CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 6(4) (2021), Pages 238-245
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Multivariate Adaptive Regression Splines and Bootstrap Aggregating Multivariate Adaptive Regression Splines of Poverty in Central Java

Ria Dhea LN Karisma[1], Juhari [2], Ramadani A. Rosa[3]

[1,2,3] Department of Mathematics UIN Maulana Malik Ibrahim Malang

Email: riadhea@uin-malang.ac.id, juhari@uin-malang.ac.id, ramadaniauiyanarosa@gmail.com

ABSTRACT

Poverty population is one of the serious problems in Indonesia. The percentage of population poverty used as a means for a statistical instrument to be guidelines to create standard policies and evaluations to reduce poverty. The aims of the research are to determine model population poverty using Multivariate Adaptive Regression Spline and Bagging MARS then to understand the most influence variable population poverty of Central Java Province in 2018. The result of this research is the Bagging MARS model showed better accuracy than the MARS model. Since, GCV in the Bagging MARS model is 0,009798721 and GCV in the MARS model is 6,985571. The most influence variable population poverty of Central Java Province in 2018 based on MARS model is the percentage of the old school expectation rate. Then, the most influentce variable based on Bagging MARS model is the number of diarrhea disease.

Keywords: Multivariate Adaptive Regression Splines; Bootstrap Aggregating; Generalized Cross-Validation; Poverty

INTRODUCTION

Poverty has concerned problem in the world even in Indonesia. In Indonesia, which is developing country, poverty has been affected in economics that it's showed level of welfare. Therefore, it has become a serious problem that must be resolved.

The growth of economics is the fundamental factor to reduce poverty. Based on BPS data, Indonesia has been able to deal with some economics global problem and succeeded in increasing economic growth. Some programs realized such as credit procurement programs, agricultural development, equitable development, infrastructure improvement, to the procurement program Inpress lagging Village (IDT) to help improve the community's living standards. The efforts considered significant because it reduced the experiencing of gaps between the rich and the underprivileged people and as an effort to realize the strategy of human Quality Development [1].

According to BPS (Badan Pusat Statistik) or Indonesian Statistics Institution, level of poverty in Indonesia has been reduced in recently. The percentage of privileged people reduced up to 0, 58% (in year-on-year) and at 2017 was the lowest poverty level rate. The

Submitted: November 25, 2020 Reviewed: February 19, 2021 Accepted: April 11, 2021
DOI: http://dx.doi.org/10.18860/ca.v6i4.10871

government succeeded in reducing poverty rate by 1.18 million people from an average. The government made a system for implement social protection based on a life cycle approach at 2018. However, in some areas poverty rate was slowed in reducing poverty [2].

MARS were introduced assumption about the relationship between the dependent and independent variables to estimate general functions of high dimensional data. Bagging MARS is a method that improved performance of in MARS method used bootstrap replicating. The past researches Karisma & Sri Harini [3] used MARS to find the classification of risk factors of ischemic and hemorrhagic patients by MARS method, Kilinc, B *et al.* [4] research to find models of metal concentrations to determine soil pollution by MARS method, etc.

MARS model used combination from spline method and recursive partition. Then, model in spline regression applied using a set of basis function to achieve q-order spline regression and estimated using least squares method. It has knot to find out the continuity basis function from one region in regression line to others. Otherwise, Bootstrap Aggregating (Bagging) used to minimize squared error value. The aimed of the research was the influenced poverty factor using MARS and Bagging MARS then it can be used for guidelines standard policies and evaluation to reduce poverty.

METHODS

Poverty is resident who have an average monthly expenditure per capita below the poverty line [5]. Poverty influenced by some factors such as human resources, employment, inflation, unemployment, population density, health facilities, income, scarcity, transportation, education, business capital [6] .

MARS method used multivariate nonparametric approaches. It has recursive partition formed, high dimensional data, and discontinuity data. Bagging MARS is a method that used for improve performance on MARS method with bootstrap replicating. MARS developed by Recursive Partitioning Regression (RPR) to estimate sub-region in each region continuous model in knots [7]. The advantage MARS is unrequired standardization, produced accurate results, used in big data, and used for regression analysis and classification simultaneously. Bagging MARS Recursive Partitioning Regression (RPR) unable to overcome the discontinuous data in knots. Therefore, the RPR algorithm used to estimate and correlate data in subregions [8]. The basis function explained the relationship between the dependent and independent variables [9] . The regression model used basis functions (BF) as follows:

$$y = \beta_0 \sum_{m=1}^{M} \beta_m h_m(x) \quad (1)$$

Where h_m is a set of basis function, and β_m is a coefficient of h_m in splines basis function defined as:

$$h_m = \prod_{k=1}^{K_m} \left[S_{km} \left(x_{v(k,m)} - t_{km} \right) \right]_+ \quad (2)$$

After modified BF with the RPR model, the MARS model obtained as follows:

$$\hat{f}(x) = a_0 + \sum_{m=1}^{M} a_m \prod_{k=1}^{K_m} \left[S_{km} \left(x_{v(k,m)} - t_{km} \right) \right]_+ \quad (3)$$

where a_0 is a coefficient, a_m is a coeefficient function basis-m M is a maximum basis, K_m is an interction degree, $x_{v(k,m)}$ is label of predictor variables, t_{km} is knot of predictor variables $x_{v(k,m)}$, and S_{km} are variables that take values ± 1 [7].

In matrix formed, MARS model defined by (4)

$$Y = Ba + \varepsilon\,, Y = (Y_1, \dots, Y_n)^T, a = (a_0, \dots, a_M)^T, \varepsilon = (\varepsilon_0, \dots, \varepsilon_n)^T \quad (4)$$

$$B = \begin{bmatrix} 1 \prod_{k=1}^{K_m}\left[S_{1m}.\left(x_{v(1,m)} - t_{1m}\right)\right] & \dots & \prod_{k=1}^{K_m}\left[S_{Mm}.\left(x_{v(M,m)} - t_{1m}\right)\right] \\ 1 \prod_{k=1}^{K_m}\left[S_{2m}.\left(x_{v(1,m)} - t_{1m}\right)\right] & \dots & \prod_{k=1}^{K_m}\left[S_{Mm}.\left(x_{v(M,m)} - t_{1m}\right)\right] \\ \vdots & \vdots & \vdots \\ 1 \prod_{k=1}^{K_m}\left[S_{nm}.\left(x_{v(1,m)} - t_{1m}\right)\right] & \dots & \prod_{k=1}^{K_m}\left[S_{Mm}.\left(x_{v(M,m)} - t_{1m}\right)\right] \end{bmatrix} \quad (5)$$

The GCV used to find the best model from MARS method, which used smaller is better. It is determined value by trial and error combining the number of basis functions (BF), maximum interaction (MI), and minimum observation (MO) [4]. The GCV defined as:

$$GCV = \frac{MSE}{\left[1 - \frac{C(\widehat{M})}{n}\right]^2} \quad (6)$$

where MSE value defined as $\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_M(x_i)\right]^2$, and $\mathrm{C}(\widehat{\mathrm{M}})$ defined as

$$\mathrm{C}(\widehat{\mathrm{M}}) = \mathrm{C}(\mathrm{M}) + \mathrm{dM} \quad (7)$$

where, C(M) is matrix trace $[\mathrm{B}(\mathrm{B}^{\mathrm{T}}\mathrm{B})^{-1}\mathrm{B}^{\mathrm{T}}] + 1$ that is the number of parameters being fit and d represents a cost for each basis function optimization [7].

The research used data from Sosial Ekonomi Nasional (Susenas), BPS (Badan Pusat Statistik) or Indonesian Statistics Institution for Java Province, and BPS Semarang Regional. Total data that used in this research was 350. It used MARS and Bagging MARS to analyze, then the steps that employed are divided data into training and testing data. Then, MARS method resolved by determined data used MARS method with a combination Basis Function (BF), Maximum Interaction (MI), and Minimal Observations (MO)[10]. Besides, obtained minimum GCV value to determine the best model in MARS and interpreted MARS model.

Bagging MARS method completed by determined Bagging MARS model using 50 replications. Then, the best model in Bagging MARS method achieved. The last is determined variable that the most influenced of poverty in Central Java Province in 2018.

RESULTS AND DISCUSSION

Statistics Descriptive

The descriptive analysis used to determine characteristic poverty in Central Java at 2018 (Badan Pusat Statistik, 2019)

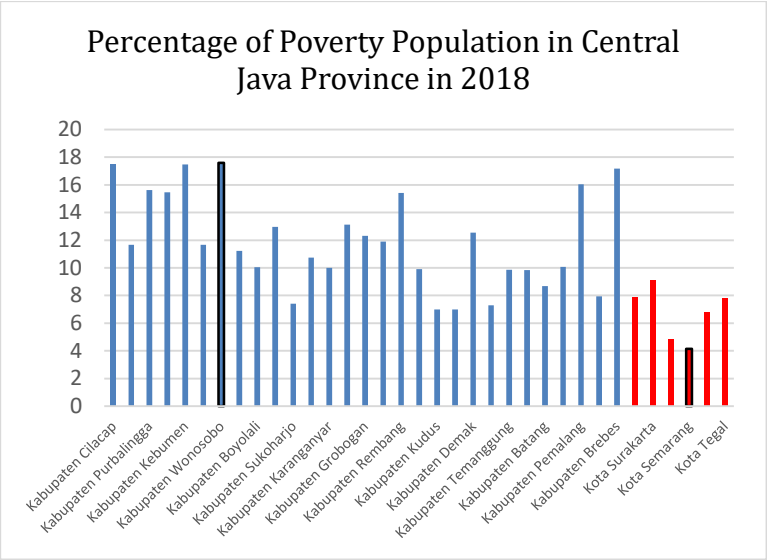

Figure 1. Descriptive Analysis Poverty Population

Figure 1 showed the percentage of population poverty in Central Java at 2018. The histogram illustrated regency areas and the percentage of poverty population in those areas. The highest poverty in those areas was Kabupaten Wonosobo with 17,58%. The total of poverty population was almost fifth percent. The percentage of population poverty occurred by some factors such as social economic, technology, health care and others. Then, the lowest population poverty was Kota Semarang from total population. It was under one in twenty percent.

Modeling Poverty Population MARS and Bagging MARS Methods

The MARS model showed in matrix pattern (*see Figure 3.2).* The matrix plot discovered relationship between response variable, which is variable the percentage of population poverty (Y), and predictor variables, which is the number of diarrhea disease (X_1), the number of life expectancy (X_2), the percentage of Human Development Index (HDI) (X_3), the percentage of expenditure per capita by non-food commodities (X_4), the percentage of open unemployement (X_5), the number of infant malnutrition (X_6), the percentage of family planning and birth control (X_7), the percentage of labor force participation rate (X_8), the percentage of expectation old school (X_9), the number of BPJS participants (X_{10}).

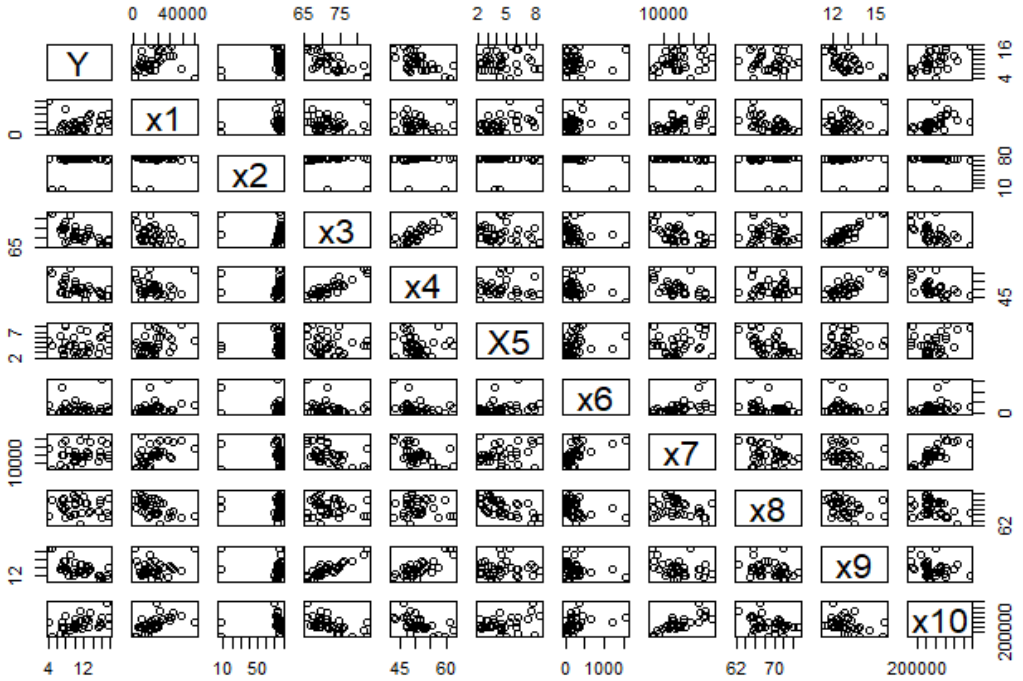

Figure 2. Matrix Plot Pattern of Poverty Population

Figure 2 illustrated that indicated unclear and difficult patterns of the relationship between variables. Then, in each variable has different characteristics on those areas and predictor variable was not able to be explained. In addition, nonparametric method used in this research which is MARS and Bagging MARS methods. The best model even in MARS and Bagging MARS methods indicated by the GCV.

The GCV in MARS model was 6.985571 and the R-Sq value was 75,7 %. Then, it was five predictor variables that significant and affected population poverty. It was X_1, X_6, X_9, X_8, X_{10} using training data 85% and testing data 15%. The MARS model obtained:

$$f(x) = 12.8 - 0.000235 * \max(0, X_1 - 19574) + 0.0107 * \max(0, 249 - X_6) - 0.514 * \max(0, X_8 - 67.5) + 7.35 * \max(0, 124 - X_9) - 1.34e - 05 * \max(0, 597322 - X_{10})$$

Then, the interpretation of MARS model is

$- 0.000235 * \max(0, X_1 - 19574)$

When, the value of X_1 was greater than 19574, for every increased number of diarrhea, it increased the percentage of the population poverty at 0.000235 in the Central Java Province with an average number of cases of diarrhea less than 19574 people.

$0.0107 * \max(0, 249 - X_6)$

when, the value of X_6 was smaller than 249, for every increased number of infant malnutrition, it increased the percentage of the population poverty at 0.0107 in the Central Java province with an average number of infant malnutrition less than 249 people.

$-0.514 * \max(0, X_8 - 68)$

when, the value of X_8 greater than 68, for every, increased in the percentage of labor force participation rate, it decreased the percentage of the population poverty by 0.514 in the Central Java province with an average percentage participation rate of a labor force more than 68 people.

$7.35 * \max(0, 12.4 - X_9)$
when, the value of X_9 is smaller than 12.4, for every increased the percentage of old school expectancy, it decreased the percentage of the population poverty at 7.35 in the Central Java province with an average percentage of the old school expectancy is less than 12.4%.

$-1,34e^{-05} * \max(0, 597322 - X_{10})$
When, the value of X_{10} was smaller than 597322, for every increased number of participants BPJS, it decreased the percentage of the population poverty of 0.0000134 in the Central Java province with an average number of participants BPJS less than 597322 people.

In Bagging MARS method that used 50 times replicate the best model obtained at the 49th replication using minimum GCV. Then, it was six predictor variables that have significant value affected population of poverty. It was $X_1, X_4, X_6, X_7, X_8, X_{10}$. The GCV was 0.009431298 and R-Sq value 0.7955023. The model was:

$\hat{f}(x) = 11.17643 - 0.0001232638 * \max(0, 13503 - X_1) + 0.0001346581 * \max(0, X_1 - 13503) + 1.637211 * \max(0, 48.96 - X_4) - 0.6424541 * \max(0, X_4 - 48.96) - 0.0250127 * \max(0, X_6 - 52) + 8.251765e - 05 * \max(0, 33664 - X_7) - 0.0001611239 * \max(0, X_7 - 33664) - 0.07994066 * \max(0, 67.03\ X_8) - 0,1345248 * \max(0, X_8 - 67.03) + 1.335112e - 05 * \max(0, X_{10} - 763837)$ (5)

Table 1. Comparison MARS and Bagging MARS Model

	Significance variables	GCV
MARS	$X_1, X_6, X_8, X_9, X_{10}$	6.985571
Bagging MARS	$X_1, X_4, X_6, X_7, X_8, X_{10}$	0.009798721

Table 1 showed that the GCV of the Bagging MARS model was 0.009798721. Then, MARS model was 6.985571. GCV in the Bagging MARS model indicated a better accuracy than the MARS model. Since, Bagging MARS model has GCV minimum than MARS model.

Best Variable In MARS and Bagging MARS Methods

The population poverty of Central Java using MARS model affected by the number of diarrhea disease (X_1), the number of infant malnutrition (X_6), the percentage of labor force participation rate (X_8), the percentage of expectation old school (X_9), and the number of participants BPJS (X_{10}). Table 2 is affected population poverty based on importance variables from MARS method.

Table 2. Importance Variables MARS Model

Variable	Importance Variables (%)
X_1	40.9
X_6	22.9
X_8	31.7
X_9	100
X_{10}	50.8

Moreover, Bagging MARS affected variable by importance variables that showed in Table 3. The variables were the number of diarrhea disease (X_1), the percentage of expenditure per capita by non-food commodities (X_4), the percentage of family planning and birth control (X_7), the percentage of labor force participation rate (X_8), the percentage of old school expectancy (X_9), and the number of participants BPJS (X_{10}).

Table 3. Importance Variables Bagging MARS Model

Variable	Importance Variables
X_1	95.32921
X_4	0.000000
X_7	60.80385
X_8	0.000000
X_{10}	0.000000

MARS and Bagging MARS method have distinction in importance variables. In MARS method the best level of importance variable was 100% which is the percentage of old school expectancy (X_9) then in Bagging MARS method was 95.33% which is number of cases of diarrhea disease (X_1).

CONCLUSIONS

Bagging MARS methods obtained better accuracy than the MARS model. The most influenced variable population of poverty in Central Java at 2018 using MARS method was the percentage of old school expectancy(X_9), then the Bagging MARS method is the variable number of cases of diarrhea disease(X_1).

REFERENCES

[1] Tjiptoherijanto, P. (1997). *Prospek Perekonomian Indonesia Dalam Rangka Globalisasi*. Rineka Cipta.

[2] Badan Pusat Statistik. (2017). *Perhitungan dan Analisis Kemiskinan Makro di Indonesia*.

[3] Karisma & Sri Harini. (2019). Multivariate Adaptive Regression Spline in Ishemic and Hemorrhagic. *Journal AIP Converence Proceedings of Symposium on BioMathematics*, 1–8.

[4] Kilinc, B. K., Malkoc, S., Koparal, A. S., & Yazici, B. (2017). Using multivariate adaptive regression splines to estimate pollution in soil. *International Journal of Advanced and Applied Sciences*. https://doi.org/10.21833/ijaas.2017.02.002

[5] Badan Pusat Statistik. (2019). Kemiskinan dan Ketimpangan. *Badan Pusat Statistik - Kemiskinan dan Ketimpangan*. https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html

[6] Kurniawan, M. D. (2017). Analisis Faktor-faktor Penyebab Kemiskinan di Kabupaten Musi Banyuasin (StudiKasus di kecamatan Sungai Lilin). *Jurnal Ilmiah Ekonomi Global Masa Kini*.

[7] Friedman, J. H. (1991). Rejoinder: Multivariate Adaptive Regression Splines. *The Annals of Statistics*. https://doi.org/10.1214/aos/1176347973

Rahmaniah, M. Nanda dkk, (2016). Bootstrap Aggregating Multivariate Adaptive

[8] Regression Spline. Jurnal Eksponensial, 7(2), 163–170. *Jurnal Eksponensial*, *7*(2), 163–170.

[9] Breiman, L. (1996). Bagging predictors. *Machine Learning*. https://doi.org/10.1007/bf00058655

[10] Shofa, B. & I. N. &. (2012). Analisis Survival dengan Pendekatan Multivariate Adaptive Regression Spline pada Kasus Demam Berdarah Dengue (DBD). *Jurnal Sains Dan Seni ITS*, *1*(1), 318–323.

[11[Badan Pusat Statistik. (2019). *https://semarangkab.bps.go.id*. Retrieved from https://semarangkab.bps.go.id/indicator/23/78/1/persentase-penduduk-miskin-kabupaten-kota-di-jawa-tengah.html