International Journal of Business and Management Invention (IJBMI)
ISSN (Online): 2319 – 8028, ISSN (Print): 2319 – 801X
www.ijbmi.org || Volume 13 Issue 11 || November, 2024 || PP: 98-104

Human Development Index Prediction in East Kalimantan uses the FFNN (Feed Forward Neural Network) optimized MLR (Multiple Linear Regression) Method

1st Elvyani Nuri Harlawati Gaffar[1], 2nd Emmilya Umma Aziza Gaffar[2],
3th* Achmad Fanany Onnilita Gaffar[3]
{elvygaffar66@gmail.com[1], emmilya.gaffars@gmail.com[2], onnygaffar212@gamil.com[3]}
Faculty of Economics and Business, 17 Agustus 1945 University, Samarinda, East Kalimantan, Indonesia[1]
Department of Economic Study, Mulawarman University, Samarinda, East Kalimantan, Indonesia[2]
Department of Information Technology, Politeknik Negeri Samarinda, East Kalimantan, Indonesia[3]
**Corresponding Author*

Abstract. *A decent standard of living is an adequate standard for each person for himself and his family, which is related to the basic needs, and the improvement of sustainable living conditions. The poverty line is the minimum income level to obtain an adequate standard of living. On the other hand, HDI (Human Development Index) is an aggregate indicator that provides a complete picture of the quality of life. Two critical points shown by HDI are the ability of humans to improve their health and education, and the ability of humans to enjoy life on the results of their efforts that are reflected by their income. Of course, the Prosperous Family has fulfilled a decent standard of living, and vice versa for Poor Families. NPF (the Number of Prosperous Families) is the number of families that meet PFI (Prosperous Family Index) requirements. The number of Poor Families contributes to %PP (the Percentage of Poor People). Therefore it could be assumed that NPF has a positive influence on HDI, while the %PP has a negative influence. The results of this study showed that the assumption has been proven, where NPF has a positive influence on HDI* $(\beta_1 = 0.0525)$*, and vice versa %PP has a negative influence* $(\beta_2 = -0.1253)$*. Both variables (NPF and %PP) simultaneously have an influence on HDI of 47.75%* $(R_{square} = 0.4775)$*. For predictive needs, the MLR (Multiple Linear Regression) model was optimized by using FFNN (Feed Forward Neural network). The results of studies have shown that the MLR model with optimization provides much better performance* $(\%R_{square} = 98.66\%)$ *compared to the MLR model without optimization* $(\%R_{square} = 54.65\%)$ *in terms of predictive activity.*

Keywords: *NPF (Number of Prosperous Families), %PP (percentage of Poor People), HDI (Human Development Index), MLR (Multiple Linear Regression), FFNN (Feed Forward Neural network)*

Date of Submission: 06-11-2024 Date of acceptance: 18-11-2024

I. Introduction

Paying attention to human quality as a development resource and as a measure of the results of development efforts encourages the measurement of the quality of the population. The first attempt is to develop indicators that reflect the quality of life of the population. The next effort is to make a measure that can easily show the level of quality of human life and can be compared to other relevant indexes. This effort in question is making the Quality of Life Index (*QLI*).Some indicators measured in *QLI* are (1). Infant Mortality Rate, (2). Life Expectancy at Age One, (3). Literacy. Some related index measurements commonly used are: (1) Social Health Index (*SHI*), (2) Family Prosperity Index (*FPI*), and (3) Human Development Index (*HDI*)[1]. *SHI* has different parameters for each age domain (children, youth, adults, elderly). In general, *SHI* consists of homicide, alcohol-related traffic fatalities, food stamp coverage, affordable housing, and income inequality [2]. FPI provides an overview of the vital and central role played by families as engines of the economy. The family is the smallest unit of the community group entitled to protection by society and the state. The main functions of the family are: (1). As a vehicle to educate, nurture, and socialize children, (2). Develop the ability of all members to carry out their functions in society properly, and (3). Provide satisfaction and a healthy social environment for the attainment of a prosperous family. In other words, *FPI* provides a complete picture of prosperity and cultural well-being [3].

A decent standard of living is an adequate standard for each person for himself and his family, which is related to adequate food, clothing, and housing, and the improvement of sustainable living conditions[4]. This

standard is the basis for measuring welfare and poverty in life. BKKBN (National Population and Family Planning Agency) Indonesia defines a Prosperous Family as a family that has fulfilled the formulated decent standard of living. Prosperous Family Index (*PFI*, almost similar to*FPI*) presents the ability of families to fulfill (1) basic needs, (2) social-psychological needs, (3) development needs, (4) the need to donate material finance and actively participate as administrators in social activities, (5) tangible and sustainable social contributions. The Pre-Prosperous Family (Poor Family) is a family that has not been able to fulfill their basic needs [1].To measure poverty, BPS (Statistics Indonesia) - Indonesia uses the concept of ability to meet basic needs. Therefore, poverty is seen as an inability from the economic side to meet basic food and non-food needs as measured by expenditure. The method used is to calculate the *Poverty Line* (*PL*), which is an aggregation of two components, namely the Food Poverty Line (*FPL*) and the Non-Food Poverty Line (*NFPL*). *Poverty Line* calculation is done separately for urban and rural areas. The poverty line is the minimum income level to obtain an adequate standard of living. Poor people are residents who have an average per capita expenditure per month below the *Poverty Line*[5]. The percentage of Poor People (%*PP*) is a comparison between the number of poor people and the population. The ability of a family to fulfill basic needs is one component of *PFI*. The number of Prosperous Families (*NPF*) is the number of families that meet *PFI* requirements.

HDI is an aggregate indicator that provides a complete picture of the quality of life. Two critical points shown by *HDI* are the ability of humans to improve their health and education, and the ability of humans to enjoy life on the results of their efforts that are reflected by their income. The three indicators measured are: (1). Life expectancy level, (2). Education level, and (3). Level of income [5, 6].*HDI* is a concise measure of the average achievement/success of the main dimensions of human development, namely: longevity and healthy living, having the knowledge, and having a decent standard of living. In other words, the *HDI* is a geometric average of the dimensions of the index of health, education, and expenditure.

Of course, the Prosperous Family has fulfilled a decent standard of living, and vice versa for Poor Families. The number of Poor Families contributes to the percentage of poor people (%*PP*). Therefore, it can be assumed that NPF has a positive influence on HDI, while the% PP has a negative influence. Multiple linear regression is a technique that can be used to test these assumptions through the estimation of the model [7, 8].

This study measured the influence of *NPF* and %*PP* on *HDI*. The analysis of the intended influence is carried out using the *MLR* (Multiple Linear Regression) method. The *MLR* model obtained is then optimized using *FFNN* (Feed Forward Neural Network). Furthermore, *HDI*predictions are made using an optimized model.

II. Methods

This study uses the assumption that *NPF* and%*PP*, respectively as independent variables, influence*HDI* as the dependent variable. This section presents how the multiple linear regression method proves this assumption and how *FFNN* optimizes the performance of the models that have been made. Furthermore, this section also presents how to predict *HDI* by using the optimized model.

2.1 Multiple Linear Regression Method

The multiple linear regression model assumes a relationship between a dependent variable $y_i = y_1; y_2; \dots; y_M$ and a set of independent variables $x_i{}' = x_{i1}x_{i2} \dots x_{iN}$ (also called a regressor).Suppose a sample of M observations, $i = 1 \dots M$.Every single observation can be statedby:

$$y_i = x_i{}'.\beta + \varepsilon_i \quad (1)$$

where $x_i{}'$ is the $(N+1)$-dimension row vector,β is the $(N+1)$-dimension column vector of model parameters, and ε_i is the estimated error as residual (error part) that is sufficient for the value of y_i. The whole sample (a number of *N* observations) can be expressed by:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_M \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \dots & x_N \\ 1 & x_{21} & \dots & x_{2N} \\ \dots & \dots & \dots & \dots \\ 1 & x_{M1} & \dots & x_{MN} \end{bmatrix} . \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_N \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_M \end{bmatrix} \quad (2)$$

$$\boldsymbol{Y}_{(M,1)} = \boldsymbol{X}_{(M,(N+1))} * \beta_{((N+1),1)} + \boldsymbol{\varepsilon}_{(M,1)}$$

Estimation of all parameters and error parts uses Ordinary Least Square (*OLS*), the method most widely used in econometrics. The *OLS* estimator is based on the following assumptions: (1). Linearity (the functional relationship between independent and dependent variables is linear in its parameters), (2). Independence (all independent and dependent variable values of observations are independently and identically distributed), (3). Homogeneity (error parts normally distributed to independent variables. error parts and its mean are independent of independent variables. error parts and independent variables are not correlated), (4). Homoscedasticity (variance of error part is constant), (5). Conditionally heteroscedasticity (the variance of the error part may depend on independent variables), (6). Identifiability (all regressors are not perfectly linear)[9].

The square distance between the observed $(y_i)$and predicted $(x_i'.\beta)$dependent variables is expressed by[10]:

$$D(\beta) = \sum_{i=1}^{M} (y_i - x_i'.\beta)^2 = (\boldsymbol{Y} - \boldsymbol{X}.\beta)'(\boldsymbol{Y} - \boldsymbol{X}.\beta) \quad (3)$$

OLS minimizes $D(\beta)$ in such a way that *OLS* estimators are obtained as follows:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \quad (4)$$

Hence, the predicted dependent variable $(\hat{y}_i)$and error part(ε_i)can be expressed by:

$$\hat{y}_i = x_i'.\hat{\beta} \quad \varepsilon_i = y_i - \hat{y}_i \quad (5)$$

2.2 Goodness of Fit Statistics

The goodness of fit of the statistical model describes how well a model fits into a series of observations. Measuring goodness of fit usually summarizes the difference between the observed value and the expected value in the model. *SSE*(Sum Square Error) is a statistic that measures the total deviation of the response value from the fit to the response value. SSE stated by:

$$SSE = \sum_{i=1}^{M} (y_i - \hat{y}_i)^2 \quad (6)$$

R-square is a statistic that measures how successful the fit is in explaining variation in data. In other words, *the R*-square is the square of the correlation between the response value and the predicted response value. *R*-square stated by:

$$R_{square} = \frac{SSR}{SST} \quad SSR = \sum_{i=1}^{M} (\hat{y}_i - \bar{y})^2 \quad SST = \sum_{i=1}^{M} (y_i - \bar{y})^2 \quad (7)$$

where *SSR* is the sum square regression, *SST* is the total sum square, and $\bar{y}$ is the mean value of the dependent variable. *R*-square can be of any value in the interval $\{0..1\}$. *An R*-square close to 1 indicates that the model has accounted for the proportion of variance.

2.3 Feed Forward Neural Network (*FFNN*)

ANN(Artificial Neural Network) is an imitation of the natural neural network where artificial neurons are connected in the same way as a brain network. *ANN* consists of processing units called neurons. Artificial neurons try to mimic the structure and behavior of natural neurons consisting of one input (dendrites), and one output (synapses through axons). Various functions are used for activation in *ANN*, one of which is the sigmoid function. In general, *ANN* architecture consists of the input layer, hidden layer, and output layer. Forward Neural Network (*FFNN*) is a type of *ANN* that often has one or more hidden layers of sigmoid neurons followed by a linear neuron output layer. Multi-layers of neurons with nonlinear transfer functions allow the network to study linear or nonlinear relationships between input and output vectors through the training process[11]. *FFNN* uses back propagation learning in its training process as shown in Fig. 1.

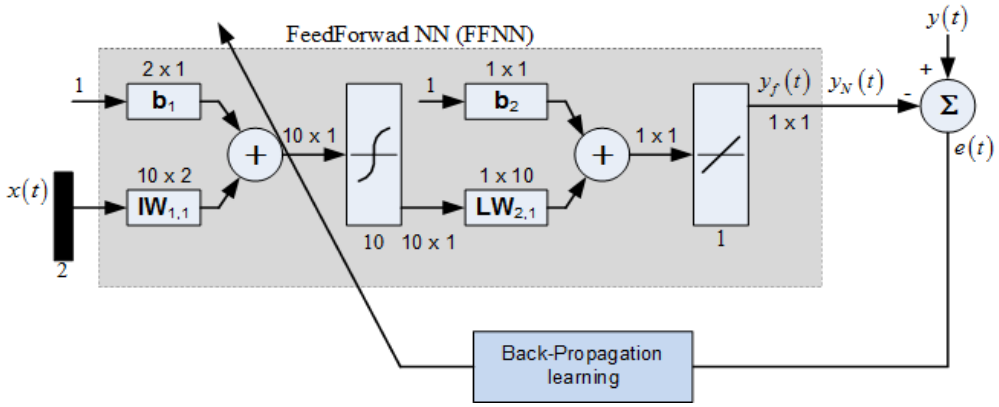


Fig. 1. Feed Forward NN with back-propagation learning.

Suppose there are a pair of observational data (x, y), $x = x_1 x_2 \ldots x_N$ is the training input data, and $y = y_1 y_2 \ldots y_N$ is the training target data. Mathematically stated by:

$$\hat{y} = f_{NN}(x) \quad e = y - \hat{y} \quad (8)$$

where $f_{NN}(x)$is an approximation function by FFNN, $\hat{y}$ is the predicted value, and e is the training error. If *FFNN* is trained such that $e \to 0$ then $f_{NN}(x) \to y$. Implementation of FFNN can be done easily by using the *Neural Network Toolbox*.

If the error part(ε_i)can be estimated using *FFNN*, such that:

$$\hat{\varepsilon}_i = f_{NN}(x_i) \quad (\varepsilon_i - \hat{\varepsilon}_i) \to 0 \tag{9}$$

then the *ANN*-optimized multiple linear regression model is stated by:

$$\tilde{y}_i = \hat{y}_i + f_{NN}(x_i) \tag{10}$$

where $f_{NN}(x_i)$ is the predicted error part by trained *FFNN*, $\hat{y}_i$ is the predicted dependent variable by *OLS*, and $\tilde{y}_i$ is the optimized dependent variable.The performance of the final model is measured using *MAPE*(Mean Absolute Percentage Error) expressed by:

$$MAPE = \frac{1}{M}\sum_{i=1}^{M}\frac{|y_i - \tilde{y}_i|}{y_i} \times 100 \tag{11}$$

2.4 Datasets

In this study, the datasets of *NPF*, *%PP*, and *HDI* from 9 districts in the period 2018-2023 have been taken from the catalog of Kalimantan Timur Province in Figures, 2024[12]. The datasets are graphically shown inFig. 2.

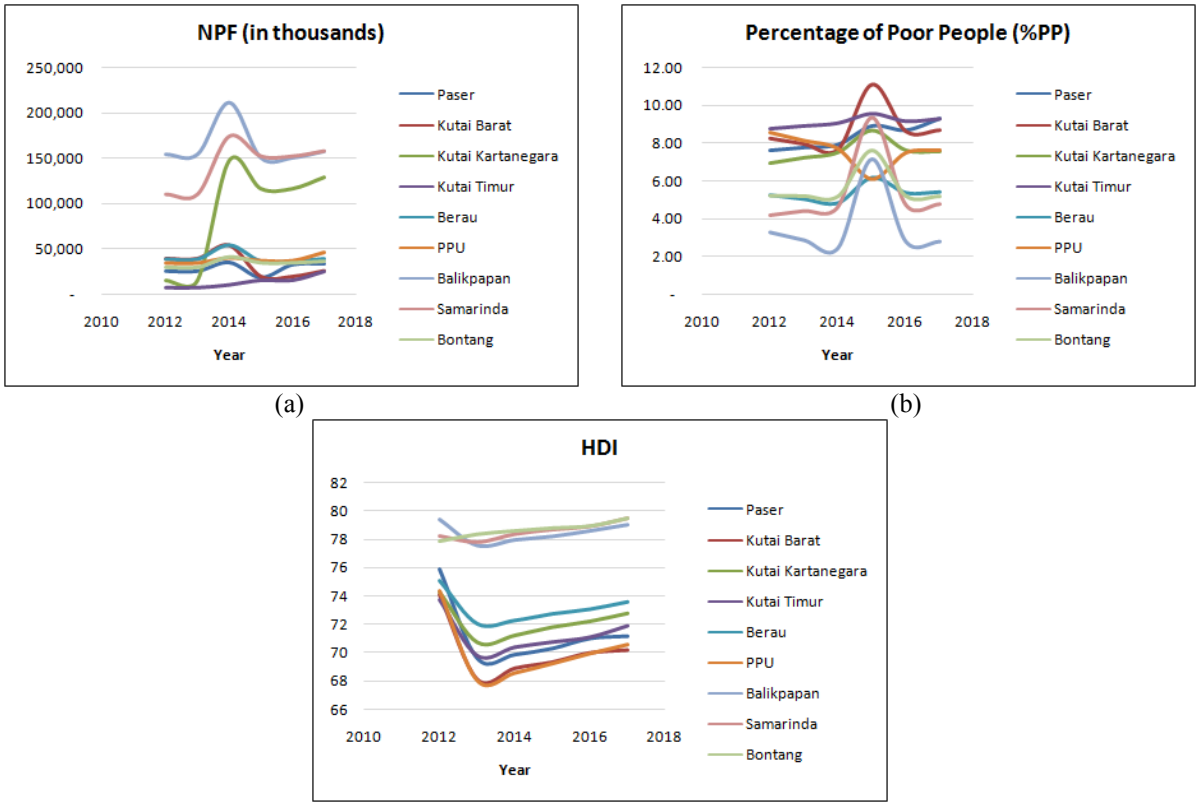


Fig. 2. Graphically Datasets

The dataset needs to be arranged in a row based on the year and number of districts so that the column headers are *NPF*, *%PP*, and *HDI* variables.An example of new datasets is shown inTable 1. All the data need to be normalized within interval {0 ... 1} by using the following formula:

$$x_{i(norm)} = x_i / max(X) \quad X = x_1, x_2, \ldots, x_N \tag{12}$$

Data no. 1-45 (years 2018-2022) are used to construct *the MLR* model, while data no. 49 - 54 (year 2023) are used for the validation of predictive results.

Table 1. An example of New Datasets

Year	District	No. of Data	$NPF(X_1)$	$\%PP(X_2)$	$HDI(Y)$
2018	Paser	1	25,965	7.64	75.85
	Kutai Barat	2	38,861	8.28	74.05
	Kutai Kartanegara	3	15,932	6.94	74.24
	Kutai Timur	4	7,798	8.77	73.75
	Berau	5	38,355	5.24	75.05
	PPU	6	34,660	8.57	74.35
	Balikpapan	7	154,169	3.30	79.38
	Samarinda	8	110,362	4.18	78.26
	Bontang	9	29,815	5.20	77.85

	.....	.....	.....	.....	.....
2023	Paser	46	34,157	9.28	71.16
	Kutai Barat	47	25,486	8.72	70.18
	Kutai Kartanegara	48	129,464	7.57	72.75
	Kutai Timur	49	24,861	9.29	71.91
	Berau	50	38,879	5.41	73.56
	PPU	51	46,020	7.63	70.59
	Balikpapan	52	157,576	2.82	79.01
	Samarinda	53	158,394	4.77	79.46
	Bontang	54	36,453	5.16	79.47

III. Results and Discussion

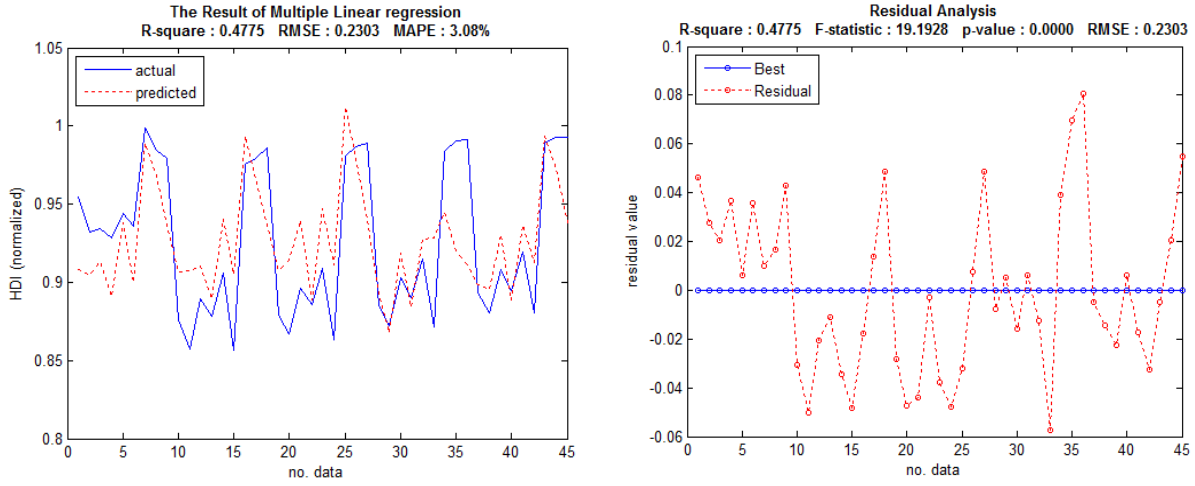
Firstly, the relationship between the independent variables (X_1 and X_2) and the dependent variable (Y) is expressed by Eq. (2) as follows:

$$Y = \beta_0 + \beta_1.X_1 + \beta_2.X_2 + \varepsilon$$

By using the OLS estimator, all parameters were found as follows:

$$\beta_0 = 0.9879 \quad \beta_1 = 0.0525 \quad \beta_2 = -0.1253$$

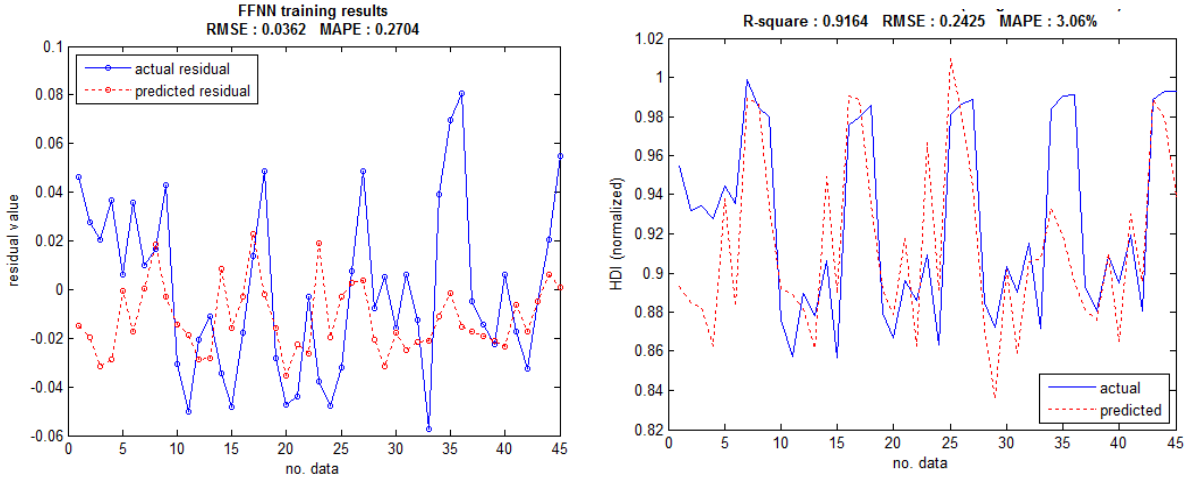
Next, all the estimated parameters were used to obtain the values of the predicted dependent variable with the results shown inFig. 3. Those figures showed that *p-value* = 0.0000 (< 0.05 significance value, [13]) indicated that the assumption of two independent variables influencing the dependent variable was acceptable. The two independent variables simultaneously affected the dependent variable of 47.75%(*R*-square=0.4775), where *NPF* had a positive influence($\beta_1 = 0.0525$), and %*PP* had a negative influence($\beta_2 = -0.1253$).Furthermore, those figures also showed the results of the residual analysis with *RMSE* = 0.2303 and *MAPE* = 3.08%.



(a) (b)

Fig. 3. The results of *MLR*

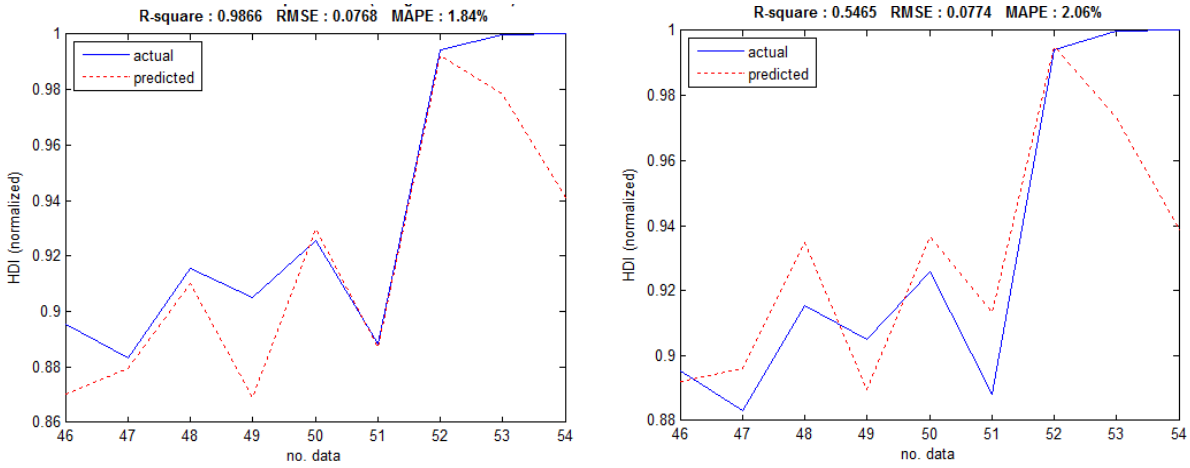
The *OLS* estimation of the *MLR*model results still contains theerror part (residual). The presence of an error part needs to be modeled to improve the performance of the model. The error part modeling is performed by using *FFNN* through training data pairs (x_i, ε_i). The values of the independent variables (x_i)were the training input data, while the error parts (ε_i)were the training target data.*FFNN* training was carried out in such a way as to comply with Eq. (9) with the results shown in **Error! Reference source not found.** (a).The model equation was then reconstructed, according to Eq. (10). Performance measurement of the *FFNN* optimized multiple linear regression model was performed by using data no. 1 - 45 (the year 2018-2022) with the results shown in **Error! Reference source not found.** (b).Those figures have shown increased model performance (*R-square*=0.9164, *MAPE*=3.06%).



(a) (b)

Fig. 4. Training and Validation results

Finally, the final model was used to predict *HDI* in 2023 with the performance of the predicted results measured by using data no. 46 -54 (year 2023). The results are shown inFig. 5. Those figures have shown a very significant difference in performance between the *MLR* models with and without *FFNN* optimization.A summary of the study results is shown in Table 2.



(a) (b)

Fig. 5. Prediction results

Fig. 6.

Table 2. The summary of the study results

Goodness of Fit	Using data from 2018-2022			Using data 2023		
	MLR Model	*FFNN*-optimized *MLR* Model		*MLR* Model	*FFNN*-optimized *MLR* Model	
		Results	Increase in performance		Results	Increase in performance
	$\beta_0 = 0.9879$ $\beta_1 = 0.0525$ $\beta_2 = -0.1253$					
$R_{square}(\%)$	47.75%	91.64%	91.92%	54.65%	98.66%	80.53%
$RMSE$	0.2303	0.2425	-5.30%	0.0774	0.0768	0.78%
$MAPE$	3.08%	3.06%	0.65%	2.06%	1.84%	10.68%

The application of the *MLR* model optimized using *FFNN* provides much better performance in terms of predictive activity, it can prove that the assumption of the influence of *NPF* and %*PP* on *HDI*, meaning that if *NPF* increases, *HDI* in East Kalimantan province will also increase, this condition is very much expected because with the increase in *NPF* means the level of welfare and quality of life of the population in East Kalimantan province will increase, but if %PP increases, it can reduce *HD*I in East Kalimantan province, this condition is very undesirable if %PP continues to increase because it can continue to increase the poor population and decrease the welfare and quality of life of the population in East Kalimantan province.

IV. CONCLUSION

This study has applied the *MLR* model to prove the assumption of the influence of *NPF* and %*PP* on *HDI*. The study results have shown that the assumption has been proven, where *NPF* has a positive influence on *HDI* ($\beta_1 = 0.0525$), and vice versa %*PP* has a negative influence ($\beta_2 = -0.1253$). Both variables (*NPF* and %*PP*) simultaneously have an influence on *HDI* of 47.75%($R_{square} = 0.4775$). For predictive needs, the *MLR* model was optimized by using *FFNN*. The results of studies have shown that the *MLR* model with optimization provides much better performance($\%R_{square} = 98.66\%$)compared to the *MLR* model without optimization($\%R_{square} = 54.65\%$) in terms of predictive activity.Performance improvement that occurred was equal to 80.53% on R_{square}, 0.78% on *RMSE*, and 10.68% on *MAPE*.This can prove that the assumption of the influence of NPF and %PP on HDI is that if NPF increases, then *HDI* in East Kalimantan province will increase, which means that the level of welfare and quality of life of the population in East Kalimantan province will increase, but if %*PP* increases, then *HDI* will decrease in East Kalimantan province. This can result in an increase in the poor population and a decrease in the welfare and quality of life of the population in East Kalimantan province.

References

[1] M. Dr. Ir. Euis Sunarti, "Prosperous Family Indicators: History of Development, Evaluation, and Sustainability," BKKBN (National Population and Family Planning Agency) - Indonesia (ISBN 978-602-8665-05-6)2006.
[2] D. Hitchcock and P. D. Marque-Luisa Miringoff, "Social Indicators: The Real Health of the States," ISSP Insight, 2008.
[3] D. I. Institute, "Family policy in a changing world: Promoting social protection and intergenerational solidarity," Department of Economic and Social Affairs - Division for Social Policy and Development - Programme on the Family, United Nations, Doha - Qatar2009.
[4] I. Frye, G. Wright, and T. Elsley, "Toward a Decen Life for All : Decent Standard of Living Index, Final Report," Labour Research Service, 2018.
[5] BPS-Jakarta, "Statistical Yearbook of Indonesia 2023," BPS-Jakarta, Ed., ed. Jakarta - Indonesia: BPS - Statistics Indonesia, 2023.
[6] P. D. Wendy P. Warcholik and M. A. J. Scott Moody, "Family Prosperity Index 2018," Family Prosperity Initiative, Farm Prosperity, 2018.
[7] G. Niedbala, M. Piekutowska, and M. Adamski, "Multiple Regression Analysis Model to Predict and Simulate Winter Rapeseed Yield," Journal of Research and Applications in Agricultural Engineering, vol. 63, 2018.
[8] E. Ostertagová, "Modelling using Polynomial Regression," Procedia Engineering, vol. 48, pp. 500-506, 2012.
[9] K. Schmidheiny, "The Multiple Linear Regression Model," Short Guides to Microeconometrics, Unversit□at Basel, 2019.
[10] Mathwork, "Statistics and Machine Learning Toolbox™ User's Guide," The Mathworks, Inc, 2019.
[11] M. T. H. Mark Hudson Beale, Howard B. Demuth, Neural Network Toolbox™. 3 Apple Hill Drive-Natick, MA 01760-2098: The MathWorks, Inc., 2015.
[12] B.-S. o. K. T. Province, "Kalimantan Timur Province in Figures, 2024," B.-S. o. K. T. Province, Ed., ed. Kalimantan Timur Province-Indonesia: BPS-Statistics of Kalimantan Timur Province, 2024.
[13] Mathwork, "Curve Fitting Toolbox™ User's Guide," The Mathworks, Inc, 2019.