MIMS Majalah Ilmiah Matematika dan Statistika 24 (2), 2024, 161 – 173
eISSN 2722-9886 pISSN 1411-6669
https://jurnal.unej.ac.id/index.php/MIMS/index

Modeling factors affecting poverty in Nusa Tenggara using the multivariate adaptive regression spline (MARS) method

Mulyasrihuda Arizal[1]*, Nurul Fitriyani[2], Bulqis Nebula Syechah[1]

[1]Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Mataram, Mataram, Indonesia
[2] Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Mataram, Mataram, Indonesia

*Corresponding: mulyasrihuda@gmail.com

Received: 04-09-2023, accepted: 15-03-2024

Abstract

Poverty is still a fundamental problem in the economy of various regions or developing countries, including Indonesia. The Nusa Tenggara Islands are one of the regions in Central Indonesia, where East Nusa Tenggara Province is in 3rd place with 20.44%, and West Nusa Tenggara Province is in 8th place with 13.83%, the highest percentage of poor people in Indonesia. A study was conducted to model the factors that influence poverty in Nusa Tenggara and determine the factors that significantly affect the percentage of poverty in Nusa Tenggara. Poverty data caused by many predictor variables that interact with each other can be said to be high-dimensional data where the relationship between the response variable and the predictor variable does not show a specific pattern, so one of the appropriate nonparametric regression methods to use for this approach is the Multivariate Adaptive Regression Spline (MARS) method. The data used in this study is secondary data with 12 predictor variables. The results of this study indicate that the best model was the model with the values of basis function (BF) of 24, maximum interaction (MI) of 2, and minimum observation (MO) of one where this model has the minimum GCV value of 0,3523701.

Keywords: Generalized cross-validation, multivariate adaptive regression spline, Nusa Tenggara Region, poverty

MSC2020: 62G08

1. Introduction

Poverty is when a person or household has difficulty meeting basic needs. At the same time, the supporting environment lacks opportunities to improve welfare on an ongoing basis or to get out of vulnerability [1]. One of the causes of poverty is a lack of income and assets to meet basic needs such as food, clothing, housing, and the level of health and education received. Besides that, poverty is also related to limited employment opportunities. Usually, they are categorized as poor without a job (unemployed), and their level of education and health is generally inadequate [2]. Poverty is still a fundamental problem where essential aspects such as the availability of basic needs are crucial factors

in everyday life. Poverty is a disease in the economy in almost every country, especially in developing countries like Indonesia, which still has a reasonably high poverty rate compared to several surrounding countries.

West Nusa Tenggara Province and East Nusa Tenggara Province are provinces in Central Indonesia which are members of Nusa Tenggara, where West Nusa Tenggara Province is in 8th place with the highest percentage of poor people in Indonesia, amounting to 13.83%. The Province East Nusa Tenggara is in 3rd place with a percentage of 20.44% [3]. According to the [4], there has been a decline in the number of poor people in West Nusa Tenggara (NTB) Province from 746,660 people (14.14%) in March 2021 to 735,030 people (13.83%) in September 2021. The decline in poor people also occurred in the Province of East Nusa Tenggara, from 1,169,310 people (20.99%) in March 2021 to 1,162,879 people (20.44%) in September 2021. This shows that the local government's poverty alleviation program in West Nusa Tenggara Province and East Nusa Tenggara (NTT) Province is quite maximal. Nevertheless, according to the Regional Medium-Term Development Plan, this achievement is still below the poverty reduction target, where the poverty rate in NTB in 2021 is at 11.75% while the poverty rate in NTT is at 12% in 2023. It is necessary to study what factors influence poverty in the two provinces.

In explaining the relationship pattern between response and predictor variables and estimating the regression curve, regression analysis can be used with a parametric or nonparametric regression approach [5], [6]. Regression analysis is a statistical method commonly used to see the effect of the predictor variable on the response variable [7]. If the parametric model assumptions are unmet, the regression curve can be performed using a nonparametric model approach. This is because the nonparametric regression method has high flexibility in estimating the regression curve [8]. The nonparametric adaptive regression approach is in demand, for example, Regression Tree, Recursive Partitioning Regression (RPR), and Multivariate Adaptive Regression Spline (MARS). The MARS method has the advantage of dealing with high-dimensional data problems (curse of dimensionality) and overcoming the weakness of RPR to produce a continuous model on knots. This method was introduced by [9]. This is a nonparametric regression method that assumes the form of the function of the relationship between the response variable and the predictor is unknown [10]. According to [11], [12], and [13], this method has been applied in various fields of knowledge, such as medicine, business, molecular biology, health, engineering, and other areas.

Based on the description above, the authors are interested in conducting a more in-depth study of what factors influence poverty in Nusa Tenggara through the variables that have been used in previous studies and see which variables have a significant effect with a statistical tool so that results in comparability with the methods that have been used so far. Previous studies on poverty have been carried out by previous researchers, including modelling the effect of the human development index on poverty in Indonesia using penalized basis spline nonparametric regression [14]. Other research was conducted to

model the poverty data in West Nusa Tenggara Province using panel data regression analysis [15]. A nonparametric regression approach is used with limited information, the form of function, and an unclear relationship pattern between the response variable, the percentage of poverty, and the factors that are thought to influence it. One approach in nonparametric regression that can be used is the Multivariate Adaptive Regression Spline (MARS). The MARS method uses the estimated regression curve with the data fitting approach; this method is very good at modelling data that has a changing pattern at certain sub-intervals, such as poverty data, by dividing the curve segmentally. Poverty data with more than three predictor variables and an amorphous pattern is very suitable for modeling using the MARS method, which can handle high-dimensional data. Therefore, this study models the factors affecting poverty in Nusa Tenggara using the Multivariate Adaptive Regression Spline (MARS) method.

2. Methods

The data used in this study is secondary data obtained from the Central Bureau of Statistics of West Nusa Tenggara and East Nusa Tenggara Provinces. The variables in this study consist of response variables (Y) and predictor variables (X). The response variable in this study was the percentage of poor people in West Nusa Tenggara and East Nusa Tenggara Provinces. This research consists of several predictor variables, namely the illiteracy rate of the population aged 15 years and over by district/city (X_1), School Participation Rate of the population aged 16-18 years by district/city (X_2), Percentage of Labor Force Participation Rate by district/city (X_3), percentage of open unemployment rate by district/city (X_4), Average monthly net income of informal workers by district/city (X_5), Gross Regional Domestic Product at current prices by district/city (X_6), Gross Regional Domestic Product at constant 2010 prices by district/city (X_7), percentage of population by district/city (X_8), percentage of households with proper sanitation by district/city (X_9), percentage of households accessing natural resources and proper drinking water by district/city (X_{10}), percentage of the population who have had health complaints during the past month by district/city (X_{11}), and the percentage of the population with BPJS PBI health insurance by district/city (X_{12}).

3. Results

Descriptive analysis was carried out to get an overview of the data used. This analysis aims to determine the characteristics of poverty data in Nusa Tenggara based on several predictor variables that have been described previously. The characteristics of the data raised in this study are the average value, minimum value, maximum value, variance value, and scatterplot. The results of calculations for the characteristics or descriptive statistics of other variables can be seen in the following table.

Table 1. Descriptive statistics of response and predictors variable

Variable	Minimum	Maximum	Average	Variance
Y	8,65	34,27	19,41	54,49
X_1	0,87	21,00	7,69	25,19
X_2	61,59	89,32	76,03	35,80
X_3	62,34	83,33	72,46	21,27
X_4	0,97	9,76	3,37	2,89
X_5	547811,00	1535896,00	950195,68	$6{,}068\times 10^{10}$
X_6	1230,40	24344,49	7866,45	46850336,99
X_7	767,26	16530,28	5226,74	22490549,63
X_8	0,813065	12,46	3,12	6,23
X_9	42,15	93,06	77,14	208,73
X_{10}	49,98	99,93	87,83	142,13
X_{11}	13,88	54,65	33,15	76,34
X_{12}	23,81	84,97	49,63	162,95

From Table 1, it can be seen that the variance value of the response variable and predictor variables, where the variance value indicates the variation or variety of the data. The variance value of the response variable (Y) is 54,492, meaning that data on the percentage of poor people in each district and city in Nusa Tenggara tends to vary. The variance value of the predictor variable, which shows that the data range is not too varied, is the variance value of the variables X_4 and X_8. The variance values that show quite a variety of data are the variance values of the variables $X_1, X_2, X_3, X_9, X_{10}, X_{11}$ and X_{12}. Whereas the variance values of the predictor variables X_5, X_6, and X_7 indicate that these variables have a very varied range of data. From Table 1 can also be seen the minimum, maximum, and average values. The minimum column indicates the smallest value of the response variable and predictor variables. While the maximum column shows the largest value of the response variable and the predictor variable. The average column shows the number of central tendency of the response variable and predictor variables.

Multicolliniaerity Check

Multicollinearity checking is carried out to see whether or not there is a high correlation between the predictor variables in a regression model. Decision-making in the multicollinearity test is based on the VIF value; where to obtain the VIF value, the R_j^2 value is first calculated by looking for a multiple linear regression model. For example, the R_j^2 value of the X_1 variable is obtained by assuming that the variable X_1 is used as the response variable, and the variables $X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11,}$ and X_{12} as predictor variables to find the multiple linear regression model. The VIF value of each predictor variable can be seen in Table 2.

Based on Table 2, it can be seen that the VIF values for $X_1, X_2, X_3, X_4, X_5, X_8, X_9, X_{10}, X_{11}$ and $X_{12} \leq 10$, it can be said that there is no correlation between predictor variables. So, based on the decision-making criteria, it can be concluded that the predictor variables do not occur in multicollinearity. While the VIF values for X_6 and $X_7 > 10$ are $X_6 = 1352{,}762$ and $X_7 = 1367{,}6611$, it can be said that there is a correlation between variables, so based on the decision-making criteria, it can be concluded that

multicollinearity occurs in the two predictor variables. Based on the multicollinearity test that has been carried out, it is found that there is multicollinearity in several predictor variables, so poverty data and its predictor variables cannot be analyzed using a parametric regression approach and are more precisely analyzed using a nonparametric regression approach with the MARS method.

Tabel 2 Multicollinearity checking results

Variable	VIF
X_1	2,199
X_2	1,420
X_3	2,725
X_4	2,674
X_5	2,853
X_6	1352,762
X_7	1367,661
X_8	3,773
X_9	3,531
X_{10}	2,936
X_{11}	1,923
X_{12}	1,992

Modeling Multivariate Adaptive Regression Spline (MARS)

MARS modeling is done by trial and error using a stepwise method (forward and backward) by combining the values of BF, MI, and MO so that the best value is obtained based on the minimum GCV value. The use of BF, MI, and MO values in this study is by the recommendations given by Friedman (1991), where the maximum number of interaction variables used is 1, 2, and 3; the minimum observations used are 0, 1, 2, and 3; while the maximum selection of the number of function bases is two to four times the number of predictor variables, namely 24, 36, and 48.

The selection of the MARS model with the minimum GCV value was carried out using the forward stepwise and backward stepwise methods. Forward stepwise process performed to obtain the maximum number of basis functions that can be written as follows.

$$BF_m(x) = \prod_{k=1}^{K_m} [s_{kM}(x_{np(k,M)} - t_{kM})]_+$$

The maximum basis function that will be included in the model is determined by the researcher, namely as many as 24, 36, and 48. The basis function obtained is used to estimate the regression coefficient α_m or estimators $\hat{\boldsymbol{\alpha}}$. Estimator $\hat{\boldsymbol{\alpha}}$ obtained by minimizing the sum of the squared errors using the least squares method or Ordinary Least Squares (OLS). Based on the nonparametric regression function, the MARS model can also be expressed in the following equation.

$$y_i = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} [s_{km}(x_{ip(k,m)} - t_{km})] + \varepsilon_i$$

which in matrix form can be written

$$\boldsymbol{Y} = \boldsymbol{B\alpha} + \boldsymbol{\varepsilon}$$

with,

$\mathbf{Y} = (y_1, y_2, \dots, y_{32})^T$,

$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_{12})^T$,

$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{32})^T$

$$\boldsymbol{B} = \begin{bmatrix} 1 & \prod_{k=1}^{K_1} [s_{k1}(x_{11(k,1)} - t_{k1})]_+ & \cdots & \prod_{k=1}^{K_m} [s_{k1}(x_{1p(k,M)} - t_{kM})]_+ \\ 1 & \prod_{k=1}^{K_1} [s_{k1}(x_{21(k,1)} - t_{k1})]_+ & \cdots & \prod_{k=1}^{K_m} [s_{kM}(x_{2p(k,M)} - t_{kM})]_+ \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \prod_{k=1}^{K_1} [s_{k1}(x_{np(k,1)} - t_{(k,1)})]_+ & \cdots & \prod_{k=1}^{K_m} [s_{kM}(x_{np(k,M)} - t_{kM})]_+ \end{bmatrix}$$

So, through estimation using the OLS method, the equation is obtained:

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{B^T B})^{-1} \boldsymbol{B^T Y}.$$

The forward stepwise process, which has been done previously, provides a model with many function bases and is very complex, so it must be done to remove some of the function bases to get a simpler model. After the forward stepwise process is done, the backward stepwise process is carried out to select the basis function that returns the minimum Generalized Cross-Validation (GCV) value. The backward stepwise process is done by removing the basis function, which has a small contribution to the estimated value of the response variable. The estimated value of the response variable can be written in the form:

$$\widehat{f_M}(x_i) = \hat{f}(x) = \sum_{m=1}^{M} \alpha_m (x) BF_m(x).$$

Based on the least squares criterion, the function basis with the smallest contribution is the function basis which, if removed from the previous model, will cause the smallest increase in ASR value. Backward stepwise done so that the model obtained produces a model that meets the parsimony concept (a simple model) and is done by minimizing the GCV value. The calculation of the GCV value can be done using the following formula.

$$GCV = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f_M}(x_i)\right)^2}{\left\{1 - \frac{C(\widetilde{M})}{n}\right\}^2}$$

Calculations are carried out similarly for all combinations of BF, MI, and MO values until the minimum GCV value is found. The calculation of the GCV value can also be done with the *software* Rstudio. The MARS modeling results obtained can be seen in the Table 3.

Table 3. The results of modeling MARS

No	BF	MI	MO	GCV	MSE
1	24	1	0	0,4623093	0,2821712
2	24	1	1	0,4876929	0,3471954
3	24	1	2	0,4044304	0,1741737
4	24	1	3	0,4223785	0,3006972
5	24	2	0	0,5737267	0,4550874
****6**	**24**	**2**	**1**	**0,3523701**	**0,1517531**
7	24	2	2	0,3814591	0,1642807
8	24	2	3	0,3772251	0,2034546
9	24	3	0	0,5737267	0,4550874
10	24	3	1	0,3899397	0,09748493
11	24	3	2	0,3814591	0,1642807
12	24	3	3	0,3540592	0,1909465
13	36	1	0	0,4623093	0,2821712
14	36	1	1	0,4876929	0,3471954
15	36	1	2	0,4044304	0,1741737
16	36	1	3	0,4223785	0,3006972
17	36	2	0	0,5737267	0,4550874
****18**	**36**	**2**	**1**	**0,3523701**	**0,1517531**
19	36	2	2	0,3814591	0,1642807
20	36	2	3	0,3772251	0,2034546
21	36	3	0	0,5737267	0,4550874
22	36	3	1	0,3899397	0,09748493
23	36	3	2	0,3814591	0,1642807
24	36	3	3	0,3540592	0,1909465
25	48	1	0	0,4623093	0,2821712
26	48	1	1	0,4876929	0,3471954
27	48	1	2	0,4044304	0,1741737
28	48	1	3	0,4223785	0,3006972
29	48	2	0	0,5737267	0,4550874
****30**	**48**	**2**	**1**	**0,3523701**	**0,1517531**
31	48	2	2	0,3814591	0,1642807
32	48	2	3	0,3772251	0,2034546
33	48	3	0	0,5737267	0,4550874
34	48	3	1	0,3899397	0,09748493
35	48	3	2	0,3814591	0,1642807
36	48	3	3	0,3540592	0,1909465

Information:
** is a model with a minimum GCV value

Based on the selection criteria for the MARS model, the best model for factors affecting poverty in Nusa Tenggara is a model with a minimum GCV value of 0,3523701 and an MSE value of 0,1517531 with a combination of BF = 24, MI = 2, MO = 1 as following.

$$\hat{f}(x) = 0{,}6077630 - 1{,}3874417BF_1 + 0{,}2129876BF_2 + 2{,}5255635BF_3.$$

Variable Importance

From the MARS model formed, it can be seen which variables significantly affect the model. The criteria used to estimate the importance of variables in the MARS model are $nsubstes$, GCV, and RSS. This study used the GCV criteria to determine the variables that had a significant effect. In the GCV criteria, if a variable that decreases the GCV value is added, then this variable is considered to have a good influence on the model and vice versa. The decrease in the GCV value is made into a scale of $0 - 100$ to facilitate interpretation, where the largest decline has a scale of 100. The level of importance of each predictor variable can be seen in the Table 4.

Table 4. Variable importance results

Variable	GCV	RSS
X_{10}	100,0	100,0
X_5	47,7	52,3
X_4	30,2	32,8
X_{12}	30,2	32,8
X_1	0,0	0,0
X_2	0,0	0,0
X_3	0,0	0,0
X_6	0,0	0,0
X_7	0,0	0,0
X_8	0,0	0,0
X_9	0,0	0,0
X_{11}	0,0	0,0

Based on Table 4, it is shown that the variable that has the dominant influence on the poverty rate in Nusa Tenggara with the GCV criteria is the percentage of households that have access to proper drinking water sources (x_{10}) with a score of 100, followed by the variable average monthly net income of informal workers (x_5) with a score of 47,7, then the variable percentage of the open unemployment rate (x_4) and the percentage of the population who have BPJS PBI health insurance (x_{12}), which has the same score of 30,2. Meanwhile, other variables do not affect the poverty rate in Nusa Tenggara because the score is 0.

The RSS column in Table 4 shows the effect of the predictor variable on the poverty rate in Nusa Tenggara with the RSS (*Residual Sum of Squares*) criteria. The deternination of variable importance using RSS criteria is done by calculating the RSS decrease for each subset. Variables that cause a greater decrease in RSS are considered more important variables and vice versa. From the RSS column in Table 4, it is shown that the variable that has the dominant influence on the poverty rate in Nusa Tenggara is the percentage of

households that have access to proper drinking water sources (x_{10}) with a score of 100, followed by the variable average monthly net income of informal workers (x_5) with a score of 52,3, then the variable percentage of the open unemployment rate (x_4) and the percentage of the population who have BPJS PBI health insurance (x_{12}), which has the same score of 32,8. Meanwhile, other variables do not affect the poverty rate in Nusa Tenggara because the score is 0.

Model Goodness Measures

The goodness of a model can be seen from the value R^2 of the coefficient of determination, the greater the value that has been obtained from a model, the better the predictor variables in the model explain the variability of the response variable. The coefficient of determination can be calculated using the following formula:

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}$$

where, from the calculations that have been done above, the value R^2 is obtained at 0,80289 or 80,289%. This means that the ability of the predictor variables in the study is able to explain the variability of the response variable of 80,289%. In comparison, the remaining 19,711% is explained by variables other than the predictor variables in the study.

MARS Model Significance Test

Simultaneous regression coefficient testing

This test was conducted to evaluate the suitability of the model. The hypothesis in this test is :

$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$ (insignificant model)

$H_1 : \exists \alpha_m \neq 0$; m = 1,2,…,M (significant model)

Decision-making on the simultaneous test is based on the calculated F_{Value}, which can be calculated using the following equation:

$$F_{Value} = \frac{\sum_{i=1}^{32} (|\hat{y}_i| - |\bar{y}|)^2 / M - 1}{\sum_{i=1}^{32} (|y_i| - |\hat{y}_i|)^2 / N - M - 1}$$

Based on the value F_{Value} and F_{Table} obtained the decision taken is reject H_0, because the value is $F_{Value} > F_{\alpha(M-1;N-M-1)}$ or equal to 72,76835 > 3,340, so it can be concluded that the MARS model obtained shows a significant relationship between the response and the predictor variables.

Partial test for regression coefficient

This test was conducted to determine the effect of the predictor variable on the response variable. The hypothesis in this test is :

$H_0 : \alpha_m = 0$ (α_m no effect on the model)
$H_1 : \alpha_m \neq 0$; for every m, where m = 1,2,…,M (coefficient α_m affects the model)

Decision-making on the partial test is based on the t_{Value}, which can be calculated using the equation as follows :

$$t_{Value} = \frac{\widehat{\alpha_m}}{\sqrt{\frac{\sum_{i=1}^{n}(|y_i| - |\hat{y}_i|)^2}{N-M-1} \times c_m}}$$

Based on the calculations that have been done, it is found that the value of $|\mathrm{t_{Value}}| > t_{(\frac{\alpha}{2};N-M)}$, that is $7{,}82519 > 2{,}045$ for $t_{(BF_1)}$, $3{,}95287 > 2{,}045$ for $t_{(BF_2)}$, and $4{,}67891 > 2{,}045$ for $t_{(BF_3)}$. Based on the decision-making criteria from the partial test, the decision that can be taken is to reject H_0 so that it can be concluded that the coefficients α_1, α_2, and α_3 have a significant effect on the model.

Testing Residual Assumptions

Normality Test

The normality test is carried out to determine whether the residuals are normally distributed. To find out the assumption of normality, a test can be used Kolmogorov-Smirnov with the following hypothesis test:
$H_0 : F = F_0$ (Residuals are normally distributed)
$H_1 : F \neq F_0$ (Residuals are not normally distributed)

The residual value used in the normality test is the difference between the original value of the response variable (Y) and its predicted value obtained through modeling using the MARS method. From the calculations that have been done, it is obtained the value $D_{maks} = 0{,}06645$, and the critical value from the Kolmogorov-Smirnov test is $D_{table(32;\,0{,}05)} = 0{,}23424$. Based on the value of test statistics and critical values Kolmogorov-Smirnov above, it is obtained that the value $D_{maks} < D_{table(32;\,0{,}05)}$ are $0{,}06645 < 0{,}23424$, namely the decision H_0 is accepted, and it can be concluded that the residuals are normally distributed.

Homoscedasticity Test

The homoscedasticity test was carried out to determine the residual variance. If the variance from the residual of one observation to another remains, it is called homoscedasticity; if the opposite occurs, it is called heteroscedasticity. The hypothesis used in the homoscedasticity test with the Glejser test is as follows:
$H_0 : \sigma_i^2 = \sigma^2$ (Residual meets identical assumption)
$H_1 :$ minimal ada satu $\sigma_i^2 \neq \sigma^2$ (Residual does not meet the identical assumption)

Based on the F_{Value} dan F_{Table} values obtained, the decision taken is H_0 is accepted because the value $F_{Value} < F_{\alpha(p-1;n-p)}$ or as big as $0{,}515787 < 2{,}31$, so it can be concluded that the residuals in the model are identical or there is no heteroscedasticity.

Independence Test (Autocorrelation)

An independence test or autocorrelation test is performed to detect the presence of autocorrelation in a model. The autocorrelation test used in the residual assumption test can use the Durbin-Watson test with the following test hypotheses:

$H_0 : \rho = 0$ (No autocorrelation occurs)
$H_1 : \rho \neq 0$ (Autocorrelation occurs)

Based on the residual value obtained from modeling using the MARS method, it can be calculated the value of the Durbin-Watson test and, based on the decision-making criteria, is obtained $4 - dU < d < 4 - dL$ namely $2{,}2677 < 2{,}379929 < 2{,}8231$ and it can be concluded that H_0 is accepted, which means there is no autocorrelation between predictor variables.

Interpretation of the MARS Model

The interpretation of the MARS model obtained in this study is:

a. $BF_1 = max(0, \quad X10 - (-0{,}42658))$
That is, the coefficient BF_1 which is $1{,}3874417$, will have meaning if the percentage of households accessing proper drinking water sources (x_{10}) is greater than $-0{,}42658$. However, if the value of the percentage of households accessing an adequate drinking water source (x_{10}) is less than $-0{,}42658$, then BF1 has no meaning; in other words, the value is 0. So that every time there is an increase of one basis BF_1 function in the value of the percentage of households accessing a source of proper drinking water (x_{10}), which is more than $-0{,}42658$, can increase the coefficient value by $1{,}3874417$.

b. $BF_2 = max(0, \ 0{,}92009 - X4) * max(0, X12 - (-0{,}29687))$
That is, the BF_2 coefficient with a value of 0,2129876 will have meaning if the percentage value of the open unemployment rate (x_4) is less than 0,92009, and the percentage value of the population who has BPJS PBI health insurance (x_{12}) is greater than -0,29687. However, if the percentage value of the open unemployment rate (x_4) is greater than 0,92009, and the percentage value of the population who has BPJS PBI health insurance (x_{12}) is less than -0.29687, then BF_2 has no meaning, or in other words, the value is 0. So that every time there is an increase in one BF_2 basis function in the percentage value of the open unemployment rate (x_4) less than 0,92009 and the percentage value of the population who has BPJS PBI health insurance (x_{12}) more than -0,29687 can increase the coefficient value by 0,2129876.

c. $BF_3 = max(0, \ -0{,}4024 - X5) * max(0, X10 - (-0{,}42658))$
That is, the coefficient BF_3 which is worth 2,5255635, will have meaning if the value of the average monthly net income of informal workers (x_5) is less than -0,4024, and the percentage value of households accessing decent drinking water sources (x_{10}) is greater than – 0,42658. However, if the value of the average monthly net income of

informal workers (x_5) is greater than -0,4024, and the value of the percentage of households accessing proper drinking water sources (x_{10}) is less than -0,42658, then BF_3 has no meaning, or in other words, the value is 0. So that every time there is an increase of one basis function BF_3 in the values of the average monthly net income of informal workers (x_5) is less than -0,4024, and the percentage value of households accessing decent drinking water sources (x_{10}) is more from -0,42658, it can increase the coefficient value by 2,5255635.

4. Conclusion

Based on the analysis and discussion that has been described previously, the following conclusions are obtained:

a.) The best MARS model is obtained by trial and error from a combination of BF = 24, MI = 2, and MO = 1, with the smallest GCV value of 0,3523701.

b.) The predictor variables that significantly affect poverty in Nusa Tenggara based on the best model obtained are the percentage of households accessing a proper drinking water source (x_{10}) with an importance level of 100%, the average monthly net income of informal workers (x_5) with a level of interest of 52.3%, the percentage of the open unemployment rate (x_4) and the percentage of the population that has BPJS PBI health insurance (x_{12}), which has the same level of interest of 32.8%.

References

[1] A. Cahyat, C. Gönner, and M. Haug, "Mengkaji kemiskinan dan kesejahteraan rumah tangga : Sebuah panduan dengan contoh dari Kutai Barat Indonesia, 978-979-1412-28-5," Bogor, 2007. [GreenVersion]

[2] Harlik, A. Amir, and Hardiani, "Faktor-faktor yang mempengaruhi kemiskinan dan pengangguran di Kota Jambi," *Jurnal Perspektif Pembiyaan dan Pembangunan Daerah*, vol. 1, no. 2, pp. 2338–4603, 2013. [CrossRef]

[3] Central Bureau of Statistics, "Provinsi Nusa Tenggara Barat Dalam Angka," Mataram, 2021. [GreenVersion]

[4] Central Bureau of Statistics, "Provinsi Nusa Tenggara Barat Dalam Angka," Mataram, 2022. [CrossRef]

[5] I.D.A.M.I. Wulandari and I.N. Budiantara, "Analisis faktor-faktor yang mempengaruhi persentase penduduk miskin dan pengeluaran perkapita makanan di Jawa Timur menggunakan regresi nonparametrik birespon spline," *Jurnal Sains dan Seni ITS*, vol. 3, no. 1, pp. 2337–3520, 2014. [CrossRef]

[6] N. Fitriyani and I.N. Budiantara, "Curve estimation and estimator properties of the nonparametric regression truncated spline with a matrix approach," *E-Jurnal Matematika*, vol. 11, no. 1, pp. 64–69, 2022. [CrossRef]

[7] P. Bilski, "Analysis of the Ensemble of Regression Algorithms for the Analog Circuit Parametric Identification," *Measurement*, vol. 160, no. 107829, pp. 1–9, 2020. [CrossRef]

[8] Y. Matdoan, "Pemodelan multivariate adaptive regression spline (MARS) pada faktor-faktor yang mempengaruhi kemiskinan di Provinsi Maluku dan Maluku Utara," *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 13, no. 1, pp. 8–14, 2020. [GreenVersion]

[9] J.H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991. [GreenVersion]

[10] Kishartini, D. Safitri, and D. Ispriyanti, "Multivariate adaptive regression spline (MARS) untuk klasifikasi status kerja di Kabupaten Demak," *Jurnal Gaussian*, vol. 3, no. 4, pp. 711–718, 2014. [CrossRef]

[11] E.K. Koc and H. Bozdogan, "Model selection in multivariate adaptive regression splines (MARS) using information complexity as the fitness function," *Mach Learn*, vol. 101, pp. 35–58, 2015. [CrossRef]

[12] N. Fitriyani, I.N. Budiantara, I. Zain, and V. Ratnasari, "Nonparametric regression spline in the estimation of the average number of children born alive per woman," in *The 1st International Conference on Science and Technology (ICST)*, Mataram: University of Mataram, 2016, pp. 169–172. [CrossRef]

[13] M. Hadijati, Irwansyah, and N. Fitriyani, "Autoregressive prewhitening on the nonparametric regression model of water discharge in the Jangkok Watershed, Lombok Island," *Global Journal of Pure and Applied Mathematics*, vol. 18, no. 1, pp. 307–318, 2022. [GreenVersion]

[14] N. Hasanah, S. Bahri, and N. Fitriyani, "The effect of human development index on poverty model in indonesia using penalized basis spline nonparametric regression," in *IOP Conf. Series: Materials Science and Engineering, 1115, 012055*, IOP Publishing, 2021, pp. 1–5. [GreenVersion]

[15] S. Aodia, N. Fitriyani, and Marwan, "Poverty data modeling in West Nusa Tenggara Province using panel data regression analysis," in *In Proceeding of the 5th International Conference on Science (ICST)*, Mataram: University of Mataram, 2021, pp. 698–708. [CrossRef]