Modeling of toddler stunting in the province of east nusa tenggara using multivariate adaptive regression splines (mars) method

D M Azizah1, E O Permatasari2
1Statistics Department, Faculty of Mathematics, Computing and Data Science, InstitutTeknologiSepuluhNopember, GebangKidul Road number 66, Surabaya 60111 Indonesia
2Statistics Department, Faculty of Mathematics, Computing and Data Science, InstitutTeknologiSepuluhNopember, GebangKidul Road number 66, Surabaya 60111 Indonesia

email: dewimuslima96@gmail.com, erma.oktania@gmail.com

Abstract. Stunting is a condition of failure to thrive in children under five (babies under five years old) as a result of chronic malnutrition so that the child is too short for his age. Malnutrition occurs since the baby is in the womb and in the early period after the baby is born. However, the condition of stunting only appears after the baby is 2 years old. Based on Basic Health Research in 2017 the province with the highest prevalence of short and very short toddlers in Indonesia is East Nusa Tenggara which is equal to 40.3 percent. This shows that the prevalence of stunting toddlers in NTT Province is still far from the overall prevalence of stunting toddlers in Indonesia. Cases of toddler stunting can be analyzed using the Multivariate Adaptive Regression Splines (MARS) method because the data used are high-dimensional data and do not show clear patterns of relationships between response variables and predictor variables. The data used is the percentage of stunting toddlers in NTT in 2017 with the district / city research unit. Based on the analysis, it was found that the district in NTT with the highest percentage of stunting children was South East Timor and the district with the lowest percentage was East Manggarai. The best MARS model is formed from 3 basis functions with 2 predictor variables included in the model, namely the percentage of pregnant women at risk of chronic lack of energy (X1) and the percentage of infants with low birth weight (X6). And after the classification results obtained the accuracy of the classification of 77.27% with an error of 22.72%.

1. Introduction
Toddler is a period that is very sensitive to the environment, so more attention is needed, especially the adequacy of nutrition. Chronic and ongoing malnutrition can cause stunting. Stunting is a chronic nutritional problem in toddlers characterized by shorter height compared to children their age. Chronic nutritional problems due to malnutrition occur since the baby is in the womb and in the early period after the baby is born, but the condition of stunting only appears after the baby is 2 years old [1].
Children who suffer from stunting will be more susceptible to disease and as adults at risk of developing degenerative diseases. In addition to influencing health, stunting can also affect children's intelligence [2].

Based on Basic Health Research in 2017 the proportion of short and very short toddlers in Indonesia reached 29.6%. Which means that almost one third of children under five in Indonesia are stunted. The province with the highest prevalence of short and very short toddlers in 2017 in Indonesia is East Nusa Tenggara Province which is equal to 40.3% [2]. Where NTT consists of 21 districts and 1 city. On the other hand the target of the prevalence of short toddlers or stunting in Indonesia in 2019 is 28%. This shows that the prevalence of stunting under five in NTT Province is still far from Indonesia's target.

Previous research on stunting toddlers was conducted by Aridiyah, Rohmawati, Ririanty (2015) by analyzing the factors that influence stunting in children under five in rural and urban areas using the chi-square test, mannwhitney test and logistic regression [3]. The study states that the factors that influence the incidence of stunting in children under five in urban areas are mother's education, family income, maternal knowledge about nutrition, exclusive breastfeeding, age of MP-ASI, zinc adequacy level, iron adequacy level, history infectious diseases as well as genetic factors from parents. Furthermore, research conducted by Anindita (2018) concerning modeling of stunting toddlers with the nonparametric spline truncated method [4]. The results showed that the variable percentage of children under five received complete immunization, the percentage of children under five received breastfeeding for 6 months, the percentage of pregnant women at risk of SEZ, the percentage of babies born with IMD and the percentage of households that had proper sanitation had a significant effect on the model. The coefficient of determination generated from this model is 80.77%.

Another study by Jihad et al (2016) about the determinant analysis of the incidence of stunting in infants aged 12-24 months in the working area of Puuwatu Health Center in Kendari City in 2016 [5]. The results of this study indicate LBW, history of exclusive breastfeeding and age of exclusive breastfeeding and maternal height are risk factors for stunting. While a history of anemia in the mother during pregnancy is not a risk factor for stunting.

Multivariate Adaptive Regression Splines (MARS) is a nonparametric regression problem solving whose main purpose is to predict a value of the response variable from several predictor variables without assuming the basic function relationship between the response variable and the predictor variable. MARS can overcome cases with high dimensions or cases with many predictor variables. This method introduced by Friedman (1991) has a flexible form of function [6]. Previous studies using the MARS method were conducted by several researchers including Permatasari (2013) about the Multivariate Adaptive Regression Splines (MARS) approach for the classification of poverty depth indexes in East Java Province [7]. The accuracy of the classification of the MARS method obtained was 89.47% with an error rate of 10.53%. While the accuracy of classification with binary MARS is 97.37% with an error rate of 2.63%. Nasuha (2016) has also conducted research on the level of open unemployment in districts / cities in Central Java Province by using the Multivariate Adaptive Regression Splines (MARS) approach [8]. The minimum GCV value obtained was 0.396 with an R-square of 86.5%.

Based on this background it is known that NTT is a province in Indonesia with the highest prevalence of stunting under five. So that in overcoming stunting toddlers, it is necessary to have research on what factors influence stunting toddlers by using the multivariate adaptive regression splines method. The multivariate adaptive regression splines method is used because the data in this study are high-dimensional data and do not show a clear relationship or pattern between the response variable and the predictor variable.
2. Method

2.1. Data Source
The data used in this study are secondary data. Data was obtained from the official report of the Ministry of Health on the results of Nutrition Status Monitoring (PSG) in 2017 on the percentage of stunted children in East Nusa Tenggara Province and East Nusa Tenggara Health Profile 2017 on Family Health and Environmental Health.

2.2. Multivariate Adaptive Regression Splines
The method used in this study is the adaptive spline regression multivariate method. MARS is an approach for multivariate nonparametric regression developed by Friedman (1991) [6]. This MARS model is useful to overcome the problem of high-dimensional data that is data that has a number of predictor variables of more than 3 (three) [12] and is able to produce accurate predictor response variables.

In general, the MARS model according to Friedman (1991) can be written in the following equation (1):

\[ y_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m B_m(x) + \varepsilon_i \]  

Where \( \alpha_0 \) is the constant coefficient of the base function \( B_0 \), and \( \alpha_m \) is the coefficient of the m-base function, and \( B_m(x) = \prod_{k=1}^{K_m}[S_{km} \cdot (x_v(k,m) - t_{km})]_+ \). So, if they are written in matrix form the following:

\[ \mathbf{y} = \mathbf{B} \mathbf{\alpha} + \mathbf{\varepsilon} \]  

where \( \mathbf{y} = (y_1, y_2, ... , y_n)^T \), \( \mathbf{\alpha} = (\alpha_1, \alpha_2, ... , \alpha_M)^T \), \( \mathbf{\varepsilon} = (\varepsilon_1, \varepsilon_2, ... , \varepsilon_n)^T \),

\[
\mathbf{B} = \begin{bmatrix}
1 \prod_{k=1}^{K_1} S_{1m}(x_{1(1,m)} - t_{1m}) & \cdots & \prod_{k=1}^{K_1} S_{Mm}(x_{1(M,m)} - t_{Mm}) \\
1 \prod_{k=1}^{K_2} S_{1m}(x_{2(1,m)} - t_{1m}) & \cdots & \prod_{k=1}^{K_2} S_{Mm}(x_{2(M,m)} - t_{Mm}) \\
\vdots & \ddots & \vdots \\
1 \prod_{k=1}^{K_n} S_{1m}(x_{n(1,m)} - t_{1m}) & \cdots & \prod_{k=1}^{K_n} S_{Mm}(x_{n(M,m)} - t_{Mm})
\end{bmatrix}
\]

Friedman’s modification to estimate the MARS model is written in the following equation:

\[ f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m}[S_{km} \cdot (x_v(k,m) - t_{km})]_+ \]  

With function,

\[
(x_v(k,m) - t_{km}) = \begin{cases} 
(x_v(k,m) - t_{km}), & x_v(k,m) - t_{km} > 0 \\
0, & x_v(k,m) - t_{km} \leq 0
\end{cases}
\]

2.3. Analysis Steps
The steps of the analysis are:
1. Collecting data on the percentage of stunted children under five in East Nusa Tenggara Province in 2017 and factors that are thought to be influential.
2. Perform descriptive statistical analysis of the response variables and predictor variables.
3. Identify the shape of the data pattern between the response variables namely the percentage of stunted children in East Nusa Tenggara Province with each predictor variable.
4. Develop a percentage model of stunting toddlers in the Province of East Nusa Tenggara using the MARS method with the following stages [6].
5. Determine the combination of the base function (BF), maximum interaction (MI) and minimum observation (MO) in the following way [6].
   - Determine the maximum BF which is 2 to 4 times the number of predictor variables used.
   - Determine the number of MI, namely 1, 2, and 3 with the assumption that, if MI > 3 will produce an increasingly complex model.
   - Determine the number of MO between knots namely 0, 1, 2, 3, 5 and 10.

6. Get models based on minimum GCV values.
7. Estimating MARS model parameters.
8. Get the best MARS models.

3. Result and Discussion

3.1. Characteristics of Toddler Stunting Data and affecting factors

The description to find out the information that can be obtained from the percentage of stunting toddlers in NTT and the factors that are suspected to influence it can be identified through descriptive statistical analysis.

Table 2 shows that the variables $X_2$, $X_3$, and $X_5$ have high variance values compared to other variables, ie above 200. This shows that the diversity of data on the percentage of pregnant women getting iron ($X_2$), the percentage of children under five receiving exclusive breastfeeding for 6 months ($X_3$), and the percentage of children under five getting complete immunization ($X_5$) between districts / cities in NTT tends to be quite high when compared to other variables.

Based on Nutrition Status Monitoring data in 2017 published by the Ministry of Health also shows that the percentage of pregnant women getting iron has an average of 60.97% with a maximum value of 90.18% located in Alor Regency and the minimum value is only 4.59 % is located in Southwest Sumba Regency. Because of the lack of mothers getting iron, the government needs to increase attention on fulfilling the nutrition of pregnant women, especially on iron needs.

For the variable percentage of children under five getting exclusive breastfeeding for 6 months, an average of 45.91% with a minimum value of 21.1% located in Timor Tengah Selatan District and a maximum value of 71.8% located in Belu District. As for the variable percentage of children under five getting complete immunization, on an average of 72.29% with a minimum value of 36.75% located in Rote Ndao, and a maximum value of 116.82% located in Malacca.

The response variable is the percentage of stunting toddlers having an average value of 39.92% with a diversity between districts / cities of 51.41. According to WHO, areas with a stunting percentage of toddlers 30% -39% are included in areas with a high percentage of stunting toddlers. This means that NTT is included in a region with a high percentage of stunting under five according to WHO. The regency / city in NTT with the highest percentage of stunting children under five is East Timor Regency, which is 53.50%. While the district with the lowest percentage of stunting children under five is East Manggarai Regency, which is 24.4% as shown in Figure 1.

| Variable | Mean | Variance | Minimum | Maximum |
|----------|------|----------|---------|---------|
| Y        | 39.92| 51.41    | 24.4    | 53.50   |
| X1       | 18.91| 89.18    | 2       | 41.70   |
| X2       | 60.97| 413.30   | 4.59    | 90.18   |
| X3       | 45.91| 273.96   | 21.1    | 71.80   |
| X4       | 11.74| 89.98    | 0.60    | 37.30   |
| X5       | 72.29| 350.21   | 36.75   | 116.82  |
| X6       | 5.61 | 6.19     | 1.4     | 10.8    |
In 2017 districts / cities in NTT based on the percentage of pregnant women with the risk of KEK \((X_1)\) has the distribution as shown in Figure 2.

Based on Figure 2, it is known that districts / cities in NTT which have a percentage of pregnant women with Chronic Energy Deficiency risk above the third quartile are only 2 districts namely Alor and South Central Timor. While districts / cities with a percentage of pregnant women with Chronic Energy Deficiency risk less than the first quartile there are 5 districts namely West Manggarai, Manggarai, East Manggarai, Balu and Central Sumba. The distribution of districts / municipalities in NTT based on the percentage of babies receiving exclusive breastfeeding \((X_3)\) is presented in Figure 3.
Figure 3 shows that almost one third or 6 districts / cities in NTT have a percentage of babies who are exclusively breastfed for 6 months above the third quartile, namely in East Manggarai, Nageko, Ende, Lembata, Belu and Central Sumba. Whereas districts / cities with the percentage of infants receiving exclusive breastfeeding for 6 months less than the first quartile were 5 districts and South Central Timor included. If reviewed based on the percentage of babies with low birth weight, the districts / cities in NTT in 2017 have the following distribution.

![Map of NTT Based on Percentage of Low Birth Weight Babies](image1)

**Figure 4** Map of NTT Based on Percentage of Low Birth Weight Babies

Based on Figure 4 it is known that districts / cities in NTT that have a percentage of babies with low birth weight above the third quartile are 3 districts namely Nageko, Ende, Sikka. While the districts / cities with the percentage of babies with low birth weight less than the first quartile there are 4 districts namely Lembata, Southwest Sumba, East Manggarai, Alor, and Malacca. After knowing the characteristics of these predictor variables, the following are the characteristics of the response variable, namely the percentage of stunting toddlers in NTT. The following is a visualization of the response variable which is the percentage of stunted children in NTT. Figure 5 can be seen that districts / cities in NTT which have a percentage of stunting under five above the third quartile there are 5 districts namely in Manggarai, Sabu Rajua, Rote Ndao, Kupang, and South Central Timor. So, it needs better attention from the government in terms of preventing the occurrence of stunting toddlers.

![Map of NTT Based on Percentage of Stunting Toddler](image2)

**Figure 5** Map of NTT Based on Percentage of Stunting Toddler

### 3.2 Formation of the MARS Model

The formation of the MARS model is done by determining the combination of the number of Base Functions (BF), Maximum Interaction (MI), and Minimum Observation (MO) between knots. Base functions are functions that are defined in each region. The maximum allowable base function is two to four times the predictor variable used. Maximum Interaction (MI) is the maximum number of correlation relationships between variables in the model. If the MI used is 1, it means there is no interaction between variables in the model. If MI is used 2, then there are interactions between the 2 variables in the model. Similarly, if the MI used is 3, then the interaction that can occur at most
between 3 variables. Whereas the minimum observation (MO) is the minimum number of observations between knots.

This study uses 6 (six) predictor variables as variables that are thought to influence the response variable (percentage of stunting toddlers), so that the maximum number of BF used is 12, 18 and 24. The maximum interaction (MI) used in this study is 1, 2 , and 3. While the minimum observations (MO) are 0, 1, 2, and 3. So that the number of possible models based on a combination of BF, MI, and MO is 36 models. The selection of the best models from the 36 models formed is seen based on the minimum GCV value.

Based on 36 models from a combination of BF, MI, and MO values, the best model with the smallest GCV value was obtained. If the GCV values obtained are the same, then the largest $R^2$ value can be considered. If the value of $R^2$ is still the same value, it can adhere to the principle of model parsimony by choosing the model that has the smallest BF, MI, and MO combination value.

The model is the second model. The second model is formed from the number of BF = 12, MI = 1, and MO = 1 with a GCV value of 39.885. The MARS model obtained is as follows.

$$\hat{f}(x) = 35.74 + 0.539 * BF_1 - 2.129 * BF_2 - 3.223 * BF_3$$

where,

$$BF_1 = \max(0, X_1 - 2) = \begin{cases} 
0, & X_1 \leq 2 \\
(X_1 - 2), & X_1 > 2
\end{cases}$$

$$BF_2 = \max(0, X_6 - 5.2) = \begin{cases} 
0, & X_6 \leq 5.2 \\
(X_6 - 5.2), & X_6 > 5.2
\end{cases}$$

$$BF_3 = \max(0, 5.2 - X_6) = \begin{cases} 
0, & X_6 \geq 5.2 \\
(5.2 - X_6), & X_6 < 5.2
\end{cases}$$

The predictor variables included in the MARS model in equation (4.1) are 2 predictor variables, namely the percentage of pregnant women at risk of chronic energy deficiency ($X_1$) and the percentage of infants with lower body weight (LBW) ($X_6$). Then the coefficient test is performed to determine that the selected MARS model is an appropriate model and shows the relationship between the predictor variable and the response variable. The following hypotheses are used for simultaneous testing.

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

$$H_1 : \text{at least one } \alpha_j \neq 0, j=1, 2, 3$$

with the rejection area: Reject $H_0$ if the value is $F > F_{table}$

| Table 2 Test Concurrent Testing Statistics |
|-------------------------------------------|
| $F_{value}$ | $F_{0.05;3;18}$ | $P_{value}$ | Decision       |
|-------------|-----------------|-------------|----------------|
| 10,622      | 3.95            | 0.301E-03   | Tolak $H_0$    |

Table 2 shows that the calculated F value is greater than $F_{table}$ and the P-value is smaller than $\alpha$ (0.05) so that the decision to reject $H_0$ is obtained, which means there is at least one significant $\alpha_j$. The model chosen is an appropriate model and shows the right relationship between the predictor variable and the response variable. Then partial test (individual) is carried out with the following hypothesis.

$$H_0 : \alpha_j = 0$$

$$H_1 : \alpha_j \neq 0, j=1, 2, 3$$
Rejected area: Reject $H_0$ if the value is $T > T_{table}$

| Parameter | Estimate | Standard Error | $T_{value}$ | $P$-value |
|-----------|----------|----------------|-------------|-----------|
| $\alpha_1$ | 0.529 | 0.111 | 4.876 | 0.0002 |
| $\alpha_2$ | -2.129 | 0.705 | -3.021 | 0.007 |
| $\alpha_3$ | -3.223 | 0.901 | -3.576 | 0.002 |

Table values with $\alpha = 0.05$ and df = 18 obtained by 1.734. So, based on Table 5 of all parameters obtained a decision reject $H_0$ which means that each predictor variable on the basis function in the model has a significant effect on the response variable.

### Table 4 Importance of Predictor Variables

| Variable | Level of Importance |
|----------|---------------------|
| $X_1$    | 100%                |
| $X_6$    | 37.775%             |
| $X_2$    | 0                   |
| $X_3$    | 0                   |
| $X_4$    | 0                   |
| $X_5$    | 0                   |

The amount of contribution or level of importance for each variable to the model can be seen in Table 4. Based on Table 4 the percentage of pregnant women at risk of chronic energy deficiency ($X_1$) has a contribution to the MARS model of 100%, and the percentage of infants with weight lower body ($X_6$) had a contribution of 37.775 to the MARS model that was formed. While the other four variables have no importance (0%) because they are already represented by two variables that enter the model.

### 4. Conclusion

Based on the analysis that has been done, obtained conclusion as follows:

1. Characteristics of data on the percentage of stunting children under five in NTT shows that South East Timor is the district with the highest percentage of stunting children under five, which is 53.50%. South East Timor has the highest percentage of pregnant women at risk of chronic energy deficiency in NTT and the lowest percentage of children under five who receive exclusive breastfeeding. While the district with the lowest percentage of stunting under five is East Manggarai Regency, which is 24.4%. Where Manggarai is the district with the lowest percentage of pregnant women with the lowest chronic energy deficiency risk and the percentage of children under five who receive exclusive breastfeeding.

2. The model that illustrates the pattern of factors that influence the percentage of stunted children in East Nusa Tenggara Province is as follows.

$$ \hat{f}(x) = 35.74 + 0.539 \times BF_1 - 2.129 \times BF_2 - 3.223 \times BF_3 $$

where,

$$ BF_1 = \max (0, X_1 - 2) $$

$$ BF_2 = \max (0, X_6 - 5.2) $$

$$ BF_3 = \max (0, 5.2 - X_6) $$

On the criteria of $BF = 12$, $Mi = 1$, $MO = 1$. Based on the model, the predictor variables that contributed to the model were the percentage of pregnant women at risk of chronic energy
deficiency ($X_1$) contributing to the MARS model of 100%, and the percentage of infants with lower body weight ($X_6$). The importance level of each variable in the model is 100% and 37.775%.

References

[1] TNP2K, "100 Kabupaten/Kota Prioritas untuk Intervensi Anak Kerdil (Stunting)," Health Research and Development Agency, Jakarta, 2017.

[2] The Indonesian Ministry of Health, "Hasil Utama RISKESDAS 2017," Health Research and Development Agency, Jakarta, 2017.

[3] F. O. Aridiyah, R. Ninna and M. R, "Faktor-faktor yang Mempengaruhi Kejadian Stunting pada Anak Balita di Wilayah Pdesaan dan Perkotaan," Pustaka Kesehatan Journal, vol. 3, pp. 163-169, 2015.

[4] S. D. Anindita, "Pemodelan Persentase Balita Stunting di Indonesia Menggunakan Regresi Non Parametrik Spline Truncated," Surabaya: Statistics Department, FMKSD ITS, 2018.

[5] J. Jihad, Ainurrafiq and L. A. Ahmad, "Analisis Determinan Kejadian Stunting pada Balita Usia 12-24 Bulan," Kendari: Haluoleo University, 2016.

[6] J. H. Friedman, "Multivariate Adaptive Regression Splines," The Annals of Statistics, vol. 19, pp. 1-141, 1991.

[7] E. O. Permatasari, "Pendekatan Boosting Multivariate Adaptive Regression Splines untuk Klasifikasi Kemiskinan di Provinsi Jawa Timur," Surabaya: Statistics Department, FMKSD ITS, 2013.

[8] F. Nasuha, "Pemodelan Tingkat Pengangguran Terbuka di Kabupaten/Kota Provinsi Jawa Tengah dengan Menggunakan Metode Multivariate Adaptive Regression Spline (MARS)," Surabaya: Statistics Department, FMKSD ITS, 2016.

[9] R. Johnson and G. Bhattacharyya, Statistics Principle and Method, United State of America: John Wiley & Sons, Inc, 2010.

[10] N. Drapper and H. Smith, Analisis Regresi Terapan, Jakarta: PT. Gramedia Pustaka Utama, 1992.

[11] T. Hastie, R. Tibshirani and J. H. Friedman, The Elements of Statistical and Addiptive Modelling, California: Stanford, 2008.

[12] R. L. Eubank, Spline Smoothing and Nonparametric Regression, New York: Cambridge University Press, 1999.