JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 9, No. 3, July 2025, pp. 699-713

A Comparison of Multivariate Adaptive Regression Spline and Spline Nonparametric Regression on Life Expectancy in Indonesia

Bagas Shata Pratama[1], Suliyanto[1*], M. Fariz Fadillah Mardianto[1], Sediono[1]
[1]Department of Mathematics, Universitas Airlangga, Surabaya, Indonesia
suliyanto@fst.unair.ac.id

Article History:
Received : 22-01-2025
Revised : 14-04-2025
Accepted : 20-04-2025
Online : 01-07-2025

Keywords:
Life Expectancy;
Regression;
Multivariate Adaptive Regression Spline;
Spline Nonparametric Regression.

ABSTRACT

Life expectancy is a key indicator of a population's overall health and well-being. It also reflects the effectiveness of government efforts in improving public welfare. Despite various initiatives by both the government and society to improve life expectancy in Indonesia, significant disparities remain. This quantitative study aims to support these efforts by analyzing factors influencing life expectancy in Indonesia using data from the Indonesian Central Agency of Statistics (BPS) in 2023. A comparative analysis was conducted using two methods: Multivariate Adaptive Regression Spline (MARS) and Spline Nonparametric Regression. The results show that the MARS model outperforms the Spline model, achieving a lower Mean Squared Error (MSE) of 1.183 and a higher R-Square of 82.7%. Key variables significantly influencing life expectancy include access to decent housing, access to safe drinking water, per capita expenditure, and the Gini ratio. The findings not only confirm the presence of complex interactions among predictor variables effectively captured by the MARS method, but also contribute to the existing literature by emphasizing the importance of socioeconomic determinants in health outcomes. From a policy perspective, the results suggest that government strategies should prioritize improving access to basic needs and reducing inequality. These insights can guide targeted, data-driven interventions aimed at enhancing life expectancy in Indonesia.

https://doi.org/10.31764/jtam.v9i3.29413

A. INTRODUCTION

The level of public health in a country can be seen through one indicator, namely life expectancy. A higher life expectancy reflects an improved public health status and indicates the success of health development programs (Bangun, 2019). According to the (BPS, 2024), life expectancy refers to the estimated average number of years a newborn is expected to live, assuming that the pattern of mortality by age at birth is the same throughout the baby's life. Life expectancy also serves as a measure of how effectively the government is working to enhance public welfare and health. A low life expectancy indicates the need for health and social programs, such as improving environmental quality, improving nutrition, and alleviating poverty. The government continues to improve life expectancy through research, improving health resources, and strengthening health services (Erlyn et al., 2022). Efforts and programs that have been carried out to improve the degree of public health align with Sustainable Development Goals (SDGs), especially in the 3rd SDGs, namely healthy and prosperous life.

Indonesia ranks 134th in the world and 8th in ASEAN when viewed based on life expectancy from highest to lowest (Worldometer, 2024). Life expectancy in Indonesia has generally increased from year to year. According to data sourced from BPS, (2023), Indonesia's life expectancy in 2023 is 72.13 years old. When compared to 10 years ago, this value has increased by 1.73 years. The province with the highest life expectancy is D. I. Yogyakarta at 75.12 and the lowest is West Sulawesi at 66.01. Therefore, it is necessary to conduct regression analysis using the Multivariate Adaptive Regression Spline (MARS) and Spline Nonparametric Regression methods to identify factors that may be influential in formulating effective policies to increase life expectancy in Indonesia.

Regression analysis is a widely used method for modelling life expectancy in Indonesia. It is one of the most commonly applied techniques in statistical modelling. The primary purpose of regression analysis is to determine the relationship pattern between response variables and predictor variables (Ali & Younas, 2021). Regression analysis can be approached through two methods: parametric and nonparametric. The parametric approach assumes a predetermined model structure. Conversely, when there is no prior knowledge about the model or curve shape, the nonparametric approach is a suitable alternative. A notable advantage of nonparametric regression is its independence from specific curve shape assumptions, allowing for greater flexibility (Mahmoud, 2021).

In this study, two nonparametric regression methods, MARS and spline nonparametric regression, will be used to model life expectancy in Indonesia. MARS offers the benefit of managing high-dimensional data and can automatically select relevant interactions between variables using adaptive basis functions (Adiguzel & Cengiz, 2023). On the other hand, spline nonparametric regression is more suitable for smooth relationships and can be used to identify flexible curve patterns without the assumption of linearity. However, this method can be less efficient in high-dimensional data or high variable complexity (Pratama et al., 2024).

Previous research on life expectancy has been conducted by Mukrom et al., (2021) using the Robust Spatial Durbin Model Regression approach. The research examined variables believed to influence life expectancy, including the average years of schooling, the proportion of households implementing clean and healthy living practices, the number of integrated health service posts, the percentage of people living in poverty, and adjusted per capita expenditure. Research on life expectancy has also been conducted by Hasanah (2017) using panel data regression analysis. The study found that the gini ratio variable has a significant effect, both simultaneously and partially on life expectancy in Indonesia. Most previous studies relied on classical or spatial regression methods and did not explore variable interactions or high-dimensional flexibility. This study fills that gap by comparing two nonparametric regression methods MARS and Spline Regression. The novelty lies in MARS ability to capture interactions among predictor variables through adaptive basis functions, offering greater flexibility and interpretability that has not been extensively explored in earlier studies.

Based on the facts in previous studies, factors that are believed to affect life expectancy such as access to decent housing, access to decent drinking water, expenditure per capita, and gini ratio. The MARS and spline nonparametric regression approach is suitable for this data because this method is relatively flexible to investigate the pattern of relationships between variables without special assumptions (Wicaksono et al., 2014). Based on this description,

researchers are interested in modelling life expectancy in Indonesia using the MARS approach and Spline Nonparametric Regression. These findings can serve as valuable references for policymakers to develop targeted health and welfare interventions. By understanding the key determinants and their interactions, the government can craft more effective and equitable policies to improve the overall health status of the Indonesian population.

B. METHODS

This research uses a comparison of two nonparametric methods, namely MARS and Spline Nonparametric Regression. The use of this Nonparametric Method is done because the initial pattern of the data does not show a certain trend. MARS was chosen in this study because of its ability to model complex and high-dimensional relationships between variables, especially when there are interactions between predictors. Nonparametric Spline Regression was chosen as a comparison to MARS because of its ability to produce smooth and flexible curves on data with nonlinear relationships but low complexity. The selection of the best model in each method is done using two indicators, namely Generalized Cross Validation (GCV) and Mean Squared Error (MSE). GCV is used to select the model with the optimal balance between complexity and prediction performance, while MSE gives an idea of how well the model minimizes the prediction error. Furthermore, the best models from each method are compared based on the coefficient of determination (R^2) to assess how much of the data variation is explained by the model. This approach provides a comprehensive and objective evaluation of model performance.

1. Data

This research is a quantitative study that uses secondary data from 2023, obtained from publications by the Indonesian Central Agency of Statistics (BPS). The observation unit used is 34 provinces in Indonesia. The data used in this research have been cleaned and preprocessed by the data provider, so they are ready to be directly used for analysis. The data utilized in this research are summarized in the following Table 1.

Table 1. Research Variables

Variable	Variable Description	Unit	Variable Type
Y	Life Expectancy	Year	Continuous
X_1	Access to Decent Housing	Percent	Continuous
X_2	Access to Decent Drinking Water	Percent	Continuous
X_3	Expenditure per Capita	Thousand Rupiah/Person/Year	Continuous
X_4	Gini Ratio	Index	Continuous

2. Nonparametric Regression

Nonparametric regression is an approach utilized to explore the relationship between response variables and predictor variables without assuming a predefined form for the regression curve. This method is flexible, allowing data modelling without the assumption of a specific curve shape. Some approaches in nonparametric regression include local linear estimators, local polynomial estimators, kernel estimators, penalized spline estimators, and Fourier series estimators (Nurhuda et al., 2022). Thus, a nonparametric regression model for *n* observations can be expressed in the following form:

$$y_i = f(x_i) + \varepsilon_i \; ; i = 1,2, \dots, n, \qquad \varepsilon_i \sim IIDN(0, \sigma^2) \tag{1}$$

where y_i represent the *i* response variable, $f(x_i)$ represents the regression function value, x_i is the predictor variable, ε_i is the error term.

3. Multivariate Adaptive Regression Spline (MARS)

MARS is a nonparametric regression approach designed to estimate value of a response variable based on predictor variables without the need to make assumptions about the shape of the relationship between them. This approach combines spline techniques with recursive partitioning regression techniques to make a regression function estimate that remains continuous at the dividing points (Lembang et al., 2019). One of the main benefit of MARS is its capacity to identify interactions between predictor variables. This method is particularly effective for high-dimensional data, which involves having between 3 and 20 predictor variables. The range reflects a balance between significant data complexity and computational efficiency, allowing the model to provide more accurate predictions (Friedman, 1991). In MARS analysis, the best model is selected through a trial-and-error process by testing different parameter combinations, including Basis Function (BF), Maximum Interaction (MI), and Minimum Observation (MO). The primary criterion for identifying the optimal model is the smallest Generalized Cross Validation (GCV) value (Yasmirullah et al., 2021). The general form of the MARS model can be represented as follows:

$$\hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[S_{km} \left(x_{v(k,m)} - t_{km} \right) \right]_+ \tag{2}$$

with,
a_m represents the coefficient of the *m* basis function
M represents the total number of basis function
K_m denotes the the degree of interaction for the *m* basis function
$x_{v(k,m)}$ is the v predictor variable, k degree of interaction, and the m basis function
t_{km} is the point of knots
S_{km} is the sign at the knot point

The model of equation (2) can be simplified into the following form.

$$\hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m \, B_m(x) \tag{3}$$

with the basis function:

$$B_m(x) = \prod_{k=1}^{K_m} \left[S_{km} \left(x_{v(k,m)} - t_{km} \right) \right]_+ \tag{4}$$

Then, equation (3) can be expressed in the matrix form shown below.

$$\boldsymbol{y} = \boldsymbol{B\alpha} + \boldsymbol{\varepsilon} \tag{5}$$

with
$\boldsymbol{y}$ is an $(n \times 1)$ dimensional vector containing $(y_1, y_2, \dots, y_n)^T$
$\boldsymbol{\alpha}$ is a vector of dimension $[(M+1) \times 1]$ containing $(\alpha_0, \alpha_1, \dots, \alpha_M)^T$
$\boldsymbol{\varepsilon}$ is an (n×1) dimensional vector containing $(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^T$
$\boldsymbol{B}$ is an $[n \times (M+1)]$ matrix, with as many columns as the number of basis functions plus one for the intercepts.

4. Spline Regression

Spline regression is a nonparametric regression aproach designed to reduce variability and estimate data patterns that exhibit significant differences (Purnama, 2020). Its strength lies in its capacity to manage data patterns with sharp fluctuations, utilizing knot points to assist in this process (Mattalunru et al., 2022). These knot points act as transition markers that signal shifts in the data pattern. Additionally, the curves generated by spline regression are typically smoother (Maharani & Saputro, 2021). Truncated spline regression model can be written in the following formula.

$$y_i = f(x_i) + \varepsilon_i \; ; i = 1,2, \dots, n \tag{6}$$

The regression function $f(x_i)$ has an unknown form and resides in the space of continuous functions, making it possible to approximate it using a truncated spline function, as represented by the equation below.

$$f(x_i) = \beta_0 + \sum_{j=1}^{q} \beta_j x_i^j + \sum_{k=1}^{K} \beta_{q+k} (x_i - t_k)_+^q \tag{7}$$

with $f(x_i)$ is the spline regression function, with $i = 1,2, \dots, n$, β_0 is a constant, β_{q+k} is the coefficient for the $(q+k)$-th spline basis component, q is the spline order, with $q \geq 1$, t_k is the k-th knot point, K represents number of knot points, $(x - t_k)_+^q$ is the truncanted power function with.

$$(x_i - t_k)_+^q = \begin{cases} (x_i - t_k)^q ; x_i \geq t_k \\ 0 \quad\quad ; x_i < t_k \end{cases} \tag{8}$$

5. Coefficient of Determination ($\boldsymbol{R^2}$)

R^2 is a statistical measure used to evaluate how effectively the regression model can account for variations in the response variable (Kuncoro, 2019). The value of R^2 is intended to measure the extent to which the predictor variables explain variations in the response variable. According to Ghozali (2016), the R^2 also serves as an indicator of the model's goodness-of-fit, meaning it gauges the degree to which the predictor variables influence the response variable. This R^2 ranges from 0 to 1. The formula R^2 is explained as follows.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{9}$$

with y_i denotes the actual value, $\hat{y}_i$ represents the estimated value, and $\bar{y}_i$ is the actual mean value at observation i.

6. Significance Test of Regression Model Parameters

The model's fit is evaluated by testing the significance of the basis function coefficients, both simultaneous and individually (Ritonga & Sutarman, 2023). The purpose of the simultaneous test is to assess whether all the basis functions in the MARS model collectively impact the response variable (Risambessy et al., 2022). Hypothesis for the simultaneous test is stated as follows:

$H_0 = a_1 = a_2 = \cdots = a_m = 0$

$H_0 =$ There is a minimum of one $a_m \neq 0; m = 1,2, \dots, M$

The calculation of F test statistics is performed using the following formula:

$$F_{hitung} = \frac{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{M}}{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{N-M-1}} \tag{10}$$

The critical area for testing the basis function coefficients simultaneously is determined by comparing the F value with $_{(M;N-M-1)}$. Another alternative is to use the p-value by comparing it against the significance treshold (α). If the F value exceeds $_{(M;N-M-1)}$ or if the p-value is smaller than α, the null hypothesis (H_0) is rejected. Partial testing is used to measure the significance of each basis function individually on the response variable, as well as ensuring that the model can accurately describe the data based on the parameters applied. To perform this, the t test statistic can be applied (Risambessy et al., 2022). Hypothesis for partial test is as follows:

$H_0 = a_m = 0; m = 1,2, \dots, M$

$H_1 = a_m \neq 0; m = 1,2, \dots, M$

Calculation of t test statistics is performed using the following formula:

$$t = \frac{\hat{a}_m}{Se(\hat{a}_m)} \tag{11}$$

with

$$Se(\hat{a}_m) = \sqrt{var(\hat{a}_m)} \tag{12}$$

The critical area is estabilished by comparing the t value to $t_{\left(\frac{\alpha}{2};N-M-1\right)}$ or by comparing the p-value. If $t_{\left(\frac{\alpha}{2};N-M-1\right)}$ or the p-value is smaller than α, the null hypothesis H_0 is rejected.

C. RESULT AND DISCUSSION

1. Descriptive Statistics

In this study, descriptive statistics are presented through summary tables and scatterplots. The summary table is used to provide an overview of the variables used in this study, including their central tendencies, dispersion, and range of values. The data used have been pre-cleaned and are ready for use in modelling without requiring additional preprocessing. Furthermore, scatterplots help to visually inspect the relationship patterns between the response variable and each predictor, serving as an initial step to evaluate the suitability of a nonparametric approach.

Table 2. Statistics Descriptive

Variable	Mean	Varian	Minimum		Maximum	
			Value	Province	Value	Province
Y	70.69	5.784	66.01	West Sulawesi	75.12	Yogyakarta
X_1	62.49	159.82	29.01	Papua	85.79	Yogyakarta
X_2	88.19	60.11	66.49	Papua	99.42	Jakarta
X_3	11470	5131865	7562	Papua	19373	Jakarta
X_4	0.338	0.00215	0.244	Bangka Belitung	0.435	Yogyakarta

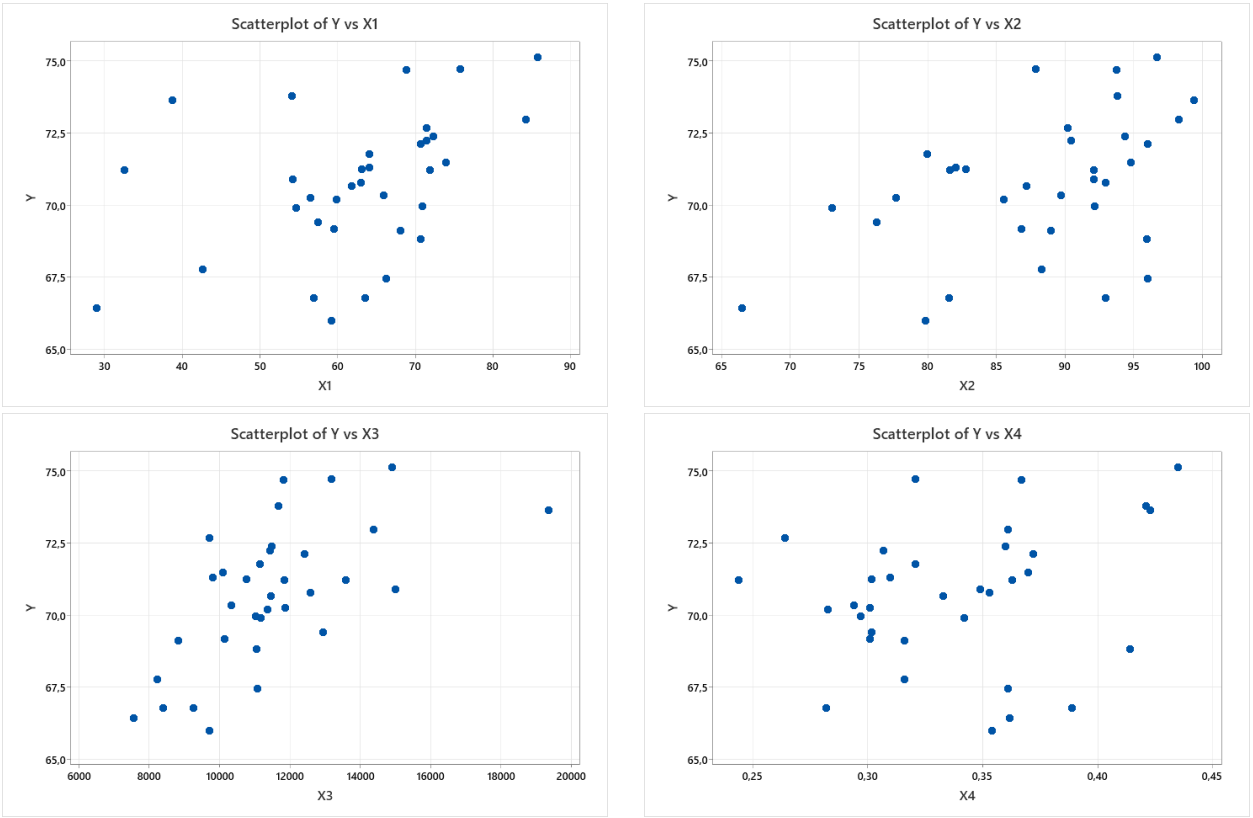

Figure 1. Scatterplot of Response Variable (Y) with Predictor Variables (X)

As illustrated in Figure 1, the distribution of the data between the response variable and all predictor variables does not exhibit a clear pattern. Therefore, these variables can be analyzed using a nonparametric regression approach.

2. Modelling with MARS

a. MARS Model Parameter Estimation

Calculations were performed using MARS software by integrating BF, MI, and MO. The BF value starts from 8 to 16, MI is 1, 2, and 3, and MO is 0, 1, 2, and 3. The optimal model is selected based on the lowest GCV value as follows Table 3.

Table 3. Best Model Combination

BF	MI	MO	GCV	R^2	MSE
12	2	0	3.723	0.576	2.61
		1	**2.867**	**0.827**	**1.183**
		2	4.461	0.484	3.177
		3	3.612	0.582	2.573

Based on Table 3, the optimal model is achieved from a combination of BF 12 with MI 2 and MO 1. The GCV result is 2.867; R^2 is 0.827; and MSE is 1.183. Then the estimation for the best model is presented below.

Table 4. Significant Model Basis Function Estimation

Basis Function (BF)	Parameter Estimation
Constant	70.535
$BF2 = \max(0, 11835 - X_3)$	-0.001
$BF3 = \max(0, X_1 - 59.280)$	0.697
$BF5 = \max(0, X_2 - 94.800) \times BF2$	-0.004
$BF7 = \max(0, X_2 - 66.490) \times BF3$	-0.020
$BF8 = \max(0, X_4 - 0.354)$	37.501

According to Table 4, the MARS model is obtained as follows:

$$\hat{Y} = 70.535 - 0.001BF2 + 0.697BF3 - 0.004BF5 - 0.020BF7 + 37.501BF8 \quad (13)$$

From equation (13) for the value of $X_3 < 11835$; $X_1 > 59.280$; $X_2 > 94.800$; and $X_4 > 0.354$ MARS model estimation is obtained as follows.

$$\hat{Y} = 70.535 - 0.001(11835 - X_3) + 0.697(X_1 - 59.280) - 0.004(X_2 - 94.800)(11835 - X_3) - 0.020(X_2 - 66.490)(X_1 - 59.280) + 37.501(X_4 - 0.354) \quad (14)$$

This interpretation shows how MARS captures nonlinearities and interactions flexibly through basis functions, particularly between access to decent housing and access to decent drinking water, as well as the expenditure per capita variable.

b. MARS Model Significance Test

1) Simultaneous Test

The outcomes of the simultaneous test are shown in the Table 5 below.

Table 5. Simultaneous Test of the MARS Model

Test Statistics	Value
F	26.679
$P-Value$	0.768931×10^{-9}

According to Table 5, the p-value is $0{,}768931 \times 10^{-9}$, smaller than the significance level α=0.05. As a result, the decision is to reject H_0, so it is concluded that there is a minimum of one α_m is not equal to zero for $m = 2, 3, 5, 7, 8$. This suggests that the model obtained is suitable and shows a relationship between the basis function coefficients and the response variable.

2) Partial Test

The outcomes of the partial test are shown in the Table 6 below.

Table 6. Partial Test of the MARS Model

Parameter	*T-Ratio*	$P-Value$	Decision
Constant	196.262	0.999201×10^{-15}	Reject H_0
$BF2$	-6.913	0.162919×10^{-6}	Reject H_0
$BF3$	4.161	0.272593×10^{-3}	Reject H_0
$BF5$	-4.514	0.104594×10^{-3}	Reject H_0
$BF7$	-3.486	0.002	Reject H_0
$BF8$	3.980	0.443808×10^{-3}	Reject H_0

According to Table 6, it is found that the p-value for each basis function in the model is also smaller than the significance level α = 0.05. Thus, the decision is to reject H_0, concluding that α_m is not equal to zero for m=2,3,5,7,8. This suggests that the obtained model accurately represents the connection between the basis function coefficients and the response variable.

c. Variable Importance Level

This is used to provide an overview of the priority of predictor variables that affect the response variablee. In modeling life expectancy in Indonesia, the importance level of predictor variables can be seen in Table 7 below.

Table 7. Variable Importance Level

Variable	Variable Description	Level of Importance	GCV Reduction
X_3	Expenditure per Capita	100%	6.148
X_2	Access to Decent Drinking Water	60.052%	4.050
X_1	Access to Decent Housing	53.004%	3.789
X_4	Gini Ratio	43.795%	3.496

According to Table 7, the predictor variable with the greatest impact on the response variable is expenditure per capita, which shows an importance level of 100% and a GCV

reduction of 6.148. Meanwhile, the gini ratio variable has the smallest influence, with an importance level of 43.795% and a GCV reduction of 3.496.

3. Modeling with Spline Nonparametric Regression

a. Selection of Optimal Knot Points

Spline nonparametric regression achieves the best model when it has optimal knot points, which are the points where the function's pattern changes. The selection of the optimal knot point is determined by minimizing the GCV value. In this study, one, two, and three knot points at the first order were used. Below is the GCV value for the access to adequate housing variable (X_1), as shown in Table 8.

Table 8. Optimum GCV Value of Variable X_1

Number of Knot	Knot Point	GCV
One Knot	66.2	4.827146
Two Knot	54.1; 55.1	4.321195
Three Knot	31; 54; 57	4.209747

According to Table 8, the minimum GCV value for variable X_1 is obtained at three knot points with knot points 31; 54; 57 and a GCV value of 4.209747. Next is the GCV value for the access to adequate drinking water variable (X_2) is as follows Table 9.

Table 9. Optimum GCV Value of Variable X_2

Number of Knot	Knot Point	GCV
One Knot	66.5	5.035866
Two Knot	66; 66.3	5.033506
Three Knot	93; 94; 95	4.991841

According to Table 9, the minimum GCV value for variable X_2 is obtained at three knot points with knot points 93; 94; 95 and a GCV value of 4.991841. Furthermore, the GCV value the expenditure per capita variable (X_3) is as follows Table 10.

Table 10. Optimum GCV Value of Variable X_3

Number of Knot	Knot Point	GCV
One Knot	11835	3.504649
Two Knot	9721; 9731	3.392449
Three Knot	11361; 11811; 11861	3.441303

According to Table 10, the minimum GCV value for variable X_3 is obtained at two knots with knot points 9721; 9731 and a GCV value of 3.392449. Furthermore, the GCV value for gini ratio variable X_4 is as follows.

Table 11. Optimum GCV Value of Variable X_4

Number of Knot	Knot Point	GCV
One Knot	0.413	5.406321
Two Knot	0.38; 0.39	5.289137
Three Knot	0.362; 0.365; 0.394	4.699213

According to Table 11, the minimum GCV value for variable X_4 is obtained at three knots with knot points 0.362; 0.365; 0.394 and a GCV value of 4, 699213. Based on the results obtained, it is found that the most optimal knot combination is (3,3,2,3), so the estimation of the spline regression model will use the knot combination (3,3,2,3) at first order.

b. Parameter Estimation for Spline Nonparametric Regression Model

The parameter estimation for spline regression model is presented as follows Table 12.

Table 12. Parameter Estimation Value for Spline Regression Model

Variable	Parameter	Estimated Value
	$\hat{\beta}_0$	0.200
X_1	$\hat{\beta}_1$	2.599
	$\hat{\beta}_2$	-2.617
	$\hat{\beta}_3$	-0.572
	$\hat{\beta}_4$	0.783
X_2	$\hat{\beta}_5$	-0.024
	$\hat{\beta}_6$	-0.079
	$\hat{\beta}_7$	-0.206
	$\hat{\beta}_8$	-0.311
X_3	$\hat{\beta}_9$	-0.001
	$\hat{\beta}_{10}$	0.325
	$\hat{\beta}_{11}$	-0.323
X_4	$\hat{\beta}_{12}$	0.150
	$\hat{\beta}_{13}$	0.016
	$\hat{\beta}_{14}$	0.014
	$\hat{\beta}_{15}$	0.003

Based on Table 12, with knot points on variable X_1 is 31; 54; 57; variable X_2 is 93; 94; 95; variable X_3 is 9721; 9731; and variable X_4 is 0.362; 0.365; 0.394, the following equation will be obtained.

$$\begin{aligned}\hat{y} = {} & 0.200 + 2.599x_1 - 2.617(x_1 - 31)_+^1 - 0.572(x_1 - 54)_+^1 + 0.783(x_1 - 57)_+^1 \\ & - 0.024x_2 - 0.079(x_2 - 93)_+^1 - 0.206(x_2 - 94)_+^1 - 0.311(x_2 - 95)_+^1 \\ & - 0.001x_3 + 0.325(x_3 - 9721)_+^1 - 0.323(x_3 - 9731)_+^1 + 0.150x_4 \\ & + 0.016(x_4 - 0.362)_+^1 + 0.014(x_4 - 0.365)_+^1 + 0.003(x_4 - 0.394)_+^1 (15)\end{aligned}$$

c. Parameter Testing of Spline Nonparametric Regression Model

1) Simultaneous Test

The outcomes of the simultaneous test are shown in the Table 13 below.

Table 13. Simultaneous Test of the Spline Regression Model

Source	DF	SS	MS	F	P-value
Regresi	15	149.781	9.985	3.511	0.006
Error	18	51.199	2.844		
Total	33	200.981			

According to Table 13, the P-value is 0.006 which means less than α=0.05. Thus, the decision is to reject H_0, suggesting that there is at least one parameter that is not equal to 0.

2) Partial Test

The outcomes of the partial test are shown in the Table 14 below.

Table 14. Partial Test of the Spline Regression Model

Variable	Parameter	Estimation	*t*	*P-value*	Description
	$\hat{\beta}_0$	0.200	6.328	0.000	Significant
X_1	$\hat{\beta}_1$	2.599	6.452	0.000	Significant
	$\hat{\beta}_2$	-2.617	-6.741	0.000	Significant
	$\hat{\beta}_3$	-0.572	-1.069	0.299	Not Significant
	$\hat{\beta}_4$	0.783	1.504	0.150	Not Significant
X_2	$\hat{\beta}_5$	-0.024	-0.388	0.703	Not Significant
	$\hat{\beta}_6$	-0.079	-0.104	0.919	Not Significant
	$\hat{\beta}_7$	-0.206	-0.609	0.550	Not Significant
	$\hat{\beta}_8$	-0.311	-0.337	0.740	Not Significant
X_3	$\hat{\beta}_9$	-0.001	-0.817	0.425	Not Significant
	$\hat{\beta}_{10}$	0.325	2.132	0.047	Significant
	$\hat{\beta}_{11}$	-0.323	-2.135	0.047	Significant
X_4	$\hat{\beta}_{12}$	0.150	1.456	0.163	Not Significant
	$\hat{\beta}_{13}$	0.016	0.664	0.515	Not Significant
	$\hat{\beta}_{14}$	0.014	0.685	0.502	Not Significant
	$\hat{\beta}_{15}$	0.003	0.313	0.758	Not Significant

According to Table 14, it is found that there are 5 parameters for which the decision is to reject H_0, indicating that the parameter is significant to the model. Meanwhile, the other 11 parameters with a p-value greater than α, lead to the decision to fail to reject H_0, meaning these parameters are not significant to the model. However, despite the presence of insignificant parameters, the variable is still included because at least one parameter is significant within each variable. Therefore, the variables of access to decent housing (X_1) and expenditure per capita (X_3) significantly affect life expectancy.

d. Coefficient of Determination (R^2)

$$R^2 = \frac{SS_{regression}}{SS_{total}}\ x\ 100\%$$
$$= \frac{149.781}{200.981}\ x\ 100\%$$
$$= 74.52\%$$

Based on these calculations, R^2 was obtained at 74.52%. This demonstrates that the model can explain 74.52% of the variation in the factors influencing life expetancy in Indonesia, while the remaining portion is attributed to other variables.

4. Comparison of the Best Method between MARS and Spline Nonparametric Regression

After selecting the best method used in MARS and spline nonparametric regression, the next step is to evaluate the two results, as shown in Table 15.

Table 15. Method Comparison

Metode	MSE	R^2
MARS	1,183	82,7%
Nonparametric Spline	2,844	74,52%

In Table 15, it can be seen that the R^2 of MARS is 82.7%. While the R^2 value obtained when using spline nonparametric regression is 74.52%. This means that the R^2 value of MARS is greater than spline nonparametric regression, making MARS the more appropriate model to use. The results of this research support previous studies, as they also found that several predictor variables exhibit interactions. The MARS method is able to accommodate such interactions through its basis function approach, making it more flexible compared to nonparametric spline regression. This aligns with prior findings and strengthens the position that MARS provides better performance in modelling data with complex variable relationships.

D. CONCLUSION AND SUGGESTIONS

According to the findings of this research, it can be concluded that the lowest life expectancy in Indonesia in 2023 is in West Sulawesi Province at 66.01 years, while the highest is in Yogyakarta Province at 75.12 years. The average life expectancy is 70.69 years with a variance of 5.784. Then, the optimal model selected based on the comparison of 2 existing methods is MARS with a combination of BF 12, MI 2, and MO 1. The results obtained R^2 of 82.7% and MSE of 1.183. The R^2 value of 82.7% indicates that 82.7% of the variation in the response variable can be explained by the predictor variable. The best model obtained is:

$$\hat{Y} = 70{,}535 - 0{,}001BF2 + 0{,}697BF3 - 0{,}004BF5 - 0{,}020BF7 + 37{,}501BF8$$

The basis functions (BF) in the MARS model help form segmented relationships in the data based on specific knot points, allowing the model to capture nonlinear patterns and interactions more effectively than conventional regression. Differences in life expectancy across provinces are influenced by factors such as healthcare access, socioeconomic conditions, education levels, and basic infrastructure availability. Therefore, the flexibility of the MARS model makes it highly relevant for informing health policy, particularly in identifying and prioritizing areas needing targeted interventions. Future research is encouraged to include additional predictor variables, such as per capita health expenditure, urbanization levels, sanitation access, or the Human Development Index, to improve model accuracy. Additionally, using data from multiple years and comparing various methods can enhance predictive performance and allow for analysis of temporal trends.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to all the lecturers of the Statistics Study Program Universitas Airlangga, parents, and friends who have helped in completing this research. Thanks also to the Central Bureau of Statistics for providing data for this research.

REFERENCES

Ali, P., & Younas, A. (2021). Understanding and interpreting regression analysis. *Evidence Based Nursing*, *24*(4), 116–118. https://doi.org/10.1136/ebnurs-2021-103425

Badan Pusat Statistik. (2023). *Angka Harapan Hidup Tahun 2021-2023*. BPS Provinsi Sulawesi Barat.

Badan Pusat Statistik. (2024). *Umur Harapan Hidup Saat Lahir (UHH)*. BPS RI. bps.go.id

Bangun, R. H. (2019). Analisis Determinan Angka Harapan Hidup Kabupaten Mandailing Natal(Life Expectations Determinants Analysis In Mandailing Natal Regency). *Jurnal Akuntansi Dan Ekonomi*, *4*(3), 22–31. https://doi.org/10.29407/jae.v4i3.13257

Bekar Adiguzel, M., & Cengiz, M. A. (2023). Model selection in multivariate adaptive regressions splines (MARS) using alternative information criteria. *Heliyon*, *9*(9). https://doi.org/10.1016/j.heliyon.2023.e19964

Erlyn, P., Hidayat, B., Cahyo, A., & Saksono, H. (2022). Investment in Human Resources to Increase Achievement Levels of Sustainable Development. *Jurnal Bina Praja*, *14*(1), 135–146. https://doi.org/10.21787/jbp.14.2022.135-146

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, *19*(1), 1–141.

Ghozali, I. (2016). *Aplikasi Analisis Multivariete Dengan Program IBM SPSS 23. Edisi 8*. Badan Penerbit Universitas Diponegoro.

Hasanah, U. (2017). Pengaruh Ketimpangan Pendapatan, Pendapatan per Kapita, dan Pengeluaran Pemerintah di Bidang Kesehatan Terhadap Sektor Kesehatan di Indonesia. *Jurnal Ilmu Ekonomi Terapan*, *2*(1), 30–43. https://doi.org/10.20473/jiet.v2i1.5504

Kuncoro, M. (2019). *Metode Riset Untuk Bisnis dan Ekonomi* (Edisi ke-3). Erlangga.

Lembang, F. K., Patty, H. W. M., & Maitimu, F. (2019). Analisis Kemiskinan di Kabupaten Maluku Tenggara Barat Menggunakan Pendekatan Multivariate Adaptive Regression Spline (MARS). *MEDIA STATISTIKA*, *12*(2), 188. https://doi.org/10.14710/medstat.12.2.188-199

Maharani, M., & Saputro, D. R. S. (2021). Generalized Cross Validation (GCV) in Smoothing Spline Nonparametric Regression Models. *Journal of Physics: Conference Series*, *1808*(1), 1–6. https://doi.org/10.1088/1742-6596/1808/1/012053

Mahmoud, H. F. F. (2021). Parametric Versus Semi and Nonparametric Regression Models. *International Journal of Statistics and Probability*, *10*(2), 90–108. https://doi.org/10.5539/ijsp.v10n2p90

Mattalunru, M. R., Annas, S., & Aidid, M. K. (2022). Aplikasi Multivariate Adaptive Regression Splines (MARS) untuk Mengetahui Faktor yang Mempengaruhi Curah Hujan di Kota Makassar. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *4*(1), 9–19. https://doi.org/https://doi.org/10.35580/variansiunm2

Mukrom, M. H., Yasin, H., & Hakim, A. R. (2021). Pemodelan Angka Harapan Hidup Provinsi Jawa Tengah menggunakan Robust Spatial Durbin Model. *Jurnal Gaussian*, *10*(1), 44–54. https://doi.org/https://doi.org/10.14710/j.gauss.10.1.44-54

Nurhuda, G. N., Wasono, W., & Nohe, D. A. (2022). Nonparametric Regression Modeling Based on Spline Truncated Estimator on Simulation Data. *Jurnal Matematika, Statistika Dan Komputasi*, *19*(1), 172–182. https://doi.org/10.20956/j.v19i1.21534

Pratama, Y. M., Fernandes, A. A. R., Wardhani, N. W. S., & Hamdan, R. (2024). Nonparametric Smoothing Spline Approach in Examining Investor Interest Factors. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, *8*(2), 425–440. https://doi.org/10.31764/jtam.v8i2.20192

Purnama, D. I. (2020). A Comparison between Nonparametric Approach: Smoothing Spline and B-Spline to Analyze The Total of Train Passangers in Sumatra Island. *EKSAKTA: Journal of Sciences and Data Analysis*, *1*(1), 73–80. https://doi.org/10.20885/EKSAKTA.vol1.iss1.art11

Risambessy, S., Aulele, S. N., & Lembang, F. K. (2022). Misclassification Analysis of Elementary School Accreditation Data in Ambon City Using Multivariate Adaptive Regression Spline. *Jurnal*

Matematika, Statistika Dan Komputasi, *18*(3), 394–406. https://doi.org/10.20956/j.v18i3.19451

Ritonga, N. A. R., & Sutarman. (2023). Estimation of Multivariate Adaptive Regression Splines (MARS) Model Parameters by Using Generalized Least Square (GLS) Method. *JMEA : Journal of Mathematics Education and Application*, *2*(2), 62–72. https://doi.org/10.30596/jmea.v2i2.13106

Wicaksono, W., Wilandari, Y., & Suparti. (2014). Pemodelan Multivariate Adaptive Regression Splines (MARS) pada Faktor-Faktor Resiko Angka Kesakitan Diare (Studi Kasus : Angka Kesakitan Diare di Jawa Tengah, Jawa Timur dan Daerah Istimewa Yogyakarta Tahun 2011). *Jurnal Gaussian*, *3*(2), 253–262. https://doi.org/https://doi.org/10.14710/j.gauss.3.2.253%20-%20262

Worldometer. (2024). *Life Expectancy of the World Population*. https://www.worldometers.info/demographics/life-expectancy/

Yasmirullah, S. D. P., Otok, B. W., Purnomo, J. D. T., & Prastyo, D. D. (2021). Parameter Estimation of Multivariate Adaptive Regression Spline (MARS) with Stepwise Approach to Multi Drug-Resistant Tuberculosis (MDR-TB) Modeling in Lamongan Regency. *Journal of Physics: Conference Series*, *1752*(1), 1–9. https://doi.org/10.1088/1742-6596/1752/1/012017