International Journal of Scientific Research and Engineering Development-– Volume 8 Issue 1, Jan-Feb 2025

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Modeling Life Expectancy in Indonesia Based on Multivariate Adaptive Regression Splines Approach

Bagas Shata Pratama[1], Suliyanto[2], M. Fariz Fadillah Mardianto[3], Sediono[4]

[1,2,3,4]Statististics Study Program, Department of Mathemathics, Airlangga University, Surabaya, Indonesia

[1]bagas.ata.pratama-2021@fst.unair.ac.id

[2]Corresponding author:suliyanto@fst.unair.ac.id

[3]m.farizfadillah.m@fst.unair.ac.id

[4]ahmad.s@fst.unair.ac.id

----------**********----------

Abstract:

Life expectancy is the average number of years a newborn is anticipated to live, assuming that the pattern of mortality by age at birth is the same throughout the baby's life. Life expectancy serves as one of the indicators that can be seen to determine the degree of public health in a country. This research focuses on identifying factors influencing life expectancy in Indonesia using the Multivariate Adaptive Regression Spline (MARS) method. This research utilizes secondary data from 2023, sourced from the Indonesian Central Bureau of Statistics, encompassing observations from 34 provinces. The optimal MARS model was established with BF 12, MI 2, and MO 1. The model achieved a Generalized Cross Validation (GCV) score of 2.86, an R-Square value of 0.82, and a Mean Square Error (MSE) of 1.18. The R-Square value indicates that 82% of the variability in life expectancy can be accounted for by the predictor variables.

Keywords — **Modeling, Life Expectancy, Multivariate Adaptive Regression Splines**

----------**********----------

I. INTRODUCTION

Health development plays a crucial role in enhancing community well-being, with one key measure being life expectancy. Life expectancy represents the average number of years a newborn is expected to live, assuming consistent mortality patterns throughout life [1]. A high life expectancy signifies successful health development efforts, whereas a low life expectancy highlights the necessity for improvements in environmental conditions, nutritional levels, and poverty reduction initiatives. [2]. The government has made various efforts, such as research, strengthening services, and developing health resources, which are in line with the 3rd goal of the SDGs on healthy and prosperous living [3]. According to [4], BPS data showing life expectancy in 2023 of 72.13 years, an increase of 1.73 years in the last decade. DIY province has the highest life expectancy (75.12 years) and West Sulawesi the lowest (66.01 years). Analysis of factors affecting life expectancy using the MARS method is needed to formulate effective policies to increase life expectancy in Indonesia.

Identifying factors that influence life expectancy is an important step to support the government in formulating policies aimed at improving people's quality of life. One approach that can be used is regression analysis. This approach seeks to explore the connection between response variables and predictor variables. Regression analysis can be categorized into two primary methods, parametric, which assumes a certain model shape, and nonparametric, which is more flexible because it does not require prior information about the model shape [5]. A nonparametric method relevant for life expectancy analysis is MARS, which is designed to handle high-dimensional data with predictor variables between 3 and 20. One of the strengths of MARS is its capability to detect interactions

between predictor variables as well as its flexibility in handling non-linear or specific relationships. This approach employs a blend of recursive splitting and truncated spline techniques during the regression process [6]. The best model can be obtained by adjusting the Maximum Basis Function (BF), Maximum Interaction (MI), and Minimum Observation (MO) parameters through experiments, by selecting model that has the lowest GCV value [7].

Studies regarding life expectancy have been carried out by [8] using panel data regression analysis. It was found that the decent housing access variable had a notable impact on life expectancy was analyzed using the Common Effect Model (CEM) regression approach. [9] examined life expectancy in Papua using the Geographically Weighted Regression (GWR) approach, it was found that the variables that had a significant effect were access to decent drinking water sources, projected schooling years, breastfeeding duration, and the midwife-to-population ratio per 10,000 individuals.

The novelty of this study compared to previous studies lies in its ability to analyze interactions between predictor variables. Some variables are known to be interrelated, such as access to safe drinking water and per capita expenditure. [10] It was found that per capita spending influences the accessibility of sanitation and clean water across different household categories. In addition, there is a relationship between access to decent housing and access to safe drinking water. Individuals who live in decent housing tend to care more about access to clean water, as these two aspects support each other in improving quality of life and health [11].

Based on the findings from previous research, some of the factors that are thought to influence life expectancy include access to adequate housing, access to adequate drinking water, gini ratio, and per capita expenditure. The MARS method is considered suitable for analyzing this data due to its flexibility in exploring the relationship between variables without requiring certain model assumptions. In addition, this method is also able to identify interactions between predictor variables visualized through basis functions [12]. Based on this background, this study focuses on modeling life expectancy in Indonesia using the MARS approach. It is expected The findings of this study may offer suggestions to the government in designing effective policies to increase life expectancy, so that it has a positive impact on the degree of public health in Indonesia.

II. RESEARCH METODOLOGY

This research utilizes secondary data for 2023 sourced from the publication of Indonesian Central Bureau of Statistics. The unit of observation analyzed includes 34 provinces in Indonesia. The variables examined in this research include life expectancy (Y), access to adequate housing (X_1), access to adequate drinking water (X_2), gini ratio (X_3), and per capita expenditure (X_4).

The data analysis method used in this study is nonparametric regression, using MARS. Nonparametric regression model for n observations can be expressed as follows:

$$y_i = f(x_i) + \varepsilon_i \; ; i = 1,2,\dots,n \qquad (1)$$

The selection of the best model in MARS analysis is done through a trial and error process by trying various combinations of parameters, namely BF, MI, and MO. The main criterion for determining the best model is the smallest GCV value. The general MARS model can be formulated as follows:

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

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

The steps in this research are as follows

1. Input data on life expectancy and the factors believed to influence it.
2. Create descriptive statistics.

3. Create a scatterplot for the response variable (Y) against each predictor variable (X).
4. Perform MARS modeling.
5. Selecting the best MARS model.
6. Testing the significance of parameters simultaneously and partially.
7. Testing residual assumptions.
8. Interpreting the model results and concluding the research findings.

III. RESULT AND DISCUSSION

A. Descriptive Statistics

This research presents descriptive statistical analysis using bar charts and scatterplots. The bar chart is used to provide an overview of life expectancy in Indonesia. On the other hand, scatterplots are used to observe the distribution pattern of the research data and serve as an initial step in identifying the potential for applying nonparametric approach methods. The following is a bar chart of life expectancy in Indonesia.

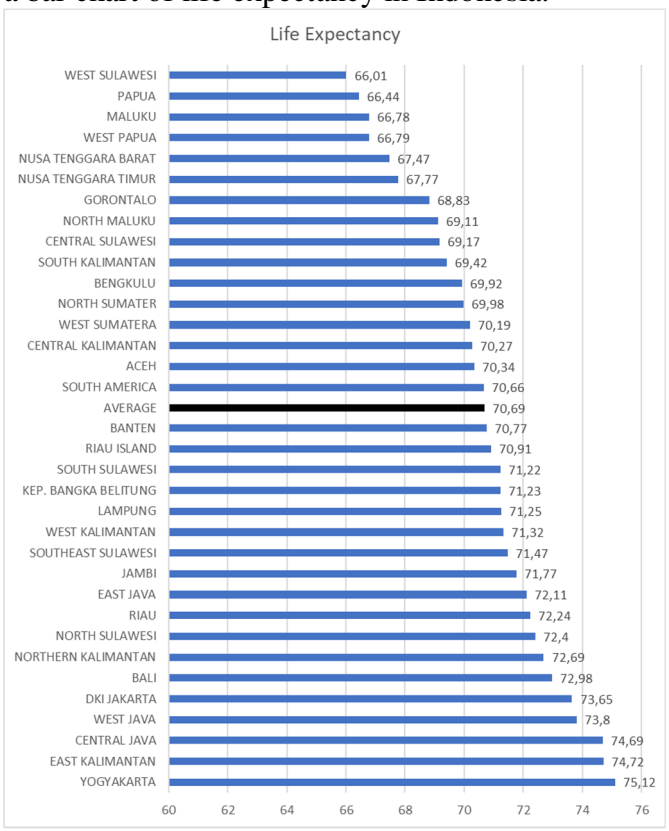

Fig. 1 Life Expectancy Diagram

As show in Fig. 1 West Sulawesi Province has the lowest life expectancy in Indonesia, at 66.01 years, while Yogyakarta Province has the highest, at 75.12 years. In addition, the average life expectancy is 70.69 years and variance of 5.78. This study finds no clear distribution pattern between the response variable and the predictor variables, making it appropriate to apply a nonparametric regression approach.

B. MARS Model Parameter Estimation

The calculation is done using MARS by combining BF, MI, and MO. The best model was derived from a combination of BF 12, MI 2, and MO 1. The GCV result is 2.86; R^2 is 0.82; and MSE is 1.18. The R^2 indicates that 82% of the variation in the response variable can be explained by the predictor variables.

The estimated value of the basis function can be calculated after determining best model. The estimation results for the best model in modeling life expectancy in Indonesia are as follows.

$$BF2 = \max(0, 11835 - X_4)$$
$$BF3 = \max(0, X_1 - 59.28)$$
$$BF5 = \max(0, X_2 - 94.8) \times BF2$$
$$BF7 = \max(0, X_2 - 66.49) \times BF3$$
$$BF8 = \max(0, X_3 - 0.354)$$

MARS model is as follows:

$$\hat{Y} = 70.54 - 0.001BF2 + 0.697BF3 - 0.004BF5 - 0.02BF7 + 37.501BF8 \quad (3)$$

From equation (3), the complete MARS model estimation is obtained as follows:

$$\hat{Y} = 70.54 - 0.001(11835 - X_4) + 0.697(X_1 - 59.28) - 0.004(X_2 - 94.8)(11835 - X_4) - 0.02(X_2 - 66.49)(X_1 - 59.28) + 37.501(X_3 - 0.354) \quad (4)$$

C. Significance Test of MARS Model

Significance test of MARS model is conducted using two approaches: simultaneous and partial. Simultaneous testing aims to evaluate if all the basis function coefficients in the MARS model have a combined effect on the response variable. The results of the simultaneous test for the model's basis function coefficients are as follows.

TABLE I
SIMULTANEOUS TEST OF MARS

Test Statistics	Value
$P-Value$	0.768×10^{-9}

According to Table 1, the p-value is 0.768×10^{-9}, smaller than significance level of $\alpha = 0.05$. So the decision rejects H_0 and the conclusion is that there is at least one $\alpha_m \neq 0$.

Then, partial testing is carried out to determine whether any of the basis function coefficient in the MARS model significantly impacts the response variable. The results of the partial test for the model's basis function coefficients are as follows.

TABLE 2
PARTIAL TEST OF MARS

Parameters	$P-Value$	Decision
Constant	0.999×10^{-15}	Reject H_0
$BF2$	0.162×10^{-6}	Reject H_0
$BF3$	0.272×10^{-3}	Reject H_0
$BF5$	0.104×10^{-3}	Reject H_0
$BF7$	0.002	Reject H_0
$BF8$	0.443×10^{-3}	Reject H_0

According to Table 2, the p-value for each basis function is smaller than the significance level $\alpha = 0.05$. So the decision taken is reject H_0, so it is concluded that α_m is not equal to zero, with the value of $m = 2, 3, 5, 7, 8$.

D. Interpretation of the Best MARS Model

Once the best model is identified, the significance of the variables in the model is tested, and the assumptions regarding the residuals are checked. Additionally, the response variable and its predicted values can be visualized through a plot to compare the two. The resulting plot is presented below.

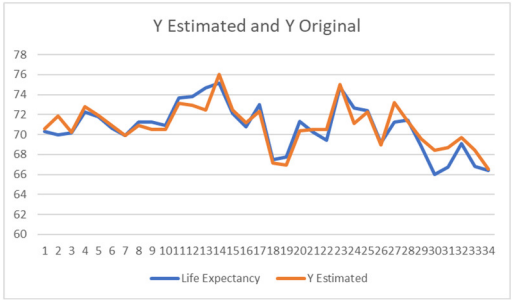

Fig. 2 Plot of Y with $\hat{Y}$

As shown in Fig. 2, the estimated value $(\hat{Y})$ is closely aligned with the actual value (Y). The results of the MARS model interpretation that has been obtained in equation (4) can be described as follows:

1. Basis Two Function $(BF2)$

$$BF2 = \begin{cases} (11835 - X_4) & ; for\ X_4 < 11835 \\ 0 & ; for\ the\ other\ X_4 \end{cases}$$

The coefficient of -0.001 on the second basis function $(BF2)$ indicates that for every one unit increase in $BF2$ will reduce life expectancy by 0.001 years, assuming $BF3$, $BF5$, $BF7$, and $BF8$ remain constant. That is, if per capita expenditure increases by 1 unit, while other predictor variables do not change, then life expectancy in provinces with per capita expenditure of less than 11835 will decrease by 0.001 years. This indicates a negative relationship between per capita expenditure and life expectancy.

2. Basis Three Function $(BF3)$

$$BF3 = \begin{cases} (X_1 - 59.28) & ; for\ X_1 > 59.28 \\ 0 & ; for\ the\ other\ X_1 \end{cases}$$

The coefficient of 0.697 on the basis function three $(BF3)$ indicates that for every-one unit increase in $BF3$ will increase life expectancy by 0.697 years, assuming that $BF2$, $BF5$, $BF7$, and $BF8$ remain constant. This means that if access to decent housing increases by 1 unit, while other predictor variables do not change, then life expectancy in provinces with access to decent housing of more than 59.28 will increase by 0.697 years.

3. Basis Five Function $(BF5)$

$$BF5 = \begin{cases} (X_2 - 94.8)(11835 - X_4); \\ for\ X_4 > 94.8\ and\ X_4 < 11835 \\ 0 \quad ; for\ the\ other\ X_2\ and\ X_4 \end{cases}$$

The coefficient value of -0.004 on the base five function $(BF5)$ indicates that for every one unit increase in $BF5$ will reduce life expectancy by 0.004 years, assuming that $BF2, BF3, BF7$, and $BF8$ remain constant. In addition, another meaning is that the province if it has a value on the variable of access to decent drinking water of more than 94.8 and the value of per capita expenditure is less

than 11835 will make a significant contribution, namely reducing life expectancy by 0.004 years.

4. Basis Seven Function ($BF7$)

$$BF7 = \begin{cases} (X_2 - 66.49)(X_1 - 59.28); \\ for\ X_2 > 66.49\ and\ X_1 > 59.28 \\ 0 \quad ; for\ the\ other\ X_2\ and\ X_1 \end{cases}$$

The coefficient value of -0.02 on the base function seven ($BF7$) indicates that for every one unit increase in $BF7$ will reduce life expectancy by 0.02 years, assuming that $BF2, BF3, BF5$, and $BF8$ remain constant. In addition, another meaning is that the province if it has a value on the decent drinking water access variable of more than 66.49 and a decent housing access value of more than 59.28 will make a significant contribution, namely reducing life expectancy by 0.02 years.

5. Basis Eight Function ($BF8$)

$$BF8 = \begin{cases} (X_3 - 0.354) \ ; for\ X_3 > 0.354 \\ 0 \qquad\quad ; for\ the\ other\ X_3 \end{cases}$$

The coefficient value of 37.501 on the basis function eight ($BF8$)) indicates that for every-one unit increase in $BF8$ will increase life expectancy by 37.501 years, assuming that $BF2, BF3, BF5$, and $BF7$ remain constant. This means that if the gini ratio increases by 1 unit, while other predictor variables do not change, then life expectancy in provinces with a gini ratio of more than 0.354 will increase by 37.501 years. This shows that the gini ratio has a positive effect on life expectancy.

IV. CONCLUSIONS

Based on the findings of this research, it can be inferred that the lowest life expectancy in Indonesia is 66.01 years (West Sulawesi Province), while the highest is 75.12 years (Yogyakarta Province). Then, the best model was derived from a combination of BF 12, MI 2, and MO 1. There are 5 significant Basis Functions, namely $BF2, BF3, BF5$, $BF7$ and $BF8$. The GCV result is 2.86; R^2 is 0.82; and MSE is 1.18. The R^2 value of 0.82 indicates that 82% of the variation in the response variable can be explained by the predictor variables.

ACKNOWLEDGMENT

The authors wish to express their gratitude to the Statistics Study Program of Universitas Airlangga, the Central Bureau of Statistics for providing the data, and everyone who contributed to this research.

REFERENCES

[1] Badan Pusat Statistik, "Umur Harapan Hidup Saat Lahir (UHH)," BPS RI. Accessed: Mar. 04, 2024. [Online]. Available: bps.go.id
[2] R. H. Bangun, "Analisis Determinan Angka Harapan Hidup Kabupaten Mandailing Natal(Life Expectations Determinants Analysis In Mandailing Natal Regency)," *Jurnal Akuntansi dan Ekonomi*, vol. 4, no. 3, pp. 22–31, 2019, doi: 10.29407/jae.v4i3.13257.
[3] Kementrian Kesehatan RI, *Rencana Aksi Program 2020-2024*. Jakarta: Badan Penelitian dan Pengembangan Kesehatan Kementrian Kesehatan RI, 2020.
[4] Worldometer, "Life Expectancy of the World Population." Accessed: Mar. 04, 2024. [Online]. Available: https://www.worldometers.info/demographics/life-expectancy/
[5] R. L. Eubank, *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker, 1999.
[6] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–141, 1991.
[7] R. Muslikah and M. Y. Darsyah, "Multivariate Adaptive Regression Splines (MARS) untuk Klasifikasi Kejadian Konstipasi Terhadap Pemberian Air Susu Ibu dan Pemberian Air Susu Formula," *Jurnal Statistika*, vol. 3, no. 2, pp. 15–21, 2015, doi: 10.26714/jsunimus.3.2.2015.%25p.
[8] A. Septianingsih, "Pemodelan Data Panel Menggunakan Random Effect Model untuk Mengetahui Faktor yang Mempengaruhi Umur Harapan Hidup di Indonesia," *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 3, no. 3, pp. 525–536, 2022.
[9] A. Tanadjaja, I. Zain, and W. Wibowo, "Pemodelan Angka Harapan Hidup di Papua dengan Pendekatan Geographically Weighted Regression," *Jurnal Sains dan Seni ITS*, vol. 6, no. 1, pp. 82–88, 2017.
[10] W. Watekhi, D. Hartono, and R. K. Dewi, "Analisis Kesediaan Membayar Air Bersih dan Sanitasi Rumah Tangga di Indonesia," *Jurnal Ekonomi dan Pembangunan Indonesia*, vol. 12, no. 1, pp. 1–14, Jul. 2012, doi: 10.21002/jepi.v12i1.01.
[11] S. Purwoko, "Indikator Air Layak Minum dan Sanitasi Layak dalam Mendukung Upaya Kesehatan Lingkungan di Rumah Tinggal," Surabaya: Prosiding Seminar Nasional GERMAS 2018, 2018, pp. 62–67.
[12] W. Wicaksono, Y. Wilandari, and Suparti, "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*, vol. 3, no. 2, pp. 253–262, 2014.

