DOI: https://doi.org/10.34123/jurnalasks.v16i1.542

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 1, June 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Performance Study of Prediction Intervals with Random Forest for Poverty Data Analysis

Nina Valentika[1], Khairil Anwar Notodiputro[2], Bagus Sartono[3*]

[1]Universitas Pamulang, South Tangerang, Indonesia, [2,3]IPB University, Bogor, Indonesia
**Corresponding Author: E-mail address:* bagusco@apps.ipb.ac.id

ARTICLE INFO

Article history:
Received 18 September, 2023
Revised 9 April, 2024
Accepted 9 April, 2024
Published 30 June, 2024

Keywords:
LM; SPI; Quant; HDR; CHDR.

Abstract

Introduction/Main Objectives: Determine the prediction interval with for analyzing poverty data at the Regency/City level in Indonesia. **Background Problems:** Poverty will be a topic in various discussion and debates in the future. **Novelty:** This study's methods for constructed prediction intervals are LM, Quant, SPI, HDR, and CHDR. This method can improve the prediction interval performance with Random Forests. **Research Methods:** The method for building forests and obtaining BOP in this study is CART with the LS splitting rule. **Finding/Results:** The results of this study are that the best method for one replication is HDR with 500 trees. The best method for 100 repetitions is LM. Based on hypothesis testing, there is sufficient evidence to say no difference between the LM, SPI, Quant, HDR, and CHDR methods for 100 replications at a 5% significance level.

1. Introduction

The goal of predictive modeling in the concept of building a model is to predict unknown responses from observations given the covariates. Prediction models in their simplest form aim to provide point predictions for new observations. However, point predictions do not contain information about their precision that could tell how close to the actual response the prediction is expected to be, which is often important in decision-making contexts. Therefore, although point prediction is often the primary goal of predictive analysis, assessing its reliability is equally important, and this can be achieved with prediction intervals. A prediction interval consists of a series of probability values for an actual response with an associated confidence level, usually 90% or 95%. Given that shorter prediction intervals are more informative, developing predictive models that can produce shorter prediction intervals along with point predictions is critical in assessing and measuring prediction error. In real-world applications, knowing the error of predictions other than point predictions can increase the practical value of those predictions. The classic and most commonly used approach to construct prediction intervals is the parametric approach. However, its main weakness is that its validity and performance depend heavily on the assumed functional relationship between covariates and responses [1]. Roy & Larocque [1] have reviewed a new method that improves the performance of prediction intervals with Random Forests. The two aspects explored by Roy & Larocque [1] are the method used to construct the forest and the method used to construct the prediction interval. Four methods for building forests, three from the Classification And Regression Tree (CART) paradigm and the transformation forest method. The

method to build a forest and get Bag of Observations for prediction (BOP) which has been studied by Roy & Larocque [1] is CART with splitting rules, namely Least-Squares (LS), splitting L1, and Shortest Prediction Interval (SPI). The prediction interval is constructed using the BOP, which is the set of nearest-neighbor observations [1]. Methods for constructing prediction intervals have been reviewed by Roy & Larocque [1] are the Classical method (LM), Quantile, Shortest Prediction Interval (SPI), Highest Density Region (HDR), and Contiguous HDR (CHDR).

The LM is calculated based on an intercept-only linear model using the BOP as a sample and produces a symmetric prediction interval around the prediction point. Similar to the Quantile Regression Forest (QRF) method, the quantile method is based on BOP quantiles. SPI corresponds to the shortest interval among the intervals containing the least $(1-\alpha)100\%$ number of observations in the BOP. As an alternative to SPI, HDR is the smallest region in the BOP, with the desired $(1-\alpha)$ coverage. Note that HDR is not necessarily one interval. If the distribution is multimodal, it can be formed with several intervals. CHDR is a way to obtain a single prediction interval from an HDR interval by constructing an interval with the minimum and maximum boundaries of the HDR interval [1].

Alakus et al. [2] created a Package RFpredInterval that implements 16 methods for constructing prediction intervals with Random Forests and Boosted Forests. Alakus et al. [2] also carried out a simulation regarding the splitting rule method least-squares (LS) and prediction interval methods, namely: LM, Quant, SPI, HDR, and CHDR using the Ranger package in building Random Forest. However, there is another package for building Random Forest that is available in the RFpredInterval package, namely randomForestSRC. This research wants to examine the prediction interval for Random Forest with the random ForestSRC package. This research creates a 95% prediction interval using variations studied by Roy & Larocque [1] the method used in building the forest, namely CART with splitting rules. LSand methods for building prediction intervals are LM, Quant, SPI, HDR, and CHDR. This study also applied Out-Of-Bag (OOB) calibration and the acceptable coverage range was set to [0.945, 0.955].

Poverty is one of the problems that exists in developing countries, such as Indonesia. The Central Bureau of Statistics of Indonesia (BPS) uses the concept of the ability to meet basic needs to measure poverty. In this approach, poverty is understood as an economic inability to meet basic food and non-food needs as measured by expenditure [3].

Various factors that influence the Percentage of Poor Population (PPP) are Gross Regional Domestic Product (GRDP), Life Expectancy Rate (LER), Mean Years of Schooling (MYS), Expected Years of Schooling (EYS), and Real Per Capita Expenditure (PPK). There are 5 main characteristics, namely area of residence, gender, education level, number of household members, and work status of the head of the household, which have the potential to cause household poverty in Central Java [4]. According to [5], a region that has a high GDP means the region has a good economy. The opposite applies. The economy in question is an economy that can support people's lives so that poverty does not arise. In the economic field, development performance in achieving prosperity is measured based on Gross Domestic Product (GDP) and its growth rate [6]. The Health Dimension is measured by the life expectancy indicator [7]. Life Expectancy is a tool for evaluating the government's performance in improving the welfare of the population in general, and improving health status in particular [8]. According to Anggadini [9], the higher the life expectancy, the higher the quality of public health. In the Circle of Poverty Theory, the quality of public health is reflected in the increase in the life expectancy rate (LER). Increasing community productivity can encourage economic growth thereby reducing the poverty rate, namely the higher the life expectancy, the lower the poverty rate. StudyPramesti & Bendesa [10], found that there is an influence on poverty where increasing education will reduce poverty. Indonesia is a developing country and has a large population. The problem of poverty in Indonesia cannot be avoided (Aulele et al. [11]). Poverty will be a topic in various discussions and debates in the future [12]. Based on the description above, the prediction interval from PPP become an interesting topic for study. Thus, this research aims to determine the prediction interval with *Random Forest* for analyzing poverty data at the Regency/City level in Indonesia.

2. Material and Methods

The data used is secondary data that comes from the Central Statistics Agency (BPS). Study This using data from 514 districts /cities in Indonesia in 2021. Table 1 presents variables used in the study. The software used in this study is R.

Table 1. Variables Study

Role of Variable	Variable	Unit	Symbol
Response	PPP	Percent	Y
Explainer	GRDP	Billion Rupiah	X_1
	LER	Year	X_2
	MYS	Year	X_3
	EYS	Year	X_4
	PPK	Thousand Rupiah	X_5

The steps taken in this study are

1. Exploration and description.
2. Create prediction intervals for one repetition and 100 repetition.
   a. Divide training data and test data. Amount trees used are {200,500,1000,5000}.
      i. Divide training data and test data. Amount trees used are {200,500,1000,5000}.
      ii. For 100 repetitions, 70% training data and 30% test data (notated 70:30); 80% training data and 20% test data (notated 80:20); as well as 90% training data and 10% test data (notated 90:19).
   b. Deternain the prediction interval for all methods. The regression model used in Random Forest is
   $$Y = \sum_{i=1}^{5} X_i + \varepsilon. \quad (1)$$
   c. Rule to create prediction interval in Random Forest is as following:
      i. Build forest and get BOP. Method used in study This is method Least Square (LS).
      ii. Calculating Prediction Intervals using BOP. Method used in study This are LM, Quant, SPI, HDR, and CHDR. Function from packages RFpredInterval to use in study is rfpi(). Packages for Random Forest used in this study is randomForestRSC The type bootstrap used moment by root active is Sampling With Replacement (SWR).
      iii. Prediction Interval Calibration with the Use of OOB i=Information Desired coverage arranged to 95% for all methods. For all methods, calibration-based validation cross done as procedure calibration main, but also checked OOB calibration. OOB calibration for finding α_w. In the second procedure calibration, the range of possible coverage accepted is arranged to [0.945, 0.955].
3. Compare results For one repetition and 100 repetitions.
   a. For one repetition, determine the amount of tree best seen from the mark of the smallest Root Mean Absolute Error (MAE). Determine the method best seen from the mean of the length of the smallest prediction interval. Then, create a plot for all the methods that own a prediction interval for every observation, that is all methods except the HDR method.
   b. For 100 repetitions, the method compare to based on mean coverage, average length of prediction interval and percentage enhancement prediction interval length. Percentage enhancement prediction interval length for method I calculated as $100\ x\ (ml_i - ml^*)/ml^*$, where ml_i is the mean length of the prediction interval of the method $i$andml^* is mean length of prediction interval shortest from all method. Smaller value of this size shows better performance [2].

3. Results and Discussion

3.1. Data Description

Table 2. Data Description

	Y	X_1	X_2	X_3	X_4	X_5
Min	2.38	1,087	55.43	1,420	3.87	3976
Q_1	7.15	5,963	67.39	7,510	12.42	8574
Q_2	10.46	13,643	69.97	8,305	12.93	10196
Mean	12.27	37,222	69.66	8,437	13.02	10325
Q_3	14.89	30,895	72.04	9,338	13.65	11719
Max	41.66	861,000	77.73	12,830	17.80	23888

Based on Table 2, it is found that the smallest percentage of poor people for districts/cities in Indonesia in 2021 is 2.38%, namely Sawah Lunto City. The largest percentage of poor people for districts/cities in Indonesia in 2021 is 41.66%, namely Intan Jaya Regency. Correlations between variables are presented in Table 3.

Based on Table 2, it is found that the smallest percentage of poor people for districts/cities in Indonesia in 2021 is 2.38%, namely Sawah Lunto City. The largest percentage of poor people for districts/cities in Indonesia in 2021 is 41.66%, namely Intan Jaya Regency. Correlations between variables are presented in Table 3.

Table 3. Correlation beetween variables

	Y	X_1	X_2	X_3	X_4	X_5
Y	1.00	-0.08	-0.54	-0.54	-0.43	-0.64
X_1	-0.08	1.00	0.21	0.17	0.09	0.34
X_2	-0.54	0.21	1.00	0.42	0.37	0.57
X_3	-0.54	0.17	0.42	1.00	0.78	0.67
X_4	-0.43	0.09	0.37	0.78	1.00	0.52
X_5	-0.64	0.34	0.57	0.67	0.52	1.00

Based on Table 3, it is found that GRDP, LER, MYS, EYS, and PPK have a negative relationship with PPP.

3.2. Prediction Interval for One Repeat

This section presents the prediction intervals in one replication with the number of trees {200, 500, 1000, 5000} and the 70% rule for training data and 30% for test data. A comparison of all methods for prediction intervals in one replication is presented in Table 4.

Table 4. Comparison of all methods for prediction intervals in one replication

Number of Trees (MAE) (RMSE)	Method	Mean length Prediction Interval	*Coverage Levels* (in %)	α_w (in %)
200	L.M	16.6	92.9	5
(3,607)	SPI	17.4	94.2	2
(4,788)	Quant	17.5	94.2	3
	HDR	16.7	94.2	4
	CHDR	16.8	94.2	4
500	L.M	18.2	97.4	5
(3,219)	SPI	18.2	97.4	2.5
(4,325)	Quant	19.2	98.1	3
	HDR	17.1	98.1	4
	CHDR	17.1	97.4	5
1000	L.M	16.3	95.5	5
(3,343)	SPI	16.1	94.2	3
(4,435)	Quant	17.6	94.2	3
	HDR	16.9	96.8	3
	CHDR	15.7	94.8	4.5
5000	L.M	17.3	96.1	5
(3,657)	SPI	20.0	97.4	1
(4.8)	Quant	18.4	94.2	3
	HDR	19.4	97.4	0.4
	CHDR	18.6	95.5	2.8

Based on Table 4, it is found that the number of trees that have the smallest MAE and RMSE is 500. The method that has the smallest mean of prediction interval length for one repetition is HDR with 500 trees. Thus, the best method for a single replicate is HDR with 500 trees. Plots for methods that have only one prediction interval for each observation (all methods except the HDR method) are presented in Figures 1 to Figure 4. HDR allows multiple prediction intervals for one observation [2].

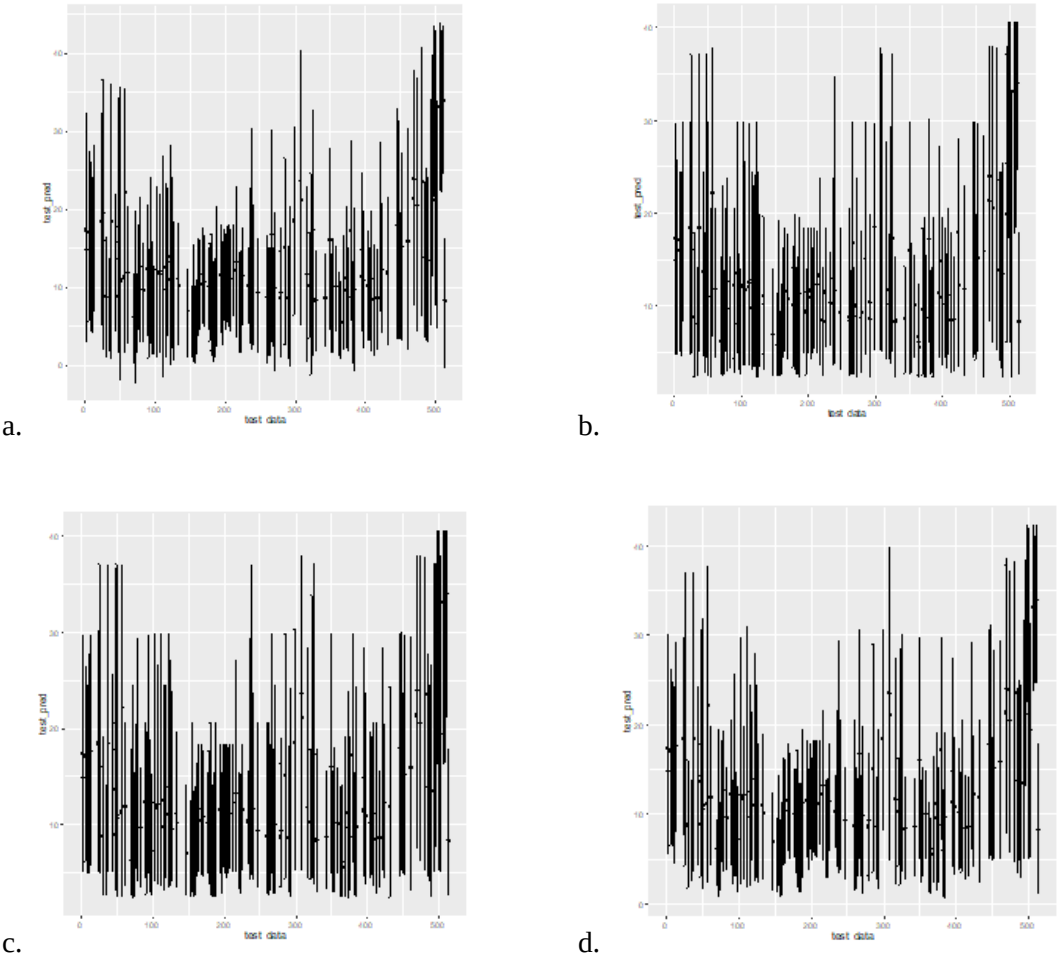

Figure. 1. (a) LS-LM method; (b) LS-SPI method; (c) LS-Quant method; (d) LS-CHDR method

Based on Figures 1, it is found that the response data in the test data is mostly within the prediction interval.

3.3. Prediction Interval for 100 Repetition

The measure used in this research to evaluate the following performance by Roy & Larocque [1] is by using mean coverage and average prediction interval length. Table 5 presents the mean level of coverage for each method from 100 replications.

Table 5. Mean level of coverage for each method from 100 replication

Number of trees (*Split)* *)	*Mean coverage* (in%)				
	L.M	SPI	Quant	HDR	CHDR
200(a)	95.2	95.9	95.3	95.6	95.5
200(b)	95.3	95.9	95.6	95.5	95.4
200(c)	95.4	95.4	95.1	95.8	95.7
500(a)	94.8	95.1	94.8	95.4	95.0
500(b)	95.0	95.3	95.2	95.2	95.2
500(c)	94.8	94.9	94.6	95.5	95.3
1000(a)	94.7	95.3	95.0	95.2	95.0
1000(b)	95.2	95.0	94.9	95.4	95.3
1000(c)	95.5	95.2	94.8	95.3	95.2
5000(a)	95.1	95.3	95.3	95.0	94.9
5000(b)	95.0	95.2	95.2	95.4	95.2
5000(c)	94.4	94.8	94.7	95.1	94.6

*)

(a) 70:30

(b) 80:20

(c) 90:10

Based on Table 5, most of the LM methods have a mean coverage difference with the desired coverage (95%) being the smallest compared to other methods. The LM method has better accuracy than other methods based on coverage. Thus, the LM method is indicated to be the best method based on mean coverage. Table 6 presents the average length of the prediction interval from 100 repetitions.

Table 6. Average length of prediction interval from 100 repetition

Number of trees (*Split)*)*	Average Prediction interval length				
	L.M	SPI	Quant	HDR	CHDR
200(a)	17.0	18.2	18.1	17.1	17.1
200(b)	16.8	17.8	17.9	17.0	17.1
200(c)	16.9	17.5	17.7	17.1	17.2
500(a)	16.7	17.4	17.6	16.8	16.8
500(b)	16.7	17.2	17.6	16.5	16.6
500(c)	16.6	16.9	17.3	16.3	16.5
1000(a)	16.7	17.3	17.5	16.9	16.7
1000(b)	16.5	17.0	17.3	16.6	16.4
1000(c)	16.4	16.6	17.0	16.1	16.3
5000(a)	16.7	17.2	17.5	18.4	17.0
5000(b)	16.5	16.9	17.3	18.3	17.0
5000(c)	16.4	16.5	17.0	17.9	16.7

*)

(a) 70:30

(b) 80:20

(c) 90:10

Based on Table 6, it is found that the LM method mostly has the smallest interval length compared to other methods. Thus, there is an indication that based on the interval length, the LM method has better accuracy compared to other methods. For all methods, for the most part, the average prediction interval length decreases as the sample size decreases. This result is different from Alakus et al. [2]. Based on Table 4, the number of trees is the best is 500. Figure 5, Figure 6, and Figure 7 illustrate mean coverage for all models with use amount tree 500 out of 100 repetitions.

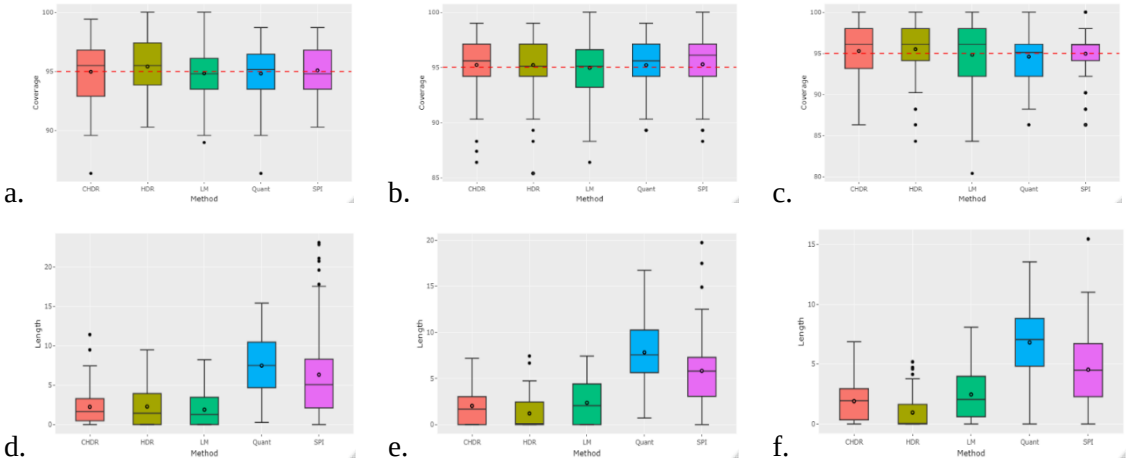

Figure. 2. (a) Mean coverage for all models from 100 replications for amount tree 500 split 70:30; (b) Mean coverage for all models from 100 replications For amount tree 500 split 80:20; (c) Mean coverage for all models from 100 replications For amount tree 500 split 90:10; (d) Mean coverage for all models from 100 replications For amount tree 500 split 90:10; (e) Percentage enhancement length of prediction interval from 100 repetitions For amount tree 500 split 80:20; (f) Percentage enhancement length of prediction interval from 100 repetitions For amount tree 500 split 90:10.

The red dotted line in Figure 2.a, Figure 2.b, and Figure 2.c is the desired level, namely 95%. The white circle is the average of the percentage increase in the length of the prediction interval from 100 replicates. Based on Table 4, the number of trees is the best is 500. Based on Figure 2.a, Figure 2.b, and Figure 2.c, it can be seen that all methods provide mean coverage that is close to the desired level. Figure 2.d, Figure 2.e, and Figure 2.f illustrate the percentage increase in the length of the prediction interval using several trees of 500 from 100 replications.

Based on Figure 2.d, Figure 2.e and Figure 2.f, the average percentage increase in the length of the smallest prediction interval in the 500 tree scenario in the 70:30 split is LM, the 80:20 split and the 90:10 split is HDR. The smaller the percentage increase, the better the method (Alakus et al., [2]. As a result, it is indicated that The best method based on the percentage increase in the length of the prediction interval is LM for split 70:30, HDR for split 80:20, and split 90:10.

Hypothesis testing is carried out with the following hypothesis:

H0: $\mu_{LM} = \mu_{SPI} = \mu_{Quant} = \mu_{HDR} = \mu_{CHDR}$(There is no difference between LM, SPI, Quant, HDR, and CHDR methods)

H1: $\mu_{LM} \neq \mu_{SPI} \neq \mu_{Quant} \neq \mu_{HDR} \neq \mu_{CHDR}$(There are differences between LM, SPI, Quant, HDR, and CHDR methods)

Table 7. Hypothesis testing results from mean coverage for 100 repetitions for each number of trees and split

Number of Trees (Split)*)	Fcount	F criteria
200(a)	1,802	2,390
200(b)	0.730	2,390
200(c)	0.711	2,390
500(a)	1,188	2,390
500(b)	0.245	2,390
500(c)	1,257	2,390
1000(a)	1,277	2,390
1000(b)	1,049	2,390
1000(c)	0.493	2,390
5000(a)	0.763	2,390
5000(b)	0.315	2,390
5000(c)	0.729	2,390
200(a)	1,802	2,390
200(b)	0.730	2,390
200(c)	0.711	2,390
500(a)	1,188	2,390
500(b)	0.245	2,390
500(c)	1,257	2,390
1000(a)	1,277	2,390
1000(b)	1,049	2,390
1000(c)	0.493	2,390
5000(a)	0.763	2,390
5000(b)	0.315	2,390
5000(c)	0.729	2,390

*)

(a) 70:30

(b) 80:20

(c) 90:10

Because Fcount is smaller than Fcriteria , then No reject H_0. So, based on Table 7, there is sufficient evidence to say that there is no difference between the LM, SPI, Quant, HDR, and CHDR methods for 100 repetitions at a 5% significance level

4. Conclusion

This research concludes that the best method for one replication is HDR with 500 trees. LM is the best method based on coverage, interval length, and percentage increase in prediction interval length for 100 repetitions. Based on hypothesis testing, there is sufficient evidence to say that there is no difference between the LM, SPI, Quant, HDR, and CHDR methods for 100 repetitions at a 5% significance level.

Ethics approval

Not required.

Acknowledgments

-

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

References

[1] M.-H. Roy and D. Larocque, 'Prediction intervals with random forests', *Statistical Methods in Medical Research*, vol. 29, no. 1, pp. 205–229, 2020.
[2] C. Alakus, D. Larocque, and A. Labbe, 'RFpredInterval: An R Package for Prediction Intervals with Random Forests and Boosted Forests', *arXiv preprint arXiv:2106. 08217*, 2021.
[3] N. Ellah, 'Analysis of the Influence of Factors that Influence Poverty in East Java', *Student Scientific Journal Feb*, vol. 4, no. 1, 2016.
[4] L. Sugiyono and S. Ningsih, 'Analysis of Potential Causes of Poor Households in Central Java', *Journal of Applications of Statistics & Statistical Computing*, vol. 10, no. 2, p. 25, 2018.
[5] R. K. Damanik and S. A. Sidauruk, 'The influence of population and GRDP on poverty in North Sumatra Province', *Journal of Darma Agung*, vol. 28, no. 3, pp. 358–368, 2020.
[6] A. Rinaldi, 'Structural equation model to analyze household welfare indicators', *Decimal: A Journal of Mathematics*, vol. 2, no. 3, pp. 281–288, 2019.
[7] A. S. Wicaksono and A. M. Yolanda, 'Grouping Regencies/Cities in East Nusa Tenggara Province Based on Human Development Index Indicators Using K-Medoids Clustering', *Journal of Applied Statistics*, vol. 1, no. 1, pp. 79–90, 2021.
[8] S. P. Sinaga, A. Wanto, and S. Solikhun, 'Implementasi Jaringan Syaraf Tiruan Resilient Backpropagation dalam Memprediksi Angka Harapan Hidup Masyarakat Sumatera Utara', *Jurnal Infomedia: Teknik Informatika, Multimedia, dan Jaringan*, vol. 4, no. 2, pp. 81–88, 2019.
[9] F. Anggadini, 'Analysis of the effect of life expectancy, literacy rate, open unemployment rate and gross regional domestic income per capita on poverty in districts/cities in Central Sulawesi Province in 2010-2013', *Catalogis E-Journal*, vol. 3, no. 7, pp. 40–49, 2015.
[10] N. L. Pramesti and I. K. G. Bendesa, 'The Influence of Socio-Economic Factors on Poverty in Bali Province', *EP Unud E-Journal*, vol. 7, no. 9, pp. 1887–1917, 2018.
[11] S. N. Aulele, V. Y. I. Ilwaru, E. R. Wuritimur, and M. Y. Matdoan, 'Analysis of the Number of Poor People in Maluku Province Using a Spatial Regression Approach', *Journal of Applications of Statistics & Statistical Computing*, vol. 13, no. 2, pp. 23–34, 2021.
[12] C. Suryawati, 'Understanding Poverty Multidimensionally', *Journal of Health Service Management*, vol. 8, no. 03, pp. 585–597, 2005.