained by substituting $\beta_0 + \beta_1 X_i$into z Equation (5) so that the logistic regression model formed is presented in Equation (6).

$$E[Y_i] = \pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} i = 1,2, \dots, n \tag{6}$$

If more than one predictor variable (X) is used, the logistic regression model shown in Equation (7).

$$\pi(x_i) = \mathrm{P}(\mathrm{Y} = 1 | X_1, X_2, \dots, X_i) \frac{e^{\beta_0 + \beta_1 X_{1\mathrm{i}} + \beta_2 X_{2\mathrm{i}} + \cdots + \beta_j X_{ji}}}{1 + e^{\beta_0 + \beta_1 X_{1\mathrm{i}} + \beta_2 X_{2\mathrm{i}} + \cdots + \beta_{ji} X_{ji}}} \tag{7}$$

with
$\pi(x_i)$: the probability of a successful event occurring
β_0 : intercept
β_i : the value of the j-th regression coefficient with $j = 1,2, \dots, p$
X_{ji} : the i-th value of the j-th predictor variable

Then a logit transformation is carried out to simplify Equation (7). So, we obtain the logit model in Equation (8).

$$\begin{aligned} \mathrm{logit}[\pi(x)] &= \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) \\ &= \beta_0 + \boldsymbol{\beta}' \boldsymbol{X} \end{aligned} \tag{8}$$

with
$\pi(x)$: the probability of a successful event occurring
β_0 : intercept
$\boldsymbol{\beta}$: $(\beta_1, \dots, \beta_k)'$ vector of model coefficients

If there are J categories in the response variable, then the cumulative probability of the response variable is shown by Equation (9).

$$\pi_j(x) = P(Y \leq j | X) = \frac{e^{\theta_j + \boldsymbol{\beta}' \boldsymbol{X}}}{1 + e^{\theta_j + \boldsymbol{\beta}' \boldsymbol{X}}} \tag{9}$$

With $j = 1,2, \dots, J - 1$
The logit model for cumulative odds in Equation (9) is presented in Equation (10).

$$\begin{aligned} \text{logit } P(Y \le j|X) &= \ln\left(\frac{P(Y \le j|X)}{1 - P(Y \le j|X)}\right) \\ &= \ln\left(\frac{\dfrac{e^{\theta_j+\boldsymbol{\beta}'\boldsymbol{X}}}{1+e^{\theta_j+\boldsymbol{\beta}'\boldsymbol{X}}}}{1-\dfrac{e^{\theta_j+\boldsymbol{\beta}'\boldsymbol{X}}}{1+e^{\theta_j+\boldsymbol{\beta}'\boldsymbol{X}}}}\right) \\ &= \theta_j + \boldsymbol{\beta}'\boldsymbol{X} \end{aligned} \tag{10}$$

with
θ_j: intercept of the logit j model ($j = 1,2,3,\dots,J-1$)
$\boldsymbol{\beta}$: $(\beta_1,\dots,\beta_k)'$vector of logit model coefficients
$\boldsymbol{X}$: $(\boldsymbol{x}_1,\dots,\boldsymbol{x}_k)'$matrix of k predictor variables

2.2.1 Parameter Estimation

Maximum Likelihood Estimation (MLE) method can be used to estimate logistic regression parameters [1]. If the response variable has an ordinal measuring scale $(y_1, y_2, \dots, y_J; n; \pi_1\pi_2, \dots, \pi_J)$ then the observation likelihood function is presented in Equation (11).

$$L(\boldsymbol{\theta},\boldsymbol{\beta}) = \prod_{i=1}^{n}\left[\left(\pi_1(x)\right)^{y_{1i}}\left(\pi_2(x)\right)^{y_{2i}-y_{1i}}\dots\left(\pi_J(x)\right)^{y_{Ji}-y_{J-1i}}\right] \tag{11}$$

with
$\boldsymbol{\pi}_j(\boldsymbol{x})$: $\left(\pi_1(x),\dots,\pi_J(x)\right)'$ vector of the probability of a successful event occurring
y_j : categorical ($j = 1,2,\dots,J$) response variable j
n : total observations
The log likelihood function for Equation (11) is presented in Equation (12).

$$\begin{aligned} \ln L(\boldsymbol{\theta},\boldsymbol{\beta}) &= l(\theta,\beta) \\ &= \sum_{i=1}^{n}\left[y_{1i}\ln(\pi_1(x)) + y_{2i}\ln(\pi_2(x)) + \dots + y_{Ji}\ln(\pi_J(x))\right] \\ &= \sum_{i=1}^{n}\left[\begin{matrix} y_{1i}\ln\left(\dfrac{e^{\theta_1+\boldsymbol{\beta}'\boldsymbol{X}}}{1+e^{\theta_1+\boldsymbol{\beta}'\boldsymbol{X}}}\right) + (y_{2i}-y_{1i})\ln\left(\dfrac{e^{\theta_2+\boldsymbol{\beta}'\boldsymbol{X}}}{1+e^{\theta_2+\boldsymbol{\beta}'\boldsymbol{X}}} - \dfrac{e^{\theta_1+\boldsymbol{\beta}'\boldsymbol{X}}}{1+e^{\theta_1+\boldsymbol{\beta}'\boldsymbol{X}}}\right) + \cdots \\ +(y_{Ji}-y_{J-1i})\ln\left(1-\dfrac{e^{\theta_{J-1}+\boldsymbol{\beta}'\boldsymbol{X}}}{1+e^{\theta_{J-1}+\boldsymbol{\beta}'\boldsymbol{X}}}\right) \end{matrix}\right] \\ &= \sum_{i=1}^{n}\left[\begin{matrix} y_{1i}\ln\left(\dfrac{e^{\theta_1+\boldsymbol{\beta}'\boldsymbol{X}}}{1+e^{\theta_1+\boldsymbol{\beta}'\boldsymbol{X}}}\right) + (y_{2i}-y_{1i})\ln\left(\dfrac{e^{\boldsymbol{\beta}'\boldsymbol{X}}\left(e^{\theta_2}-e^{\theta_1}\right)}{\left(1+e^{\theta_2+\boldsymbol{\beta}'\boldsymbol{X}}\right)\left(1+e^{\theta_1+\boldsymbol{\beta}'\boldsymbol{X}}\right)}\right) + \cdots \\ +(y_{Ji}-y_{J-1i})\ln\left(1-\dfrac{e^{\theta_{J-1}+\boldsymbol{\beta}'\boldsymbol{X}}}{1+e^{\theta_{J-1}+\boldsymbol{\beta}'\boldsymbol{X}}}\right) \end{matrix}\right] \end{aligned} \tag{12}$$

The value of $\boldsymbol{\theta} = [\theta_1\ \theta_2 \dots \theta_{J-1}]$ and $\boldsymbol{\beta}' = [\beta_1\ \beta_2 \dots \beta_k]$ is obtained from the derivative of the log likelihood from Equation (12) which is maximized when $\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\beta}} \sim N((\boldsymbol{\theta},\boldsymbol{\beta}),\boldsymbol{I}^{-1}(\boldsymbol{\theta},\boldsymbol{\beta}))$. Looking for the values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ requires numerical methods. This is intended to simplify the calculations because the log likelihood function is a non-linear function. The numerical method used is Newton-Raphson in Equation (13).

$$\widehat{\boldsymbol{\phi}}^{\boldsymbol{k+i}} = \boldsymbol{\phi}^{\boldsymbol{k}} - \boldsymbol{I}^{-\boldsymbol{1}}\boldsymbol{l}'(\boldsymbol{\phi}) \tag{13}$$

Where k is the number of iterations and $\boldsymbol{\phi} = \begin{bmatrix}\boldsymbol{\theta}\\ \boldsymbol{\beta}\end{bmatrix}$. $\boldsymbol{l}'(\boldsymbol{\phi})$ is the first derivative of the log likelihood function $\boldsymbol{\phi}$. Iterative value estimation $\boldsymbol{\phi}$ will stop until it reaches a convergent condition or when the value of $\left|\boldsymbol{\phi}^{\boldsymbol{k}} - \boldsymbol{\phi}^{\boldsymbol{k-1}}\right| < \boldsymbol{\varepsilon}$, where the expected value of $\boldsymbol{\varepsilon}$ is very small ($\boldsymbol{\varepsilon} = 10^{-5}$).

2.2.2 Hypothesis Testing

The hypothesis test carried out is the same as the hypothesis test in ordinary logistic regression. Testing is carried out simultaneously and partially. The simultaneous test uses the G-test statistic and the partial test uses the Wald test.

a. G-test statistics (Simultaneous test)
The hypothesis used is as follows [7].

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$
H_1: There is at least one $\beta_j \neq 0$ $(i = 1,2,\dots,p)$

The test statistic used is the likelihood ratio of the likelihood ratio test) with the formula in **Equation (14)**.

$$G = -2\ln\left(\frac{L_0}{L_p}\right) \tag{14}$$

with
L_0: likelihood value of the model without exogenous variables
L_p: likelihood value of the model with exogenous variables

The G-test statistic uses a distribution approach $\chi^2$with degrees of freedom p (many parameters). So, a decision to reject can be made H_0 if the G-test statistical value is > critical point χ_p^2or the G-test statistical p-value is $< \alpha$ $(0{,}05)$.

b. Wald test (Partial test)
The hypothesis used in the Wald test is as follows [7].
H_0 : $\beta_j = 0$ $(j = 1,2,\dots,p)$
H_1 : $\beta_j \neq 0$ $(j = 1,2,\dots,p)$

Wald test statistics are presented in the **Equation (15)**.

$$W = \frac{\hat{\beta}_j}{Se\left(\hat{\beta}_j\right)} \tag{15}$$

with
$\hat{\beta}_j$: estimator of the j-th β parameter $(j = 1,2,\dots,p)$
$Se(\hat{\beta}_j)$: standard error estimator for j-th β parameter (j $= 1,2,\dots,p)$

Wald test statistic uses a standard normal distribution approach if the sample used is large. So, a decision to reject H_0 can be taken if the Wald test statistic value $(W) >$ critical point (Z) or the Wald test statistic p-value $< \alpha$ $(0{,}05)$.

2.3 Bagging Ensemble

Ensemble is a machine learning algorithm where some models are weak trained to solve problems and combined to get better results [6]. Ensemble techniques are able to provide predictions with very good accuracy [8]. The main idea of an ensemble is to combine multiple sets of models that solve the same problem to obtain a more accurate model. [9]. One method of ensemble bagging is bootstrap aggregating (bagging). Algorithm of bagging shows in **Figure 1**.

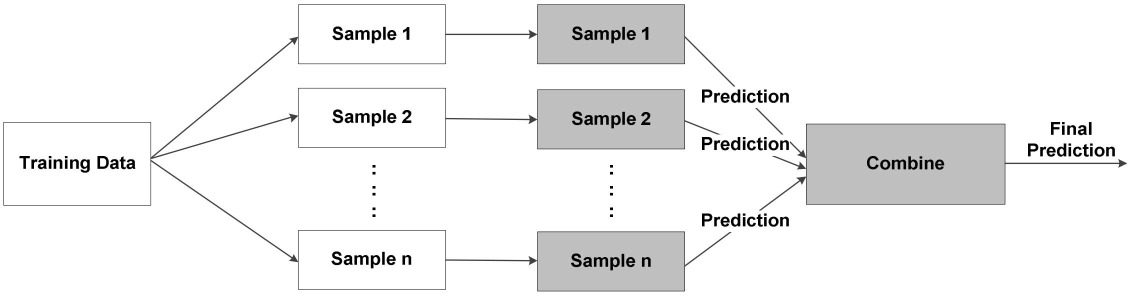

Figure 1. **Bagging Ensemble Learning**

2.4 Classification Accuracy

Classification accuracy is measured based on four criteria, including: accuracy, sensitivity, specificity, and F1-Score. For example, in the classification of toddler nutritional status based on Height/Age, there are three categories, namely 1) Stunted, 2) Normal, and 3) Tall. Which way the most common way to show classification results is by presenting them in the form of a confusion matrix to obtain accuracy, sensitivity, specificity, and F1-Score values as in **Table 1**.

Table 1. Confusion Matrix

Actual	**Prediction**		
	category 1	**category 2**	**category 3**
category 1	a	b	c
category 2	d	e	f
category 3	g	h	i

Classification accuracy is calculated through accuracy using **Equation (16)**, sensitivity is calculated using **Equation (17)**, specificity is calculated using **Equation (18)**, and F1-Score is calculated using **Equation (19)**.

$$\text{Accuracy} = \frac{\text{a+e+i}}{\text{a+b+c+d+e+f+g+h+i}} \tag{16}$$

$$\text{Sensitivity} = \frac{\frac{a}{a+(d+g)} + \frac{e}{e+(b+h)} + \frac{i}{i+(c+f)}}{3} \tag{17}$$

$$\text{Specificity} = \frac{\frac{a}{a+(b+c)} + \frac{e}{e+(d+f)} + \frac{i}{i+(g+h)}}{3} \tag{18}$$

$$\text{F1} - \text{Score} = 2\left(\frac{\text{Sensitivity} \times \text{Specificity}}{\text{Sensitivity} + \text{Specificity}}\right) \tag{19}$$

with
a : number of observations from group 1 and correctly classified to group 1
b : number of observations from group 1 that are classified into group 2
c : number of observations from group 1 that are classified into group 3
d : number of observations from group 2 that are classified into group 1
e : number of observations from group 2 and correctly classified into group 2
f : number of observations from group 2 that are classified into group 3
g : number of observations from group 3 that are classified into group 1
h : number of observations from group 3 that are classified into group 2
i : number of observations from group 3 and correctly classified into group 3

2.5 Research data

The data used is secondary data from research [10]. The study population included mothers with young children in Wajak district. The sample for this study consisted of mothers with young children in the village of Sumberputih. The sampling method was stratified random sampling. Stratified random sampling is a probability sampling method where the population is divided into smaller subgroups or strata based on specific characteristics. Each stratum is then sampled randomly to obtain a representative sample of the entire population. The sample size is determined using the Slovin formula with an expected error was 10% and an anticipated response rate was 80% and the number of populations was 389.

$$n = \frac{N}{1+N(e^2)}\frac{1}{rr} = \frac{389}{1+389(0{,}8^2)}\frac{1}{0{,}8} = 99{,}437 \approx 100$$

The number of samples was 100 respondents. The research instrument was tested for validity and reliability. Validity and reliability are two critical concepts in questionnaire design, ensuring that the instrument accurately measures what it intends to measure and produces consistent results. Validity refers to the extent to which a questionnaire measures what it is intended to measure. Reliability refers to the consistency of the questionnaire's results over time, across different items, or with different raters. The predictor variables in this study were economic level, health services, children's diet, and environment. The response variable in this study is the nutritional status of toddlers in the stunted, normal, and tall groups which are respectively given categories number 1, 2, and 3.

2.6 Steps

Secondary data processing uses ordinal logistic regression and bagging algorithms ordinal logistics.

a. Stages of ordinal logistic regression analysis
   1) Split the data into training data and testing data with a ratio of 80:20 (80% for training data and 20% for testing data).
   2) Form an ordinal logistic regression model according to **Equation (9)**.
   3) Estimate ordinal logistic regression parameters on training data in **Equation (13)**.
   4) Test the significance of parameters using the G test of **Equation (14)** and the Wald test of **Equation (15)**.
   5) Test data with the Accuracy, Sensitivity, Specificity, and F1-Score formulas using **Equation (16)** to **Equation (19)**.

b. Stages of ordinal logistic regression bagging analysis
   1) Split the data into training data and testing data with a ratio of 80:20.
   2) Perform resampling bootstrapping $\mathcal{L}_B$ as many as n from the training data with bootstrapping carried out as many as 100, 500, and 1000.
   3) Form an ordinal logistic regression model according to **Equation (9)**.
   4) Test data with the Accuracy, Sensitivity, Specificity, and F1-Score formulas using **Equation (16)** to **Equation (19)**.

3. RESULTS AND DISCUSSION

3.1 Results of Logistic Regression Analysis

The results of estimating parameters and hypothesis test of ordinal logistic regression are presented in **Table 2**.

Table 2. Parameter Estimation

Variable Relationships	Parameter	Parameter Estimation	t Test Statistics	p-Value
Intercept for $\boldsymbol{\pi_1(x)}$	β_0	18,689	37,378	<0,001
Intercept for $\boldsymbol{\pi_2(x)}$	β_0	23,973	62,297	<0,001
$\mathbf{X_1}$ to $\mathbf{Y}$	β_{X_1Y}	1.709	2.558	0.010
$\mathbf{X_2}$ to $\mathbf{Y}$	β_{X_2Y}	2.891	2.430	0.015
$\mathbf{X_3}$ to $\mathbf{Y}$	β_{X_3Y}	−0.262	−0.272	0.784
$\mathbf{X_4}$ to $\mathbf{Y}$	β_{X_4Y}	2,055	2.701	0.006

Based on the coefficient estimation results, an ordinal logistic regression model Equation can be formed in this research. Logistic regression model of the influence of exogenous variables: economic conditions (X_1), health services (X_2), children's eating patterns (X_3), and the environment (X_4) on Toddler Nutrition Consumption (Y) as follows.

$$\pi_1(x) = \frac{\exp(18.689 + 1.709X_1 + 2.891X_2 - 0.262\mathrm{X}_3 + 2.055\mathrm{X}_4)}{1 + \exp(18.689 + 1.709X_1 + 2.891X_2 - 0.262\mathrm{X}_3 + 2.055\mathrm{X}_4)}$$
$$\pi_2(x) = \frac{\exp(23.973 + 1.709X_1 + 2.891X_2 - 0.262\mathrm{X}_3 + 2.055\mathrm{X}_4)}{1 + \exp(23.973 + 1.709X_1 + 2.891X_2 - 0.262\mathrm{X}_3 + 2.055\mathrm{X}_4)}$$

Interpretation of the model formed through the odds ratio value obtained from Exp(β). The odds ratio for $\beta_{X_1Y}, \beta_{X_2Y}, \beta_{X_3Y}, \beta_{X_4Y}$ are 12.910, 11.359, 0.761, and 14.895. Thus, adding one unit to the economic conditions (X_1) score will increase the risk of stunting by 12.910 times compared to the risk of normal or tall toddlers. Adding one unit to the health services (X_2) score will increase the risk of stunting by 11.359 times compared to the risk of normal or tall toddlers. Adding one unit to the health children's eating pattern (X_3) score will increase the risk of stunting by 0.761 times compared to the risk of normal or tall toddlers, and adding one unit to the environment (X_4) score will increase the risk of stunting by 14.895 times compared to the risk of normal or tall toddlers

a) Simultaneous test

Simultaneous tests are carried out using the G test statistic. The G test statistic value (36.903) > *Chi-square value* (5.991) so that a Reject decision on H_0 can be taken. So, at a real level 5% it can be concluded that economic conditions (X_1), health services (X_2), children's eating patterns (X_3)and the environment (X_4) together have a significant effect on the nutritional status of toddlers (Y).

b) Partial test

The partial test is carried out using the Wald Test. The test results are shown in Table 2. Based on the results of the Wald Test, the economic conditions (X_1) regarding the Nutritional Status of Toddlers (Y) are obtained $p -$ value < 0.05 so that a decision to reject can be taken H_0. With a real level of 5% it can be concluded that economic conditions (X_1) partially has a significant effect on the nutritional status of toddlers (Y). Health services (X_2)regarding the Nutritional Status of Toddlers (Y) is obtained $p -$ value < 0.05 so that a decision to reject H_0 can be made. With a real level of 5% it can be concluded that Health Services (X_2) partially has a significant effect on the nutritional status of toddlers (Y). Children's Eating Patterns (X_3) on Toddler Nutritional (Y) Status is obtained $p -$ value < 0.05 so that acceptance H_0 decisions can be made. With a real level of 5%, it can be concluded that children's eating patterns (X_3) partially has a significant effect on the nutritional status of toddlers (Y). Environment (X_4) regarding the Nutritional Status of Toddlers (Y) is obtained $p -$ value < 0.05 so that a decision to reject H_0 can be made. With a real level of 5%, it can be concluded that children's eating patterns (X_3) partially has a significant effect on the nutritional status of toddlers (Y).

Based on the classification results of the testing data, confusion matrices were obtained as in Table 3.

Table 3. Confusion Matrix Ordinal Logistic Regression

Actual	Prediction		
	Stunted	Normal	Tall
Stunted	2	2	0
Normal	3	11	0
Tall	0	1	1

Based on Table 2, information regarding classification accuracy was obtained using logistic regression analysis. Logistic regression mode was able to correctly classify the nutritional status of 14 toddlers out of a total of 20 toddlers.

3.2 Results of Logistic Regression Bagging Analysis

The number of bootstraps carried out on the training data are 100, 500, and 1000 bootstraps. The classification results based on the number of bootstraps are presented in Table 4.

Table 4. Confusion Matrix Bagging Ordinal Logistic Regression

Number of Bootstrap	Actual	Prediction		
		Stunted	Normal	Tall
100	**Stunted**	2	2	0
	Normal	1	13	0
	Tall	0	1	1
500	**Stunted**	3	1	0
	Normal	1	13	0
	Tall	0	1	1
1000	**Stunted**	2	2	0
	Normal	2	12	0
	Tall	0	1	1

Based on Table 3, information regarding classification accuracy was obtained using bagging logistic regression analysis. Bagging logistic regression using 100 bootstrap was able to correctly classify the nutritional status of 16 toddlers out of a total of 20 toddlers. Bagging logistic regression using 500 bootstrap was able to correctly classify the nutritional status of 17 toddlers out of a total of 20 toddlers. Bagging logistic regression using 1000 bootstrap was able to correctly classify the nutritional status of 16 toddlers out of a total of 16 toddlers.

3.3 Classification Accuracy

Based on the classification results, the values of accuracy, sensitivity, specificity, and F1-Score obtained are presented in Table 5.

Table 5. Classification Results Accuracy

Classification Methods	Number of Bootstrap	Accuracy	Sensitivity	Sensitivity	F1-Score
Ordinal Logistic Regression	Without bootstrap	0.700	0.729	0.595	0.655
Bagging Ordinal Logistic Regression	100	0.800	0.826	0.643	0.723
	500	0.850	0.872	0.726	0.793
	1000	0.750	0.767	0.619	0.685

Based on the classification results, the ensemble method is bagging logistic regression with 500 bootstrap is the best method for classifying stunting cases in toddlers in Sumberputih Village because it has the highest accuracy, sensitivity, specificity, and F1-Score values. This is enough to show that Bagging Logistic Regression is the best classification method for imbalanced data in stunting cases.

3.4 Discussion

The results of logistics regression shows that Economic Condition, Health Services, and Environment influential significant on the Nutritional Status of Toddlers. This is in line with previous research which states that the family's economic condition is one of the factors that can influence stunting conditions. Economic conditions, especially household income, are one of the factors that can influence the incidence of stunting in toddlers [11][12]. Good family economic conditions will increase the opportunity for a better standard of living so that it can increase the family's nutritional consumption which in turn can increase the opportunity to prevent stunting in toddlers. Unfavorable environmental conditions allow various diseases to occur. Diseases caused by unfavorable environments such as poor hygiene and sanitation can affect children's growth. If this happens for a long time and is not accompanied by adequate intake, then this can happen cause stunting [13]. Environmental sanitation is a factor that can influence stunting in toddlers. Unhealthy environmental sanitation can influence and increase the risk of stunting [14][15]. Besides that, [16] state that the influence of children's diet, especially the provision of animal protein, has a significant effect on the incidence of stunting in toddlers. This is also in line with research [17] which states that household socio-economic status, access to health services, children's consumption of complementary foods, and environmental conditions such as the availability of toilets and clean water are determining factors for stunting in children in Indonesia.

Based on the classification results, it also shows that the accuracy, sensitivity, specificity, and F1-Score values in bagging logistic regression are higher than the values in ordinary logistic regression. This shows that the ensemble bagging method is better used to classify data with unbalanced proportions compared to ordinary logistic regression. The best method was obtained, namely the bagging logistic regression using 500 bootstraps. In the other hand, an accuracy value of 0.850 was obtained, which means that the Bagging logistic regression using 500 bootstrap methods can correctly predict 85 % of the total data. Sensitivity can measure the extent to which the model correctly detects positive cases. The higher the sensitivity, the fewer positive cases are missed (false negatives), which means the model has a good ability to identify positive cases. With a sensitivity value of 87.2 %, it can be concluded that the Bagging logistic regression using 500 bootstrap method is able to classify normal toddlers very well. Meanwhile, Specificity measures as far as the model can go differentiate with Correct between case negative and positive. Specificity can measure the extent to which the model can correctly distinguish positive and negative cases. The higher the specificity, the fewer negative cases are detected as positive cases (false positives). With a specificity value of 7 2.6 %, it can be concluded that the Bagging logistic regression using 500 bootstrap methods can classify stunted toddlers very well. F1-Score measures the balance between a model's ability to correctly identify positive cases (sensitivity) and its ability to avoid classifying negative cases as positive cases (precision). The higher the F1-Score value, the better the model performs in achieving this balance. With an F1-Score value of 79.3%, this value is stable enough to classify nutritional status.

4. CONCLUSIONS

Variables that affect the nutritional status of toddlers include Economic Conditions, Health Services, and the Environment. Classification performance is seen based on the accuracy, sensitivity, specificity, and F1-Score values calculated from the three-category confusion matrix. Bagging logistic regression is better at classifying data with unbalanced proportions compared to ordinary logistic regression. The best classification method for classifying edit cases with unbalanced data proportions is Bagging logistic regression using 500 bootstrap. This is because the classification accuracy value produced by Bagging logistic regression using 500 bootstrap is the highest with accuracy, sensitivity, specificity, and F1-Score values of 85%, 87.2%, 72.6%, and 79.3%, respectively. In further research, simulation studies are expected to be carried out with various unbalanced proportions and many different samples.

REFERENCES

[1] J. Jajang, N. Nurhayati, and S. J. Mufida, "ORDINAL LOGISTIC REGRESSION MODEL AND CLASSIFICATION TREE ON ORDINAL RESPONSE DATA," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 16, no. 1, pp. 075–082, Mar. 2022, doi: 10.30598/barekengvol16iss1pp075-082.

[2] R. D. Fitriani, H. Yasin, and Tarno, "PENANGANAN KLASIFIKASI KELAS DATA TIDAK SEIMBANG DENGAN RANDOM OVERSAMPLING PADA NAIVE BAYES (Studi Kasus: Status Peserta KB IUD di Kabupaten Kendal)," *Jurnal Gaussian*, vol. 10, no. 1, pp. 11–20, 2021.

[3] G. Ngo, R. Beard, and R. Chandra, "Evolutionary bagging for ensemble learning," *Neurocomputing*, vol. 510, pp. 1–14, Oct. 2022, doi: 10.1016/j.neucom.2022.08.055.

[4] F. Aziz, "Peningkatan performance Logistic Regression menggunakan teknik Ensemble Bagging pada kasus Credit Scoring," *Journal of System and Computer Engineering (JSCE)*, vol. 1, no. 1, pp. 21–27, Jul. 2020, doi: 10.47650/jsce. v1i1.75.

[5] E. Rahmawati and C. Agustina, "Implementasi Teknik Bagging untuk Peningkatan Kinerja J48 dan Logistic Regression dalam Prediksi Minat Pembelian Online," *Jurnal Teknologi Informasi dan Terapan*, vol. 7, no. 1, pp. 16–19, Jun. 2020, doi: 10.25047/jtit. v7i1.123.

[6] L. M. Cendani and A. Wibowo, "Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes," *JURNAL MASYARAKAT INFORMATIKA*, vol. 13, no. 1, pp. 33–44, May 2022, doi: 10.14710/jmasif.13.1.42912.

[7] A. B. Astuti, A. Efendi, S. Astutik, and E. Sumarminingsih, *Analisis Data Kategorik Menggunakan R: Teori dan Aplikasinya pada Berbagai Bidang*. Malang: UB Press, 2020.

[8] A. Efendi, R. Fitriani, H. I. Naufal, and B. Rahayudi, "ENSEMBLE ADABOOST IN CLASSIFICATION AND REGRESSION TREES TO OVERCOME CLASS IMBALANCE IN CREDIT STATUS OF BANK CUSTOMERS," *J Theor Appl Inf Technol*, vol. 98, no. 17, pp. 3428–3437, 2020.

[9] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors)," *The Annals of Statistics*, vol. 28, no. 2, Apr. 2000, doi: 10.1214/aos/1016218223.

[10] A. A. R. Fernandes and Solimun, "Menelisik Faktor-Faktor Penyebab Stunting pada Anak di Kecamatan Wajak: Integrasi Cluster dengan Path Analysis dengan Pendekatan Statistika dan Sains Data," 2023.

[11] D. Wahyuni and R. Fitrayuna, "PENGARUH SOSIAL EKONOMI DENGAN KEJADIAN STUNTING PADA BALITA DI DESA KUALU TAMBANG KAMPAR," *PREPOTIF Jurnal Kesehatan Masyarakat*, vol. 4, no. 1, pp. 20–26, 2020.

[12] R. A. Utami, A. Setiawan, and P. Fitriyani, "Identifying causal risk factors for stunting in children under five years of age in South Jakarta, Indonesia," *Enferm Clin*, vol. 29, pp. 606–611, Sep. 2019, doi: 10.1016/j.enfcli.2019.04.093.

[13] N. B. Brahmana, V. S. Manalu, D. Nababan, T. R. Sinaga, and F. L. Tarigan, “Faktor-Faktor yang Berhubungan dengan Kejadian Stunting Pada Balita di Desa Marbun Tonga Marbun Dolok Kecamatan Hasundutan Tahun 2021,” *Journal of Healthcare Technology and Medicine*, vol. 7, no. 2, pp. 2615–109, 2021.

[14] L. Cameron, C. Chase, S. Haque, G. Joseph, R. Pinto, and Q. Wang, “Childhood stunting and cognitive effects of water and sanitation in Indonesia,” *Econ Hum Biol*, vol. 40, p. 100944, Jan. 2021, doi: 10.1016/j.ehb.2020.100944.

[15] M. R. Putri, T. Y. Handayani, and D. P. Sari, “Pengaruh Sanitasi Lingkungan Terhadap Kejadian Stunting Pada Balita,” *JURNAL KESEHATAN MERCUSUAR*, vol. 5, no. 1, pp. 63–68, Apr. 2022, doi: 10.36984/jkm.v5i1.260.

[16] T. Mahmudiono, S. Sumarmi, and R. R. Rosenkranz, “Household dietary diversity and child stunting in East Java, Indonesia,” *Asia Pac J Clin Nutr*, vol. 26, no. 2, pp. 317–325, Jan. 2017, [Online]. Available: https://search.informit.org/doi/10.3316/ielapa.688058173877148

[17] T. Beal, A. Tumilowicz, A. Sutrisna, D. Izwardy, and L. M. Neufeld, “A review of child stunting determinants in Indonesia,” *Matern Child Nutr*, vol. 14, no. 4, Oct. 2018, doi: 10.1111/mcn.12617.