CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 571-587
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Modeling Risk Factors of Acute Respiratory Infections using Logistic Regression and Multivariate Adaptive Regression Splines

Ardi Kurniawan*, Nathania Fauziah, Arinda Mahadesyawardani , Syifa' Azizah Putri Gunawan, and Aurellia Calista Anggakusuma

Statistics Study Program, Department of Mathematics, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia

Abstract

Acute Respiratory Infections (ARI) remain a leading cause of morbidity among children, particularly in regions with limited healthcare access. This study aimed to model the risk factors of ARI in children using Binary Logistic Regression and Multivariate Adaptive Regression Splines (MARS). Using secondary data from Southeast Aceh, seven predictor variables were analyzed, including maternal characteristics, breastfeeding status, and household conditions. Both models were statistically significant in identifying key predictors. Logistic regression showed superior performance with 86.96% accuracy, 85.00% precision, 91.89% recall, 81.25% specificity, and 88.30% F1-score. In contrast, MARS achieved a higher recall (97.30%) but lower specificity (62.50%), indicating higher sensitivity but a greater likelihood of false positives. Exclusive breastfeeding, home ventilation, and housing density were significant predictors in both models. Overall, logistic regression was found to be the more reliable and interpretable method, offering better balance in classification metrics. These findings support the use of logistic regression for identifying ARI risk factors in similar contexts and contribute to improved data-driven public health strategies aimed at reducing ARI incidence among vulnerable populations.

Keywords: Acute Respiratory Infections; Binary Logistic Regression; Multivariate Adaptive Regression Splines; Classification Model; children.

1 Introduction

Acute Respiratory Infections (ARI) are contagious respiratory diseases that represent a leading cause of death worldwide, accounting for 2.6 million deaths annually [1]. These infectious diseases are caused by viruses, fungi, bacteria, or a combination of all three, and are generally characterized by symptoms such as fever, sore throat, cough, and flu-like conditions [2]. According to reports from the World Health Organization (WHO), approximately 3.5% of the total global burden of diseases is attributed to ARI. Globally, ARI are responsible for about 15% of all deaths among children, with the majority of these cases occurring in low- and middle-income countries. Respiratory infections contribute to 33% of total deaths among toddler, particularly in Southeast Asia [3].

*Corresponding author. E-mail: ardi-k@fst.unair.ac.id

Submitted: May 15, 2025 Reviewed: June 10, 2025 Accepted: June 27, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.33833

As a developing country and part of Southeast Asia, Indonesia is not exempt from the burden of ARI. Data from the Ministry of Health show that in 2023, there was an increase in ARI cases, reaching 285,623 in July alone, with a prevalence rate of 3.5 per 100 toddlers. The highest prevalence by age group was recorded among children aged 1 to 4 years [4]. children are especially vulnerable to ARI due to their still developing immune systems. This condition is exacerbated by various factors, including environmental conditions, socioeconomic disparities, and unequal access to healthcare services across regions. One area that deserves special attention is Aceh Tenggara Regency, which is geographically located in a hilly region with limited access in several sub-districts [5]. According to the 2024 Aceh Province Health Profile, pneumonia cases among toddler accounted for 10% of the total under-five population in the region, totaling 45,280 cases. Meanwhile, the number of ARI cases in Aceh Tenggara Regency is relatively high compared to several other regencies in Aceh Province.Although the Ministry of Health has introduced programs such as the Integrated Management of Childhood Illness (IMCI) and the revitalization of Posyandu (integrated health posts) to detect and manage ARI cases, operational barriers in remote areas have hindered the effectiveness of these initiatives in reaching vulnerable groups, particularly toddler in Southeast Aceh

The incidence of acute ARI in toddler is influenced by a variety of interconnected and complex risk factors. A study conducted by Titi Saparina L and Rasni Intan [6] stated that the size of ventilation openings and housing density have a significant relationship with the incidence of ARI in toddler. Furthermore, [7] found that children who live in highly crowded rooms are at greater risk of developing ARI. Research by [8] also showed that exclusive breastfeeding is associated with a reduction in infectious diseases, including ARI among toddler. Exclusive breastfeeding provides natural antibodies that help strengthen the child's immune system. Socioeconomic factors such as low maternal education, unstable parental employment, and limited access to healthcare facilities, also contribute to increased vulnerability to this disease. [9] found that mothers with good knowledge and positive attitudes toward child health are more likely to recognize ARI symptoms early and take appropriate preventive actions. Therefore, maternal education also has an indirect impact by improving health literacy and access to medical services.

A data driven approach is becoming increasingly important for detecting and understanding risk factors in greater depth. Binary logistic regression as a parametric statistical method is considered more effective than standard linear approaches in identifying health risk factors, as it can accommodate various types of predictor variables, both categorical and numerical, making it more adaptable to the complex nature of health related data [10]. Meanwhile, Multivariate Adaptive Regression Splines (MARS) is a nonparametric method that builds regression models using combinations of basis functions, enabling it to capture nonlinear relationships and interactions between variables without rigid prior assumptions [11].

A study by [12] applied binary logistic regression to ARI cases in toddlers, analyzing the influence of knowledge, environmental, and behavioral factors at a health center in Southeast Aceh. However, the model was not explicitly presented, and classification metrics were not included, limiting its role as a predictive tool. This study addresses those gaps by using the same dataset to construct an explicit logistic regression model with performance evaluation, and compares it to a nonparametric method, namely Multivariate Adaptive Regression Splines (MARS). The use of both approaches is supported by [13], who compared logistic regression and MARS in modeling hypertension risk in Indonesia, showing that MARS can serve as an effective alternative in health-related risk modeling.

The novelty of this study lies in the application of the MARS (Multivariate Adaptive Regression Splines) method to model ARI risk factors in children. This nonparametric technique has not been previously applied to this particular health issue in Southeast Aceh. In addition, this study reconstructs the binary logistic regression model using an existing dataset and includes classification performance metrics, which were not presented in previous research.

The main contribution of this research is the comparative evaluation of parametric (logistic regression) and nonparametric (MARS) approaches in identifying ARI risk factors, which provides a more comprehensive understanding of their predictive capabilities and interpretability. The study highlights that logistic regression yields a better balance in classification metrics, whereas MARS reveals key interaction effects between predictor variables. These insights support the development of data-driven public health interventions and contribute to the achievement of SDG Target 3.2 aimed at reducing preventable child mortality by 2030.[14].

2 Methods

This section presents an overview of the data, modeling methods, and evaluation procedures, structured to support a clear understanding of the analytical approach used.

2.1 Data Source and Variables

This study utilizes secondary data obtained from the thesis by [12]. This study employs variables suspected to be risk factors for the incidence of ARI in toddler at the Deleng Pokhkisen Health Center in Aceh Tenggara District. A detailed description of the response and predictor variables used in this study is presented in Table Table 1 below.

Table 1: Research Variables

Variable Type	Variable Name	Description	Data Scale
Y	ARI Incidence in children	1: Suffering from ARI 0: Not Suffering from ARI	Nominal
X_1	Mother's Age	0: 26–32 years 1: 18–25 years	Nominal
X_2	Mother's Education	1: No Schooling 2: Elementary School 3: Junior High School 4: Senior High School 5: Higher Education	Ordinal
X_3	Mother's Knowledge	0: Poor 1: Good	Nominal
X_4	Mother's Attitude	0: Negative 1: Positive	Nominal
X_5	Exclusive Breastfeeding	0: Not Given 1: Given	Nominal
X_6	House Ventilation	0: Substandard 1: Standard	Nominal
X_7	Housing Density	0: Dense 1: Not dense	Nominal

2.2 Binary Logistic Regression

Logistic regression employs the logit function to estimate the probability of an event occurring [15]. In logistic regression, it is assumed that binary responses are independent and follow a binomial distribution. The probability of success for the i-th observation, given the predictor variables x_i, is modeled using a binomial distribution as expressed in Eq. 1.

The parameters in Eq. 1 are estimated using Maximum Likelihood Estimation (MLE), which identifies the values of the coefficients that maximize the likelihood of observing the given sample data.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)} \tag{1}$$

The interpretation of β in logistic regression refers to the log odds, which is obtained by exponentiating the coefficient e^{β_i} [16]. The odds in a binary logistic model are defined as the ratio between the probability of success and the probability of failure.

The odds ratio (OR) quantifies how the odds of the outcome change with a one-unit increase in the predictor variable. An OR greater than 1 indicates an increased risk, whereas an OR less than 1 indicates a decreased risk, holding other variables constant. For a binary predictor variable, the odds ratio comparing individuals with $x_i = 1$ to those with $x_i = 0$ is defined as:

$$OR = \frac{odds_{x_i=1}}{odds_{x_i=0}} = \exp(\beta_i) \tag{2}$$

If x_i is a continuous variable, then the odds ratio for an increase of m units in x_i is given by:

$$OR = \exp(m \cdot \beta_i) \tag{3}$$

These expressions allow for a more interpretable assessment of how each predictor influences the probability of the outcome. To assess the simultaneous influence of the predictors, the Likelihood Ratio Test (G-test) is employed at a 5% significance level. The hypotheses are formulated as follows:

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$
H_1: At least one $\beta_i \neq 0$ for $i = 1, 2, 3, 4, 5, 6, 7$

The G-test statistic is defined as:

$$G = -2 \log \left(\frac{L_0}{L_1} \right) \tag{4}$$

The null hypothesis H_0 is rejected if $G > \chi^2_{(\alpha;p-1)}$ or the p-value $< \alpha$, indicating that the predictor variables jointly have a significant effect on the incidence of ARI in toddler. To test the partial effect of each predictor variable, the Wald test is applied. The hypotheses are:

$H_0 : \beta_i = 0$ for $i = 1, 2, 3, 4, 5, 6, 7$
$H_1 : \beta_i \neq 0$ for $i = 1, 2, 3, 4, 5, 6, 7$

The Wald test statistic is given by Eq. 5:

$$W = \left(\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \right)^2 \tag{5}$$

Here, $\hat{\beta}_i$ represents the estimated parameter for the ith variable and $SE(\hat{\beta}_i) = \sqrt{(\sigma^2(\beta_i))}$ denotes the standard error associated with that estimate. The null hypothesis is rejected if $W \geq Z_{\alpha/2}$ or the p-value $< \alpha$, indicating that the i predictor has a statistically significant effect on the probability of ARI occurrence in toddler.

The model's fit in logistic regression is assessed using the Hosmer–Lemeshow test, which compares predicted and observed outcomes to evaluate Goodness of Fit (GoF). A model is considered well-fitted if it meets this criterion. The test statistic is calculated using Eq. 6:

$$\hat{C} = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i(1 - E_i)} \tag{6}$$

Where O_i denotes the observed count in the ith group, E_i is the expected count in the same group, and g refers to the total number of groups. The model is considered to have a good fit with the data when the test statistic satisfies $\hat{C} < \chi^2_{(\alpha,g-2)}$ or p-value $> \alpha$.

2.3 Multivariate Adaptive Regression Splines (MARS)

MARS is a nonparametric regression method developed by Friedman in 1991, as expressed in Eq. 7 [17].

$$y_i = a_0 + \sum_{m=1}^{M} a_m B_{mi}(x,t) + \varepsilon_i \tag{7}$$

In this study, the dependent variable is binary, meaning it consists of two categories. Therefore, the outcome can be interpreted as the probability of a particular event occurring. To identify the predictor variables that significantly influence the binary outcome, one may analyze the odds ratio [13]. The probability that $Y = 1$ is given in Eq. 8.

$$P(Y = 1 \mid X_1, X_2, ..., X_p) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}} \tag{8}$$

Accordingly, this probability can also be expressed using the basis functions of MARS, as shown in Eq. 9 below:

$$P = \frac{\exp(BF_0 + BF_i(X_i))}{1 + \exp(BF_0 + BF_i(X_i))} \tag{9}$$

where i denotes the basis function components obtained from the optimal MARS model.

In the MARS algorithm, a basis function (BF) captures either the main effect of a predictor or its interaction with others. Its piecewise linear form around a knot τ. is typically represented by reflected pairs, as shown in Eq. 10.

$$c^{+}(x,\tau) = \max(0, x - \tau), \quad c^{-}(x,\tau) = \max(0, \tau - x) \tag{10}$$

These functions capture directional changes in the predictor's influence on the response, centered around the knot location τ [18]. When modeling interactions between predictors, MARS constructs a basis function by multiplying several piecewise linear terms corresponding to different variables. For a given observation $(\bar{x}_i, y_i)$, where $i = 1, 2, ..., N$, the interaction-based basis function for the m-th term is represented as Eq. 11.

$$B_m(x^{(m)}) := \prod_{j=1}^{K_m} \left[s_{k_j^m} \cdot (x_{k_j^m} - \tau_{k_j^m}) \right]_+ \tag{11}$$

In this expression, K_m is the number of truncated linear terms in the m-th basis function. The variable $x_{k_j^m}$ represents the input for the j-th component, while $\tau_{k_j^m}$ denotes the corresponding knot that activates the function. The coefficient $s_{k_j^m} \in \{+1, -1\}$ indicates the direction of the piecewise term. The operator $[q]_+ := \max(0, q)$ ensures only positive contributions are included. This interaction basis function is active only when all variables meet their knot conditions, thus capturing the joint effect of multiple predictors on the response [18].

The modeling process in MARS involves forward selection and backward elimination, which are iteratively performed until the Generalized Cross Validation (GCV) score is minimized, indicating the most optimal model. The GCV criterion is defined in Eq. 12 [19].

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2}{\left[1 - \frac{C(M)}{n}\right]^2} \tag{12}$$

As explained by [19], significance testing is also conducted in MARS to examine parameter significance and assess model adequacy. A simultaneous test is performed using the F-statistic shown in Eq. 13:

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(N - M - 1)} \tag{13}$$

In this study, the hypotheses for the simultaneous test are as follows:

H_0: $a_1 = a_2 = \cdots = a_M = 0$ (The model is not significant)
H_1: At least one $a_m \neq 0$ (the model is significant)

The null hypothesis is rejected if $F_{\text{stat}} > F_\alpha(M, N - M - 1)$ or p-value $< \alpha$, indicating that at least one predictor variable significantly affects the response variable. For partial significance testing of individual parameters, $t - test$ shown in Eq. 14:

$$t = \frac{\hat{a}_m}{\sqrt{\operatorname{var}(\hat{a}_m)}} \tag{14}$$

The hypotheses for the partial test are:

H_0: $a_m = 0$ (basis function $B_m(x)$ has no effect on the model)
H_1: $a_m \neq 0$ (basis function $B_m(x)$ has a significant effect on the model)

The null hypothesis is rejected if $t > t_{\alpha/2(N-M-1)}$ or p-value $< \alpha$.

2.4 Evaluation Matrix

The evaluation matrix serves as a tool to measure how well a classification model performs in identifying specific factors. In health-related studies, typical performance indicators include accuracy, precision, recall, and the F1-score. These indicators are typically derived from the confusion matrix, as illustrated in Table 2 below:

Table 2: Classification evaluation metrics and their formulas

Classification Evaluation	Calculation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

where:
TP (True Positive): Positive cases that were accurately identified as positive.
TN (True Negative): Negative cases that were accurately identified as negative.
FP (False Positive): Negative cases that were mistakenly classified as positive.
FN (False Negative): Positive cases that were mistakenly classified as negative.

2.5 Research Stage

1. The factors influencing ARI were modeled using binary logistic regression with *SPSS* software, following these steps:
   (a) Testing the independence among predictor variables.
   (b) Estimating logistic regression parameters and evaluating their significance both simultaneously using the Likelihood Ratio test and individually using the Wald test.
   (c) Evaluating the overall model fit using the Hosmer–Lemeshow goodness-of-fit test.
   (d) Interpreting the model outcomes and calculating epidemiological association measures.
2. The factors influencing ARI were modeled using MARS with the *MARS* software, through the following procedure:

(a) Seven predictor variables were used, with the number of basis functions limited to 14, 21, and 28. The maximum number of interactions was set to 1 and 2, while the minimum number of observations between knots was set at 0, 1, and 2.

(b) The optimal MARS model was selected based on the lowest Generalized Cross Validation (GCV), the minimum Mean Squared Error (MSE), and the highest R-squared (R^2) values from various combinations of basis functions, interaction levels, and knot intervals.

(c) A binary MARS model was built using the parameter estimates from the previously selected best basis function model.

(d) The significance of the MARS model parameters was evaluated to assess model fit using both simultaneous and partial regression coefficient tests.

3. Comparing the classification performance of the binary logistic and MARS models based on evaluation metrics.

3 Results and Discussion

The results are presented in several subsections, covering descriptive statistics, model construction using logistic regression and MARS, evaluation, and a comparison to identify the most effective method.

3.1 Descriptive Analysis of Data

The proportion of children suffering and not suffering from ARI is presented in Figure 1 below.

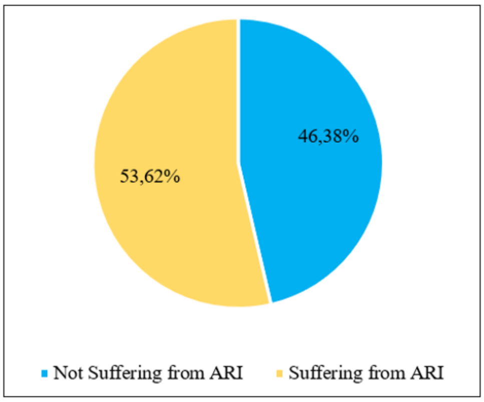

Figure 1: Distribution of children Based on ARI Status

Based on the distribution in Figure 1, indicates that more than half of the observed children experienced ARI, which may be influenced by various environmental and maternal factors. These proportions provide an overview of the prevalence of ARI in the studied population and justify further investigation into its risk factors.

Figure 2 illustrates the distribution of ARI cases in children based on seven predictor variables. The figure reveals that higher incidences of ARI are observed among children of younger mothers (X_1), mothers with lower education levels (X_2), poor maternal knowledge (X_3), and negative maternal attitudes (X_4). Furthermore, children who were not exclusively breastfed (X_5), lived in homes with inadequate ventilation (X_6), and resided in areas with high housing density (X_7) also exhibited a noticeably higher proportion of ARI. These patterns suggest that both maternal factors and living conditions play a critical role in the risk of developing acute respiratory infections in toddler.

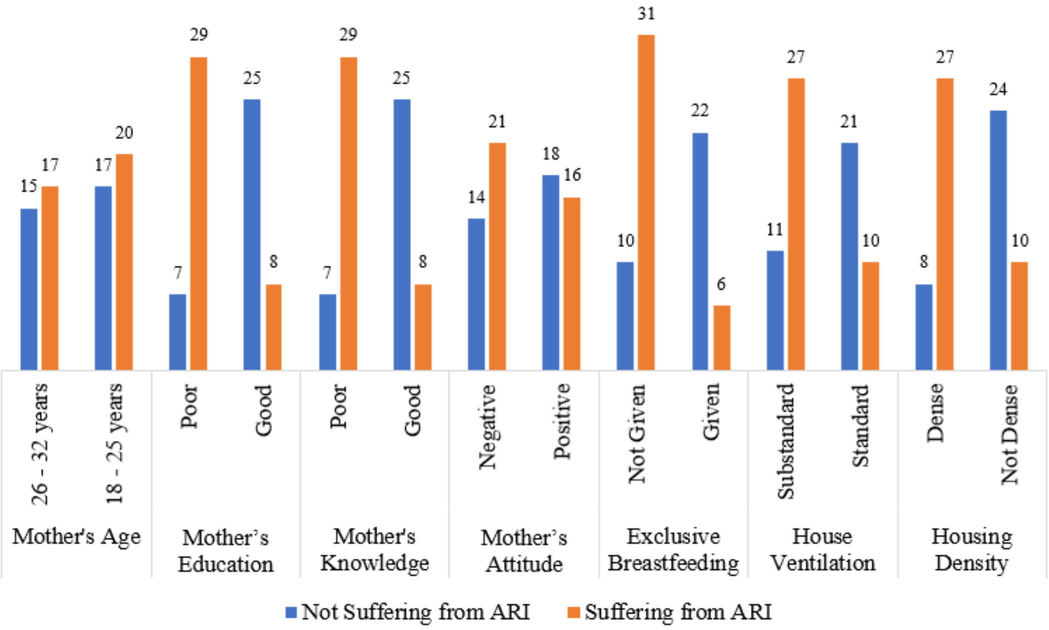

Figure 2: Distribution of ARI Status by Predictor Variables

3.2 Binary Logistic Regression Modelling

3.2.1 Multicollinearity Test

Multicollinearity testing is essential in binary logistic regression to ensure that predictor variables are not highly correlated, as such correlation can inflate the variance of estimated coefficients, reduce statistical power, and lead to unstable interpretations [20]. A key assumption in logistic regression is the independence among predictors; violation of this results in indistinguishable predictor effects. The *Variance Inflation Factor* (VIF) is commonly used to detect multicollinearity by quantifying the extent to which variance is increased due to intercorrelation. The hypotheses tested are as follows:

H_0: There is no multicollinearity among the predictor variables.
H_1: There is multicollinearity among the predictor variables.

The critical region is defined such that H_0 is rejected if $VIF > 10$. The results of the multicollinearity test are presented in Table 3 below:

Table 3: Multicollinearity Test Results for Factors Affecting ARI Cases in children

Predictor Variable	Tolerance	VIF
Mother's Age	0.813	1.230
Mother's Education	0.459	2.179
Mother's Knowledge	0.473	2.114
Mother's Attitude	0.514	1.944
Exclusive Breastfeeding	0.782	1.278
Home Ventilation	0.795	1.258
Household Density	0.800	1.249

Based on Table 3, it can be concluded that all VIF values are below the threshold of 10. Therefore, we fail to reject H_0, indicating that there is no multicollinearity among the predictor variables used in the model. This confirms that the independence assumption among predictors is met, and the regression estimates can be interpreted with greater confidence.

3.2.2 Significance Testing of Binary Logistic Regression Model

The significance test evaluates how much the predictor variables explain the variation in ARI incidence. Parameter testing is conducted simultaneously using the G test and partially using the Wald test. The results of the simultaneous test are shown in Table 4.

Table 4: Simultaneous Parameter Test

	Chi-Square	**df**	**P-Value**
Step	54.492	10	0.000
Block	54.492	10	0.000
Model	54.492	10	0.000

Based on Table 4, the p-value obtained is 0.000, which is less than $\alpha = 5\%$. Therefore, the decision is to reject H_0, and it can be concluded that there is a significant effect of the predictor variables on the incidence of ARI in toddler. Another side, the results of the partial parameter significance test are presented in Table 5.

Table 5: Partial Parameter Test

	B	**S.E.**	**Wald**	**df**	**P-Value**
Constant	23.502	28296.171	0.000	1	0.999
Mother's Age	1.234	0.919	1.801	1	0.180
Mother's Education (1)	-21.599	28296.171	0.000	1	0.999
Mother's Education (2)	-20.136	28296.171	0.000	1	0.999
Mother's Education (3)	-20.895	28296.171	0.000	1	0.999
Mother's Education (4)	-41.058	31289.017	0.000	1	0.999
Mother's Knowledge	-1.265	1.109	1.302	1	0.340
Mother's Attitude	0.852	0.894	0.909	1	0.340
Exclusive Breastfeeding	-2.497	1.027	5.908	1	0.015
Home Ventilation	-2.073	0.966	4.610	1	0.032
Housing Density	-2.208	0.955	5.351	1	0.021

Based on Table 5, it was found that the variables *exclusive breastfeeding*, *home ventilation*, and *housing density* have p-values less than α (5%), leading to the decision to reject H_0. Therefore, it can be concluded that these three variables have a partial effect on the incidence of ARI in toddler. Meanwhile, the other variables do not have a partial effect, as their p-values are greater than α (5%).

3.2.3 Goodness of Fit Test

A model fit test was conducted to assess the suitability of the binary logistic regression model, specifically to determine whether there is a significant difference between the observed and predicted values. The Hosmer–Lemeshow test was used for this evaluation. The results of the model fit test are presented in Table 6 below.

Table 6: Model Goodness-of-Fit Test Results

Chi-Square	**df**	**P − Value**
12.899	8	0.115

Based on Table 6, the *p-value* is 0.115, which is greater than α (5%). Therefore, we fail to reject H_0 and conclude that the binary logistic regression model fits the data adequately.

3.2.4 Coefficient of Determination Test

One approach to assess how well a logistic regression model fits the data is through the Nagelkerke R Square test. A higher value of this statistic indicates that a larger portion of the variance in the dependent variable is accounted for by the independent variables in the model. A value near 1 suggests strong explanatory capability, whereas a value closer to 0 implies weak explanatory power. The outcomes of the model's goodness-of-fit assessment using the Nagelkerke R Square are shown in Table 7.

Table 7: Determination Coefficient Test Results

-2 Log Likelihood	Cox & Snell R Square	Nagelkerke R Square
12.899	8	0.115

Based on Table 7, the Nagelkerke R Square value is 0.115. This indicates that the predictive variables in the model are able to explain 11.5% of the variation in the response variable. In other words, 88.5% of the variation in the response variable is influenced by other factors that are not included in this logistic regression model.

3.2.5 Parameter Estimation of Binary Logistic Regression Model

The model estimation was performed using analysis results that included only predictor variables with a statistically significant effect on the incidence of ARI among toddler. The coefficient estimates of these significant variables are shown Table 8 below.

Table 8: Binary Logistic Regression Model Estimation

Variable in the Equation	B	S.E	Wald	df	Sig.
Exclusive Breastfeeding	-2.598	0.781	11.062	1	0.001
Home Ventilation	-2.409	0.792	9.247	1	0.002
Housing Density	-1.985	0.715	7.705	1	0.006
Constant	3.292	0.810	16.525	1	0.000

Based on Table 8, the parameter values β used to construct the initial binary logistic regression model were obtained by considering the factors that influence the incidence of ARI in toddler. The logistic regression model obtained is:

$$\pi(x_i) = \frac{\exp(3.292 - 2.598x_5 - 2.409x_6 - 1.985x_7)}{1 + \exp(3.292 - 2.598x_5 - 2.409x_6 - 1.985x_7)} \tag{15}$$

Since Eq. 15 is nonlinear, a logit transformation is required to form the logistic regression model:

$$g(x) = 3.292 - 2.598x_5 - 2.409x_6 - 1.985x_7 \tag{16}$$

Based on the resulting model, it can be concluded that children who receive exclusive breastfeeding have approximately 92.6% lower odds of developing ARI compared to those who do not, based on an odds ratio of 0.074. Likewise, improvement in home ventilation and reduction in housing density are associated with 90.9% and 86.3% lower odds of ARI, respectively.

3.2.6 Classification Results of the Binary Logistic Regression Model

The previous estimation results were evaluated for classification performance in the form of the following confusion matrix shown in Table 9.

Table 9: Classification Results of the Binary Logistic Regression Model

Observed	Predicted	
	Not Suffering from ARI	Suffering from ARI
Not Suffering from ARI	26	6
Suffering from ARI	3	34

3.3 Multivariate Adaptive Regression Splines (MARS) Modeling

3.3.1 Best MARS Model

Before estimating the MARS model, the first step is to determine the optimal combination of parameters. This study used 7 predictor variables, with the number of basis functions (BF) set to 14, 21, and 28; the maximum interaction (MI) set to 1 and 2; and the minimum number of observations (MO) set to 0, 1, and 2. All possible combinations of BF, MI, and MO were evaluated to identify the best model configuration based on the lowest GCV and MSE values, and the highest R^2 value. The calculation results are presented in Table 10 below.

Table 10: GCV Calculation Results for Each Model Combination

Model	BF	MI	MO	GCV	MSE	R^2
1	14	1	0	0.173	0.142	0.323
2	14	1	1	0.180	0.137	0.292
3	14	1	2	0.180	0.137	0.292
4	**14**	**2**	**0**	**0.155**	**0.123**	**0.393**
5	14	2	1	0.155	0.123	0.393
6	14	2	2	0.155	0.123	0.393
7	21	1	0	0.181	0.142	0.292
8	21	1	1	0.192	0.137	0.273
9	21	1	2	0.181	0.142	0.292
10	21	2	0	0.155	0.123	0.393
11	21	2	1	0.155	0.123	0.393
12	21	2	2	0.155	0.123	0.393
13	28	1	0	0.192	0.142	0.248
14	28	1	1	0.205	0.155	0.201
15	28	1	2	0.192	0.142	0.248
16	28	2	0	0.155	0.123	0.393
17	28	2	1	0.155	0.123	0.393
18	28	2	2	0.155	0.123	0.393

Based on Table 10, several model parameter combinations produce identical evaluation results, notably models 4–6, 13–15, and 22–24. All of these models yield the minimum GCV value of 0.155, MSE of 0.123, and the highest R^2 value of 0.393, meeting the criteria for the best MARS model. According to the parsimony principle—preferring the simplest model among those with equivalent predictive performance. Model 4 is selected as the optimal model, with BF = 14, MI = 2, and MO = 0. This model not only demonstrates strong performance but is also structurally simpler, making it easier to interpret and apply.

3.3.2 Parameter Estimation of MARS Model

Based on the best MARS model previously obtained, estimation can be conducted to model ARI cases in toddler as shown in Table 11 below.

Table 11: Parameter Estimation Results of the MARS Model

Basis Function (BF)	Parameter Estimation
BF_0	0.840
$BF_3 = (X_7 = 1)BF_1$; where $BF_1 = (X_3 = 1)$	-0.553
$BF_8 = \max(0, X_2 - 1)BF_2$; where $BF_2 = (X_5 = 1)$	-0.117

Based on Table 11, the MARS model for estimating ARI cases in toddler at the Deleng Pokhkisen Health Center, Aceh Tenggara Regency is as follows:

$$Y = 0.840 - 0.553BF_3 - 0.117BF_8 \tag{17}$$

Based on Eq. 17, the MARS model indicates two main interaction terms occurring in BF_3 and BF_8, with the interpretation of each basis function described as follows:

a. Basis function $BF_3 = (X_7 = 1)(X_3 = 1)$ captures the interaction between housing density (X_7) and maternal knowledge about ARI (X_3). BF_3 equals 1 only if the child lives in a low-density household and the mother has good knowledge of ARI. With a negative coefficient of -0.553, this suggests that such a combination significantly reduces the risk of ARI in toddler at Deleng Pokhkisen Health Center.

b. Basis function $BF_8 = \max(0, X_2 - 1)(X_5 = 1)$ represents the interaction between maternal education level (X_2) and exclusive breastfeeding practice (X_5). BF_8 equals 1 when the mother has education above primary school and provides exclusive breastfeeding. Its coefficient of -0.117 implies a moderate reduction in ARI risk, although less impactful than BF_3.

Since the response variable is binary, the probability model for ARI occurrence is expressed as:

$$P(Y = 1) = \pi(x) = \frac{\exp(0.840 - 0.553BF_3 - 0.117BF_8)}{1 + \exp(0.840 - 0.553BF_3 - 0.117BF_8)} \tag{18}$$

Based on Eq. 18, the negative coefficients of BF_3 and BF_8 indicate that fulfillment of these conditions decreases the log-odds of ARI. Specifically, when $BF_3 = 1$, the logit decreases by 0.553, showing a substantial protective effect. When $BF_8 = 1$, the logit decreases by 0.117, indicating a milder, yet still beneficial effect.

3.3.3 Significance Testing of MARS Model

The following are the hypotheses formulated to test whether the set of basis functions significantly influences the response variable in the MARS model:

H_0: $a_3 = a_8 = 0$ (The model is not significant)
H_1: At least one $a_M \neq 0$, $m = 3$ and $m = 8$ (the model is significant)

At a significance level of 5%, the critical region for the above hypotheses is: reject H_0 if the test statistic $F > F_{\text{critical}}$ or if the p-value $< \alpha$.

Table 12: Simultaneous Significance Test Results for the MARS Model

F-Statistic	36.826
Standard Error of Regression	0.351
P-value	0.181332×10^{-10}
Residual Sum of Squares (RSS)	8.110
(MDF, NDF)	(2, 66)
Regression Sum of Squares (SSR)	9.050

Based on Table 12, it can be seen that the F-statistic is 36.826, which is greater than $F_{(2,66)} = 3.13592$. In addition, the p-value of 0.181332×10^{-10} is less than 0.05. Therefore, we reject H_0 and conclude that the model is statistically significant, indicating that at least one basis function with $\beta \neq 0$ significantly influences the incidence of ARI among toddler at the Deleng Pokhkisen Public Health Center in Aceh Tenggara Regency.

The next step is to perform a partial significance test to examine whether each basis function has a significant effect on the response variable in the MARS model. The hypotheses used are as follows:

H_0: $a_m = 0$ (basis function $B_m(x)$ has no effect on the model)
H_1: $a_m \neq 0$, $m = 3$ and 8 (basis function $B_m(x)$ has a significant effect on the model)

At a 5% significance level, H_0 is rejected if the the p-value $< \alpha$. The results of the partial significance tests for each basis function are as follows:

Table 13: Partial Significance Test Results for the MARS Model

Parameter	**Estimate**	**Standard Error**	**t-Ratio**	**P-value**
Constant	0.840	0.056	15.088	0.99201×10^{-3}
Basis Function 3	-0.553	0.099	-5.565	0.514515×10^{-6}
Basis Function 8	-0.117	0.030	-3.873	0.249783×10^{-3}

Based on Table 13, the results of the partial significance test for each basis function are as follows:

a. For basis function 3, the test statistic is $t = -5.565$ and the p-value is 0.514515×10^{-6}. Since the p-value is less than 0.05, we reject H_0 and conclude that basis function 3 significantly affects the incidence of ARI among toddler at the Deleng Pokhkisen Public Health Center in Aceh Tenggara Regency.

b. For basis function 8, the test statistic is $t = -3.873$ and the p-value is 0.249783×10^{-3}. Again, the p-value is below 0.05, leading us to reject H_0 and conclude that basis function 8 also significantly affects the incidence of ARI in the same setting.

3.3.4 Predictor Variable Importance Level

To explain the extent of contribution of each predictor variable to the probability model of ARI (Acute Respiratory Infection) incidence in children, the output of relative variable importance is viewed in terms of the importance level and cost of omission, as shown in Table 14 below.

Table 14: Relative Variable Importance Output

Variable	**Importance Level**	**Cost of Omission**
Mother's Knowledge	100.000	0.201
Household Density	100.000	0.201
Mother's Education Level	51.947	0.168
Exclusive Breastfeeding	51.947	0.168
Mother's Age	0.000	0.155
Mother's Attitude	0.000	0.155
House Ventilation	0.000	0.155

Based on Table 14, the following interpretation is obtained:

a. *Mother's knowledge* and *household density* are the most influential predictors, each with 100% importance and a cost of omission of 0.201, indicating a strong impact on ARI incidence. Removing either would significantly reduce model performance, making them essential to retain.

b. *Mother's education level* and *exclusive breastfeeding* have 51.947% importance and a cost of omission of 0.168. While less impactful than the top predictors, they still contribute meaningfully and should be maintained to reduce model error.

c. *Mother's age*, *attitude*, and *house ventilation* show 0% importance, with a cost of omission of 0.155. Removing them may be considered, as their exclusion would not significantly affect the model's effectiveness.

3.3.5 Classification Results of MARS Model

The evaluation of a MARS model involves several performance metrics that indicate how well the model can distinguish between categories in the dependent variable. Table 15 presents the classification results of MARS model.

Table 15: Classification Results of MARS Model

Observed	**Predicted**	
	Not Suffering from ARI	**Suffering from ARI**
Not Suffering from ARI	20	12
Suffering from ARI	1	36

3.4 Selection of the Best Regression Method

Based on Table 9 and Table 15, the accuracy, precision, recall, specificity, and F1-score are calculated to determine the best regression model between binary logistic regression and MARS. The comparison results are presented in Table 15 as follows:

Table 16: Comparison of Classification Accuracy between Logistic Regression and MARS

Method	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1-Score**
Binary Logistic Regression	86.96	85.00	91.89	81.25	88.30
MARS	81.16	75.00	97.30	62.50	84.70

Based on Table 16, It is clear that the Binary Logistic Regression model performs better than MARS in terms of accuracy, precision, specificity, and F1-score, while MARS shows better performance only in recall. This suggests that Logistic Regression gives a more balanced classification of acute respiratory infection (ARI) cases in children, with fewer classification errors, especially in avoiding false positive results. Whereas for MARS method is more sensitive in detecting ARI cases but has lower accuracy in predicting them correctly. Therefore, in this study, Binary Logistic Regression is considered the most suitable method for classifying ARI cases in children because it provides more consistent and reliable results. This study uses all sample data in its tests to utilize the available information in comparing models, considering the relatively small sample size. As a result, this approach limits the ability to generate Receiver Operating Characteristic (ROC) curves, which typically require separate testing data. Although this is a methodological decision, the evaluation results remain valid for the study's purpose, despite the limitation in assessing ROC-based performance.

4 Conclusion

This study aimed to model the risk factors of Acute Respiratory Infections (ARI) in children by comparing Binary Logistic Regression and Multivariate Adaptive Regression Splines (MARS). The findings showed that both models were statistically significant and capable of identifying key predictors of ARI. However, logistic regression demonstrated superior classification performance, with higher accuracy, precision, specificity, and F1-score. MARS model achieved a higher recall indicating better sensitivity to actual positive case. Overall, binary logistic regression emerged as the more reliable and interpretable method. Importantly, exclusive breastfeeding, home ventilation, and housing density emerged as significant predictors in the logistic regression model, while MARS highlighted the interaction effects between maternal knowledge and housing conditions, as well as between maternal education and breastfeeding. These findings offer new insights into how individual and environmental factors interact in influencing ARI incidence in children.

This study used the entire sample due to limitations in generating ROC-based evaluations. Therefore, future research is encouraged to use a larger dataset that can be divided into training and testing sets. This will enable a more comprehensive assessment of model performance using ROC curves and other validation metrics. Such improvements are especially important for health-related classifications, like detecting ARI cases in children under five, where model accuracy and generalizability are essential for practical use and decision-making in clinical or public health contexts.

CRediT Authorship Contribution Statement

Ardi Kurniawan: Conceptualization. **Nathania Fauziah:** Methodology, Formal Analysis. **Arinda Mahadesyawardani:** Software, Validation. **Syifa' Azizah Putri Gunawan:** Visualization, Data Curation. **Aurellia Calista Anggakusuma:** Writing–Review & Editing.

Declaration of Generative AI and AI-assisted technologies

AI-assisted technologies were used to assist in language refinement and initial draft structuring. All scientific content and interpretations were developed and validated by the authors.

Declaration of Competing Interest

The authors declare no competing interests.'

Funding and Acknowledgments

This research received no external funding. The authors would like to express their sincere gratitude to Department of Mathematics, Faculty of Science Technology, Universitas Airlangga for the academic and institusional support provided throughout the reseacrh process.

Data Availability

The dataset used in this study was obtained from the Master's thesis by Rahman Sabri (2019), titled *"Faktor yang Memengaruhi Balita terhadap Penyakit ISPA di Puskesmas Deleng Pokhkisen Kabupaten Aceh Tenggara"*, submitted to the Master of Public Health Program, Faculty of Public Health, Helvetia Institute of Health, Medan, Indonesia [1]."

References

[1] X. Yu, H. Wang, S. Ma, W. Chen, L. Sun, and Z. Zou, "Estimating the global and regional burden of lower respiratory infections attributable to leading pathogens and the protective effectiveness of immunization programs," *International Journal of Infectious Diseases*, vol. 149, p. 107 268, 2024. DOI: 10.1016/j.ijid.2024.107268.

[1] https://helvetia.ac.id/

[2] R. Mirino, D. Dary, and R. Tampubolon, “Identification of factors causing acute respiratory infection (ari) of under-fives in community health center work area in north jayapura sub-district,” *Journal of Tropical Pharmacy and Chemistry*, vol. 6, no. 1, pp. 15–20, 2022. DOI: 10.30872/j.trop.pharm.chem.v6i1.214.

[3] H. Zhu, K. Huang, X. Han, *et al.*, “The burden of acute respiratory infection in children under 5 attributable to economic inequality in low- and middle-income countries,” *BMJ Global Health*, vol. 10, no. 3, pp. 1–9, 2025. DOI: 10.1136/bmjgh-2024-017409.

[4] S. Y. Nurjamillah and C. M. Dwiriani, “Status gizi dengan kejadian ispa balita di wilayah kerja puskesmas unyur kota serang sebelum dan selama pandemi covid-19,” *Jurnal Ilmu Gizi dan Dietetik*, vol. 1, no. 2, pp. 95–102, 2022. DOI: 10.25182/jigd.2022.1.2.95-102.

[5] M. Irwansyah, S. Zuliansyah, and I. Hasan, “Sustainable landscape for high urban temperature mitigation in the disaster-prone coastal city of banda aceh, indonesia,” in *IOP Conference Series: Earth and Environmental Science*, vol. 630, IOP Publishing, 2021, p. 012 010. DOI: 10.1088/1755-1315/630/1/012010.

[6] R. Intan, “Relationship of the physical environment with the incidence of ari in toddlers,” *Miracle Journal of Public Health*, vol. 4, no. 2, pp. 176–186, 2021. DOI: 10.36566/mjph/Vol4.Iss2/268.

[7] E. Saputri, P. Eka Sudiarti, and Z. R. Z., “Hubungan kepadatan hunian kamar dan jenis bahan bakar memasak dengan kejadian ispa pada balita di desa pulau rambai wilayah kerja upt puskesmas kampa tahun 2023,” *Jurnal Ners*, vol. 7, no. 2, pp. 20 234–21 841, 2023. DOI: /10.31004/jn.v7i2.16997.

[8] A. Intiyati, R. D. Y. Putri, I. S. Edi, *et al.*, “Correlation between exclusive breastfeeding, complementary feeding, infectious disease with wasting among toddlers: A cross-sectional study,” *Amerta Nutrition*, vol. 8, no. 2SP, pp. 1–8, 2024. DOI: 10.20473/amnt.v8i2SP.2024.1-8.

[9] I. Yasmin, I. A. Pramesty, T. T. Affandi, and Y. Naldi, “Hubungan antara tingkat pengetahuan, tingkat pendidikan ibu, serta status gizi balita terhadap kejadian infeksi saluran pernapasan akut (ispa) pada balita di puskesmas kesunean kota cirebon jawa barat,” *Jurnal Kedokteran & Kesehatan*, vol. 5, no. 1, pp. 1–8, 2024. Available online.

[10] E. T. Yuniarsih, M. Salam, M. H. Jamil, and A. N. Tenriawaru, “Determinants determining the adoption of technological innovation of urban farming: Employing binary logistic regression model in examining rogers’ framework,” *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 10, no. 2, p. 100 307, 2024. DOI: 10.1016/j.joitmc.2024.100307.

[11] J. Sharma and S. K. Mitra, “Developing a used car pricing model applying multivariate adaptive regression splines approach,” *Expert Systems with Applications*, vol. 236, no. 1, p. 121 277, 2024. DOI: 10.1016/j.eswa.2023.121277.

[12] R. Sabri, I. Effendi, and N. Aini, “Factors affecting the level of ispa disease in children in deleng pokhkisen health center aceh tenggara district,” *Contagion :Scientific Periodical of Public Health and Coastal Health*, vol. 1, no. 2, pp. 69–82, 2019. DOI: 10.30829/contagion.v1i2.6883.

[13] N. Chamidah, A. T. Hendrawan, F. S. Ardiyanto, M. S. Hammami, N. Izzah, and S. N. Hariadi, “Modeling hypertension disease risk in indonesia using multivariate adaptive regression spline and binary logistic regression approaches,” *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 18, no. 4, pp. 2217–2230, 2024. DOI: 10.30598/barekengvol18iss4pp2217-2230.

[14] J. E. Lawn, Z. A. Bhutta, C. Ezeaka, and O. Saugstad, “Ending preventable neonatal deaths: Multicountry evidence to inform accelerated progress to the sustainable development goal by 2030,” *Neonatology*, vol. 120, no. 4, pp. 491–499, 2023. DOI: 10.1159/000530496.

[15] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd. New York: John Wiley & Sons, 2013. DOI: 10.1002/9781118548387.

[16] J. Wilson, K. Lorenz, and L. Selby, “Introduction to binary logistic regression,” in *Modeling Binary Correlated Responses*, ser. ICSA Book Series in Statistics, Cham: Springer, 2024. DOI: 10.1007/978-3-031-62427-8_1.

[17] H. Dukalang and B. Otok, “Modified partial least square structural equation model with multivariate adaptive regression spline: Parameter estimation technique and applications,” *MethodsX*, vol. 14, p. 103 381, 2025. DOI: 10.1016/j.mex.2025.103381.

[18] T. Ö. Kurtulmuş, F. Yerlikaya–Özkurt, and A. Askan, “Modeling of kappa factor using multivariate adaptive regression splines: Application to the western türkiye ground motion dataset,” *Natural Hazards*, vol. 120, no. 8, pp. 7817–7844, 2024. DOI: 10.1007/s11069-024-06535-y.

[19] S. Risambessy, S. Aulele, and F. Lembang, “Misclassification analysis of elementary school accreditation data in ambon city using multivariate adaptive regression spline,” *Jurnal Matematika, Statistika dan Komputasi*, vol. 18, no. 3, pp. 394–406, 2022. DOI: 10.20956/j.v18i3.19451.

[20] M. S. Sulaiman, M. M. Abood, S. K. Sinnakaudan, M. R. Shukor, G. Q. You, and X. Z. Chung, “Assessing and solving multicollinearity in sediment transport prediction models using principal component analysis,” *ISH Journal of Hydraulic Engineering*, vol. 27, no. sup1, pp. 343–353, 2019. DOI: 10.1080/09715010.2019.1653799.