JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Harumeka, A., & Purwa, T. (2023). Does the physical type of house still affect household poverty in Indonesia? An entropy-based fuzzy weighted logistic regression approach. *Journal of Information and Communication Technology, 22*(3), 337-361. https://doi.org/10.32890/jict2023.22.3.2

Does the Physical Type of House Still Affect Household Poverty in Indonesia? An Entropy-based Fuzzy Weighted Logistic Regression Approach

*[1]Ajiwasesa Harumeka & [2]Taly Purwa
[1]Central Bureau of Statistics, East Java Province, Indonesia
[2]Central Bureau of Statistics, Bali Province, Indonesia

*ajiwasesa@bps.go.id
taly@bps.go.id
*Corresponding Author

Received: 13/4/2022 Revised: 21/2/2023 Accepted: 22/3/2023 Published: 24/7/2023

ABSTRACT

Poverty is one of the biggest challenges facing the world nowadays. Numerous studies have concentrated on the characteristics that determine poverty to identify poor households. One of the most important factors is the physical type of the house. The physical type of houses includes floor type, wall type, roof type, and floor area per inhabitant in Indonesia, especially Surabaya, one of Indonesia's big cities and the capital of East Java Province. This factor gave promising results to the country. Therefore, it was assumed that these variables could no longer distinguish between those in wealth and those in poverty. Poor household data are one example of imbalanced data in terms of classification, which is a concern. The Rare Event Weighted Logistic Regression (RE-WLR) and Entropy-based Fuzzy Weighted

Logistic Regression (EFWLR) methods were utilised to overcome these problems. As a result, the only factor, including the physical design of a house, which had a substantial impact on the likelihood that a household would be classified as poor, was the floor area per capita. The other three variables were not statistically significant, namely floor type, wall type, and roof type. In addition, the elimination of the physical type of house would reduce the Area Under the Curve of the RE-WLR and EFWLR methods by 6.78 percent and 6.85 percent, respectively.

Keywords: Classification, imbalanced data, poor household, weighted logistic regression.

INTRODUCTION

Poverty is one of the biggest challenges facing the worlds. The world is committed to reducing the poverty rate as much as possible by including the issue in the first goal of Sustainable Development Goals (SDGs) (United Nations, 2021), i.e., "ending poverty in all its form everywhere". To identify poor households, many studies aimed to determine the factors that cause poverty at the macro and micro levels. At the macro level, the influencing factors include the Human Development Index (HDI), investment, Gross Regional Domestic Product (GRDP), level of health, employment opportunities, and education (Lilik Andrietya et al., 2020; Sakinah & Pudjianto, 2018). At the micro level, the influencing factors comprise education level, number of families, age of the head of household, house-building material, and asset ownership (Utariyanto et al., 2020; Purwa, 2019; Triasmoro et al., 2018; Santosa, 2018; Mathiassen & Roll-Hansen, 2007).

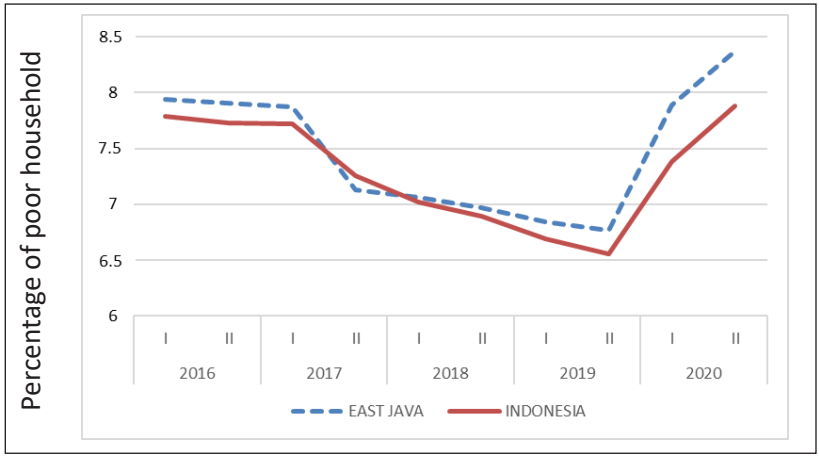
Poor households are more likely to be found in rural areas. In 2020, the percentage of poor Indonesian households in rural areas was 13.20 percent, almost two times larger than poor households in urban areas, i.e., 7.88 percent (BPS, 2021c). Although the percentage of poor households in rural areas was greater, the characteristics of poor households in urban areas still have an interesting side, namely getting closer to the "poverty crust" term (Krisliani & Setyari, 2020). This term is associated with poor households that are very difficult to eradicate. For example, people who are old do not have enough energy to access the economy and have physical limitations.

Poverty in urban areas is also associated with urbanisation (Zainal et al., 2012). The urbanisation of low-income groups contributes to an increase in urban poverty rates (Zainal et al., 2012). The poverty rate in Indonesia and East Java's urban areas decreased from 2016 until 2019 (BPS, 2021c), as presented in Figure 1. At the end of 2019, there was an outbreak of COVID-19, which caused the percentage of poor households in Indonesia and East Java to increase in 2020 (BPS, 2021c). This negative impact was also experienced by most countries in the world (United Nations, 2021). The United Nations (2021) stated that COVID-19 caused poverty alleviation to be off the target for 2030.

The physical characteristics of a home are one of the variables contributing to poverty (Utariyanto et al., 2020; Triasmoro et al., 2018; Purwa, 2019). Triasmoro et al. (2018) explained that the lower the quality of the type of floor and roof, the greater the tendency for households to be categorised as poor in Bali. Poverty in urban areas generally forms slums and squatter areas (Zainal et al., 2012). These areas have characteristics, such as readily damaged and fire-resistant house-building materials. In the multidimensional poverty approach, the physical type of the house is also a factor in impoverished households (UNDP & OPHI, 2020).

Figure 1

Percentage of Poor Households in East Java and Indonesia, 2016–2020



In contrast, the physical type of house possessed by East Java households indicated that it would be better in 2020 (BPS, 2021c). In urban areas, the proportions of families having non-fibre/thatched roofs, non-soil floors, and non-bamboo walls were 92.05 percent, 97.62 percent, and 99.67 percent, respectively (BPS, 2021c). In East Java's urban area, especially in Surabaya, it is feared that this factor will no longer be able to distinguish between households that are poor or not, and there will be households that are classified as poor but have a good physical type of house. This concern is due to the high percentage of urban households that have decent housing physical conditions.

Surabaya, one of Indonesia's big cities and the capital of East Java Province, the country's second-most populous province, was chosen for this study (BPS, 2021a). Surabaya has 154 urban villages categorised as urban areas (BPS, 2021b). The percentage of poor households in the city reached 5.02 percent (BPS, 2020).

In terms of binary classification, the poverty condition in Surabaya is classified as imbalanced data since the number of data on poor households is substantially smaller, or a minority class, compared to non-poor households or the majority class (Ali et al., 2015). Imbalanced data cause accuracy in the minority class to be small or the exclusion error to be high because the classification algorithm learns less from the minority class than the majority class (Jamaluddin & Mahat, 2021; Ali et al., 2015). Poor households are crucial for poverty alleviation programmes since they have a great urge to be noticed (Sri Kusumawati, 2019). In this case, exclusion error is more important than inclusion error (Gebremedhin et al., 2018). Moreover, Sen (1995) mentioned that these poor households are not only targets but also objects.

The algorithm-level approach is one alternative that can be used to overcome imbalanced data problems without changing the distribution of the original data. In binary classification, the methods included in this approach are Fuzzy Support Vector Machine (Lin & Wang, 2002), Entropy-based Fuzzy Support Vector Machine (Fan et al., 2017), Rare Event Weighted Logistic Regression (Maalouf & Siddiqi, 2014), and Entropy-based Fuzzy Weighted Logistic Regression (Harumeka et al., 2021). The logistic regression-based method offers the advantage of solving the optimisation problem without regard to constraints (Maalouf & Siddiqi, 2014). Rare Event Weighted Logistic Regression (RE-WLR) and Entropy-based Fuzzy Weighted Logistic

Regression (EFWLR) can be optimised using Truncated Newton with Linear Conjugate Gradient (Maalouf & Siddiqi, 2014; Harumeka et al., 2021). Meanwhile, EFWLR was superior to RE-WLR in terms of accuracy (Harumeka et al., 2021).

Based on the aforementioned issues, the purpose of this study is to compare the effect of the physical kind of housing on impoverished households using two methods: RE-WLR and EFWLR. This study began with an introduction to the subject and a review of the literature, followed by methodology, results and discussion, and a conclusion with potential future research.

LITERATURE REVIEW

Household Poverty Determinants

The Central Bureau of Statistics Indonesia determined the poor population based on the poverty line. The poverty line consists of the food and non-food poverty lines. The food poverty line is determined by dividing the population's food expenditure by 2,100 kcal per capita per day, while the non-food poverty line is determined using the 2004 Basic Needs Commodity Package Survey's minimum need value (BPS, 2016).

In general, the definition of poverty contains elements of powerlessness, shackled ideas, vulnerability, and fear (The World Bank, 2001). Poverty encompasses not simply poor income or consumption, but also limited access to education, health, nutrition, and other human development chances. According to the World Bank's research, there are five non-monetary factors that determine poverty: 1) education, 2) type of work, 3) gender, 4) access to basic services and infrastructure, and 5) geographical location.

In 2010, the United Nations Development Programme (UNDP) put forward the concept of multidimensional poverty. Poverty, according to the concept, is more than just a financial issue. It also implies that someone who is poor has difficulties accessing three fundamental aspects of life, namely health, education, and living standards (UNDP & OPHI, 2020). These aspects were further described in 10 non-monetary aspect indicators: 1) adequate nutrition, 2) infant mortality rate, 3) length of schooling, 4) education status, 5) cooking fuel, 6) sanitation, 7) drinking water, 8) electricity facilities, 9) housing, and 10) assets (UNDP & OPHI, 2020).

Physical Type of House as the Determinant of Poverty in Indonesia

In Indonesia, poverty alleviation is a government priority. One of the initiatives is to establish a social protection system. The social protection system, which includes the rice programme for the poor, the family hope programme (PKH), and community health insurance, directly targets poor households. Consequently, the programme's target families are required. The Population and Family Planning Agency (BKKBN) and the Central Bureau of Statistics (BPS) collected data to determine programme-target families in Indonesia. BKKBN conducted nuclear family-based data collection while BPS conducted household-based data collection, including the Population Socio-Economic Data Collection for 2005 (PSE05), Social Protection Programme Data Collection for 2008 (PPLS2008), Social Protection Programme Data Collection for 2008 (PPLS2011), and the Integrated Database Update for 2015 (PBDT2015).

The Indonesian Population and Family Planning Agency (BKKBN) has operated a data collection system for disadvantaged households since 1994. BKKBN divided family welfare status into five categories, namely Pre-Prosperous Families (Pre-KS), Prosperous Families 1 (KS1), Prosperous Families 2 (KS2), Prosperous Families 3 (KS3), and Prosperous Families 3 Plus (KS3 plus). There were a total of 23 indicators used to categorise this status. A family was classified as Pre-KS if it did not meet the first through fifth indicators, which are as follows:

1. Family members could worship according to their religion;
2. All family members could eat at least twice a day;
3. All family members had different clothes for home, school, work, and travel;
4. The widest type of floor was not land;
5. When a child becomes ill, they are sent to a medical facility.

The physical type of the house was represented by two of the 23 indicators, namely 1) the type of floor that is the largest non-soil; and 2) the inability to attain a floor area per capita of at least eight square metres.

BPS held socio-economic data collection for 2005 (PSE05). This programme aimed to obtain data on households targeted for direct

cash assistance (BLT) distribution. The BLT programme was created for low-income households to protect them from the rising fuel costs. PSE05 used indicators of household characteristics, as opposed to the monetary approach, which employs a consumption strategy. There were 14 indicators employed, which are as follows:

1. The floor area of the house;
2. Type of floor of the house;
3. Type of wall of the house;
4. Facilities for defecation;
5. Source of drinking water;
6. The lighting used;
7. Fuel used;
8. Frequency of eating in a day;
9. Habit of buying meat/chicken/milk;
10. Affordability to buy clothes;
11. Affordability to visit the health centre/polyclinic;
12. Employment of the head of household;
13. Education of the head of household;
14. Ownership of assets.

Three indicators specifically related to the physical type of the house emerged from all other indicators: 1) the floor area of the house, 2) type of house floor, and 3) type of house walls.

The 2008 Social Protection Programme Data Collection (PPLS2008) had the following objectives: 1) update the target household database for PSE05 results; 2) update information related to the socio-economic conditions of the target households; and 3) add data on the target household members, such as name, gender, school status, occupation of household members, and additional information related to housing. The indicators collected were complete compared to PSE05. The physical type of house was also included as an indicator that was collected to determine the target household, namely 1) floor area, 2) floor type, and 3) type of wall.

In 2011, the Indonesian government conducted another social protection programme data collection. In that year, the target households' coverage was expanded to include 40 percent of families with middle-class or lower incomes. It was expected that by expanding the coverage by 40 percent, social assistance programmes would not just target the destitute but also those who were at risk of falling into poverty. The information would then be utilised to create an Integrated

Database (BDT). The National Team for the Acceleration of Poverty Reduction oversaw BDT (TNP2K). When compared to the previous PPLS, the indicators collected were complete. The physical type of the house was still used as an indicator to determine the target household, such as: 1) floor area, 2) floor type, 3) wall type, and 4) type of roof.

The Integrated Database Update for 2015 (PBDT2015) was carried out to improve the BDT from the PPLS2011 results. The mechanism for carrying out this data collection was different from the previous activities. PBDT2015 included community involvement through Public Consultation Forum (FKP) activities. This FKP was a forum for discussion between regional officials, community leaders, facilitators, and assistant facilitators at the village level to identify targeted households for assistance. Complete field data collection of the target houses was done after the FKP activities. Even more comprehensive indicators were gathered than in PPLS2011. The physical characteristics of the home, including the 1) floor area, 2) floor type, 3) wall type, and 4) kind of roof, were still utilised to identify poor households.

Poverty-Related Studies

Several studies are concerned with poverty at the micro level. The summary of these studies is presented in Table 1. All of the variables used to determine whether a household is poor are connected to its members' characteristics, some of which include the physical type of house as utilised by Alatas et al. (2012), Triasmoro et al. (2018b), and Purwa (2019). A study by Utariyanto et al. (2020) also incorporated an additional variable related to the land area ownership of households. Based on the method used to determine the household poverty status, most of these studies employed the logistic regression analysis (Utariyanto et al., 2020; Triasmoro et al., 2018b; Purwa, 2019). As stated in the previous section, the household poverty data represent the imbalanced data where the poor household group is the minority group. Therefore, rather than using the standard logistic regression, Triasmoro et al. (2018b) and Purwa (2019) used different approaches to handle the imbalanced data. The former study used a modified algorithm approach by applying Rare Event Weighted Logistic Regression (RE-WLR) and Truncated Regularised-Prior Correction (TR-PC). The latter chose a re-sampling technique to obtain balanced data, including undersampling, oversampling, and a combination of sampling schemes. This study also employed other methods, such as

Random Forest for classification. Alatas et al. (2012) employed the Proxy Mean Test Regression (PMT) to examine poor targeting in 640 Indonesian villages.

Table 1

Poverty Studies

Research Author	Method	Predictor Variables Used
Utariyanto et al. (2020)	Logistic Regression	1. Household head age 2. Number of family 3. Land area ownership 4. Household head education
Purwa (2019)	Logistic Regression, Random Forest, and sampling	1. Number of family 2. Household head sex 3. Household head education 4. Household head job 5. Floor area per capita 6. Roof type 7. Wall type 8. Floor type 9. Toilet availability 10. Main source of drinking water 11. Main source of light 12. Type of cooking fuel 13. Assets ownership
Triasmoro et al. (2018b)	RE-WLR and Truncated Regularised-Prior Correction (TR-PC)	1. Household head age 2. Household head education 3. Household head job 4. Roof type 5. Floor type 6. Floor area per capita 7. Toilet ownership 8. Main source of drinking 9. Assets ownership 10.Residential area
Alatas et al. (2012)	Proxy Mean Test Regression (PMT)	There are 49 indicators similar to those used in PPLS2008. Some of these are physical house types, such as: 1. Floor area per capita 2. Roof type 3. Wall type 4. Floor type

METHODOLOGY

Data Source and Variables Description

This study aimed to analyse the effect of the physical type of house on households categorised as poor in Surabaya, Indonesia. The data used came from the Central Bureau of Statistics-March Indonesia's 2020 national socio-economic survey (Susenas). Susenas is conducted twice a year, in March and September. Since more samples were collected in March than in September, the Susenas survey in March can be used to estimate up to the regency level, while the survey in September can only be used for provincial-level estimation. In the Susenas March 2020 survey in Surabaya, 1,084 household samples participated, of which 43 were from poor households, and 1,041 were not. The variables used in the analysis process are listed in Table 2.

Table 2

Variable Description

Variable	Description	Code
Y	Household poverty category (response variable)	0 : Non-poor 1 : Poor
X_1	Gender of the head of household (predictor variable)	0 : Male 1 : Female
X_2	Number of household member (predictor variable)	
X_3	Head of household diploma (predictor variable)	0 : junior high school and above 1 : have no diploma/ primary school
X_4	Drinking water facility (predictor variable)	0 : branded bottled/ refillable water 1 : plumbing/ well/ water springs/ surface water/rain water/other
X_5	Gold asset ownership (predictor variable)	0 : Yes 1 : No
X_6	Floor area per capita (predictor variable)	…….

(continued)

Variable	Description	Code
X_7	Roof type (predictor variable)	0 : concrete/ roof tile 1 : zinc/ asbestos/ bamboo/ wood/ straw/ leaves/other
X_8	Wall type (predictor variable)	0 : plastering wall 1 : wood/ woven bamboo/ logs/ bamboo/ other
X_9	Floor type (predictor variable)	0 : marble/ granite/ parquet/ vinyl/ carpet /tile 1 : wooden planks/ cement/red bricks/ bamboo, soil/other

The operational definitions of some of the variables mentioned above are as follows:

1. Household poverty category (Y)
   A poor household is a household that has a monthly per capita expenditure below the poverty line. The poverty line of the city of Surabaya in March 2020 was Rp592,137.
2. Number of household members (X_2)
   Household members are all people who usually reside in a household (head of household, husband/wife, children, in-laws, grandchildren, parents/in-laws, other relatives, stay-at-home assistants, or other household members), both currently living at home or temporarily not at home.
3. Head of household diploma (X_3)
   A diploma is a sheet or proof of graduation given to someone who has completed all academic requirements at a certain level of formal education.
4. Drinking water facility (X_4)
   The drinking water facility is the source of water used by household members for drinking. If there are multiple sources of drinking water, the one with the greatest amount of water utilised by families is chosen.

The variables included in the physical type of house are floor area per capita, roof type, wall type, and floor type.

Rare Event Weighted Logistic Regression (RE-WLR)

Suppose there are n independent samples. From the sample set, there are data pairs $\vec{\boldsymbol{X}}_i$ and y_i, $i = 1,2,\dots,n$. $\vec{\boldsymbol{X}}_i$ is the vector of the i-th sample predictor variables while y_i is the response variable of the i-th sample, which is categorical, with a value of 0 or 1. Let p be the number of predictor variables. The vector and matrix notation are written as follows:

$$\vec{\mathrm{Y}} = \begin{bmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \vdots \\ \mathrm{y}_\mathrm{i} \\ \vdots \\ \mathrm{y}_\mathrm{n} \end{bmatrix}, \qquad \mathrm{X} = \begin{bmatrix} 1 & \mathrm{x}_{11} & \mathrm{x}_{12} & \dots & \mathrm{x}_{1\mathrm{p}} \\ 1 & \mathrm{x}_{21} & \mathrm{x}_{22} & \dots & \mathrm{x}_{2\mathrm{p}} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \mathrm{x}_{\mathrm{i}1} & \mathrm{x}_{\mathrm{i}2} & \dots & \mathrm{x}_{\mathrm{ip}} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \mathrm{x}_{\mathrm{n}1} & \mathrm{x}_{\mathrm{n}1} & \dots & \mathrm{x}_{\mathrm{np}} \end{bmatrix} = \begin{bmatrix} \vec{\mathrm{X}}_1^{\,\mathrm{T}} \\ \vec{\mathrm{X}}_2^{\,\mathrm{T}} \\ \vdots \\ \vec{\mathrm{X}}_\mathrm{i}^{\,\mathrm{T}} \\ \vdots \\ \vec{\mathrm{X}}_\mathrm{n}^{\,\mathrm{T}} \end{bmatrix} \tag{1}$$

where y_i is a mutually exclusive random variable with Bernoulli distribution. Equation 2 represents the results of the numerical iteration technique used to generate the logistic regression parameter from the natural logarithmic likelihood function (Hoshmer & Lemeshow, 2000).

$$\ln(\mathrm{L}(\vec{\beta})) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(\mathrm{y_i} \ln\left(\frac{e^{\vec{\mathrm{X}}_\mathrm{i}^{\,\mathrm{T}}\vec{\beta}}}{1+e^{\vec{\mathrm{X}}_\mathrm{i}^{\,\mathrm{T}}\vec{\beta}}} \right) + (1-\mathrm{y_i}) \ln\left(\frac{1}{1+e^{\vec{\mathrm{X}}_\mathrm{i}^{\,\mathrm{T}}\vec{\beta}}} \right) \right) \tag{2}$$

where $\vec{\boldsymbol{\beta}}$ is a vector of logistic regression parameters with size $p) + 1$.

Komarek and Moore (2005) added an element of ridge regularisation to the natural logarithmic likelihood function to overcome the overfitting problem. Overfitting is a problem that occurs if the model is good on the training data but bad on the testing data. The addition of ridge regularisation changes the form of Equation 2 to Equation 3:

$$Reg.ln\left(L(\vec{\boldsymbol{\beta}})\right) = \sum_{i=1}^{n} \left(ln\left(\frac{e^{y_i \vec{\boldsymbol{X}}_i^{\,\mathrm{T}}\vec{\boldsymbol{\beta}}}}{1+e^{\vec{\boldsymbol{X}}_i^{\,\mathrm{T}}\vec{\boldsymbol{\beta}}}} \right) \right) - \frac{\lambda}{2}\left\|\vec{\boldsymbol{\beta}}\right\|^2 \tag{3}$$

where $\frac{\lambda}{2}\left\|\vec{\boldsymbol{\beta}}\right\|^2$ is the ridge regularisation, λ is the regularisation parameter, and $\left\|\vec{\beta}\right\|^2 = \sum_{\mathrm{j}=0}^{\mathrm{p}} \beta_\mathrm{j}{}^2$.

One of the numerical iterations that can be used to optimise Equation 3 is the Newton Raphson iteration. Nevertheless, the iteration requires a long computational process if the number of samples is large. Therefore, Komarek and Moore (2005) performed optimisation using Truncated Newton with a linear conjugate gradient to cut the number of iterations of Newton Raphson.

The method developed by Komarek and Moore (2005), Truncated Regularised-Iteratively Reweighted Least Square (TR-IRLS), is built on the assumption of balanced data. King and Zheng (2001) provided a weighting on logistic regression to overcome the case of classification on imbalanced data. The weighting system was adopted by Maalouf and Siddiqi (2014) to develop the RE-WLR method that combined TR-IRLS, weighting, and bias correction. Bias correction is added to overcome the bias that occurs due to the addition of ridge regularisation. When a weight is added, the form of Equation 3 becomes Equation 4:

$$Reg.ln\left(L_W(\vec{\boldsymbol{\beta}})\right) = \sum_{i=1}^{n}\left(w_i\, ln\left(\frac{e^{y_i \vec{X}_i^{\mathrm{T}}\vec{\boldsymbol{\beta}}}}{1+e^{\vec{X}_i^{\mathrm{T}}\vec{\boldsymbol{\beta}}}}\right)\right) - \frac{\lambda}{2}\left\|\vec{\boldsymbol{\beta}}\right\|^2 \quad (4)$$

where $w_i = \left(\frac{\tau}{\bar{y}}\right) y_i + \left(\frac{1-\tau}{1-\bar{y}}\right)(1-y_i)$, τ is the proportion of successful events, usually coded 1, on the population data, and $\bar{y}$ is the proportion of successful events on the sample data.

The optimisation of Equation 4 using Truncated Newton requires Gradient Vector and Hessian Matrix (Maalouf & Siddiqi, 2014). Equation 5 represents the Gradient Vector obtained from the first derivative of Equation 4 with respect to $\vec{\boldsymbol{\beta}}$.

$$\boldsymbol{G}(\vec{\boldsymbol{\beta}}) = \frac{d\, ln\left(L(\vec{\boldsymbol{\beta}})\right)}{d\vec{\boldsymbol{\beta}}} = \boldsymbol{X^T W}(\vec{\boldsymbol{Y}} - \vec{\boldsymbol{\pi}}) - \lambda\vec{\boldsymbol{\beta}} \quad (5)$$

where $\boldsymbol{W} = diag(w_1, w_2, \ldots, w_n)$, $\vec{\boldsymbol{\pi}}$ is a vector whose component is the probability of an event, and π_i is described in Equation 6:

$$\pi_i = \frac{e^{\vec{X}_i^{\mathrm{T}}\vec{\beta}}}{1 + e^{\vec{X}_i^{\mathrm{T}}\vec{\beta}}} \tag{6}$$

While the Hessian Matrix is the second derivative of Equation 3 with respect to $\vec{\boldsymbol{\beta}}$. The Hessian Matrix is depicted in Equation 7:

$$\boldsymbol{H}(\vec{\boldsymbol{\beta}}) = \frac{d\ ln\left(L(\vec{\boldsymbol{\beta}})\right)}{d^2\vec{\boldsymbol{\beta}}} = -\boldsymbol{X^T D X} - \lambda \boldsymbol{I} \tag{7}$$

where $v_i = \pi_i(1 - \pi_i)$ and $\boldsymbol{D} = diag\{v_1 w_1, v_2 w_2, \ldots, v_n w_n\}$. The Gradient Vector and Hessian matrix, respectively in Equations 5 and 7, are used to obtain $\vec{\hat{\boldsymbol{\beta}}}$ using the Truncated Newton method (Komarek & Moore, 2005; Maalouf & Siddiqi, 2014). This study applied the Conjugate Gradient algorithm as the inner iteration in Truncated Newton, as used by Rahayu et al. (2012) and Harumeka et al. (2021).

Logistic regression will produce biased parameters if applied to imbalanced data (King & Zeng, 2001). The addition of regularisation will also add to the bias (Maalouf & Siddiqi, 2014). For this reason, Maalouf and Siddiqi (2014) added an element of bias correction to the parameter estimation. The bias correction is shown in Equation 8:

$$\boldsymbol{B}\left(\vec{\hat{\boldsymbol{\beta}}}\right) = (\boldsymbol{X^T D X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X^T D} \vec{\xi} \tag{8}$$

where $\vec{\xi}$ is a vector whose *i*-th element has a value based on Equation 9:

$$\xi_i = 0{,}5 Q_{ii}\left((1 + w_1)\hat{\pi}_\iota - w_1\right), i = 1,2, \ldots, n \tag{9}$$

where w_1 is the weighting value for the minority class, $\hat{\pi}_\iota$ is an estimator of π_i, and Q_{ii} is the main diagonal element of the $\boldsymbol{Q}$ covariance matrix as in Equation 10:

$$\boldsymbol{Q} = \boldsymbol{X}(\boldsymbol{X^T D X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X^T} \tag{10}$$

The bias equation in Equation 8 is also solved by using Truncated Newton with Conjugate Gradient based on Rahayu et al. (2012). Therefore, the bias-corrected parameter is obtained by Equation 11:

$$\vec{\tilde{\boldsymbol{\beta}}} = \vec{\hat{\boldsymbol{\beta}}} - \boldsymbol{B}\left(\vec{\hat{\boldsymbol{\beta}}}\right) \tag{11}$$

Entropy-based Fuzzy Weighted Logistic Regression (EFWLR)

Entropy-based Fuzzy Weighted Logistic Regression (EFWLR) adopted the use of Entropy-based Fuzzy membership in Entropy-based Fuzzy Support Vector Machine (EFSVM) as a substitute for weighing in RE-WLR (Harumeka et al., 2021). The regularised weighted natural logarithmic likelihood function of EFWLR has the same form as Equation 4. Nonetheless, the weights are assessed based on entropy-based fuzzy membership by Fan et al. (2017). Therefore, the weight becomes Equation 15:

$$w_i = \begin{cases} 1\,, if\ y_i = 1 \\ FM_l\,, if\ y_i = 0\ and\ \vec{\boldsymbol{X}}_i \epsilon\ l-th\ subset \end{cases} \quad (15)$$

$$i = 1,2,\dots,n.\ and\ l = 1,2,\dots,m$$

where FM_l is the fuzzy membership value of the l-th subset. To get w_i, it is necessary to tune the hyperparameters of $m, \varphi,$ and k (Fan et al., 2017). In this case, m is the number of subsets, φ is the fuzzy membership parameter, and k is the number of nearest neighbours.

The Truncated Newton method with linear Conjugate Gradient was used to obtain $\vec{\boldsymbol{\beta}}$. The Conjugate Gradient algorithm was utilised as the inner iteration in Truncated Newton as also done by Rahayu et al. (2012) and Harumeka et al. (2021). Equations 5 and 7, respectively, have the same form for the Gradient Vector and Hessian Matrix.

Bias correction was also adopted in EFWLR. The bias correction vector had the same form with Equation 8. Based on Equation 15, samples in the minority class would be given a value of 1. So, Equation 8 became Equation 16:

$$\xi_i = 0{,}5Q_{ii}(2\hat{\pi}_i - 1) \quad (16)$$

where Q_{ii} is the main diagonal element of the $\boldsymbol{Q}$ covariance matrix as in Equation 10. The bias-corrected parameter was obtained by Equation 11 and partial beta testing was done using the Wald Test, which had statistics as in Equation 13.

Procedures for Analysing the Effect of Physical Type of House

Figure 2 provides a summary of the procedures involved in utilising RE-WLR and EFWLR to determine how the physical type of a house affects household poverty.

1. The dataset was divided into training and testing data using Stratified 5-fold Cross-Validation. This approach divided the dataset into five groups, called folds, with the same number of samples for each fold and the same proportion of minority classes with the entire dataset (Gareth et al., 2013). If the first fold was treated as the testing data, then the remaining fold was treated as the training data. This procedure was repeated five times.
2. $\bar{y}$ was calculated for each training data and τ was determined based on the estimated percentage of Surabaya's poor population, which was 5.02 percent (BPS, 2020). Furthermore, this study also determined the entropy-based fuzzy membership value of each sample in the training data.
3. $\bar{y}$ and τ were used to obtain the value of w_i on RE-WLR, while entropy-based fuzzy membership was applied to determine the value of w_i on EFWLR.
4. RE-WLR had one hyperparameter, which was λ. EFWLR had four hyperparameters, namely $m, \varphi, k,$ and λ. The study established λ based on the set of {0,00001; 0,0001; 0,001; 0,01; 0,1; 1;10; 100; 1000; 100000}, k based on the set of $\{3,5,7,9,11\}$, while m and φ were set to 10 and 0.05, respectively. The optimal hyperparameter was obtained from the model that produced the largest Stratified 5-fold Cross-Validation's Area Under the Curve (AUC) from all predefined hyperparameter combinations. AUC represents classification accuracy. The Stratified 5-fold Cross-Validation's AUC was computed by averaging the AUC value of each fold (Gareth et al., 2013) as in Equation 17:

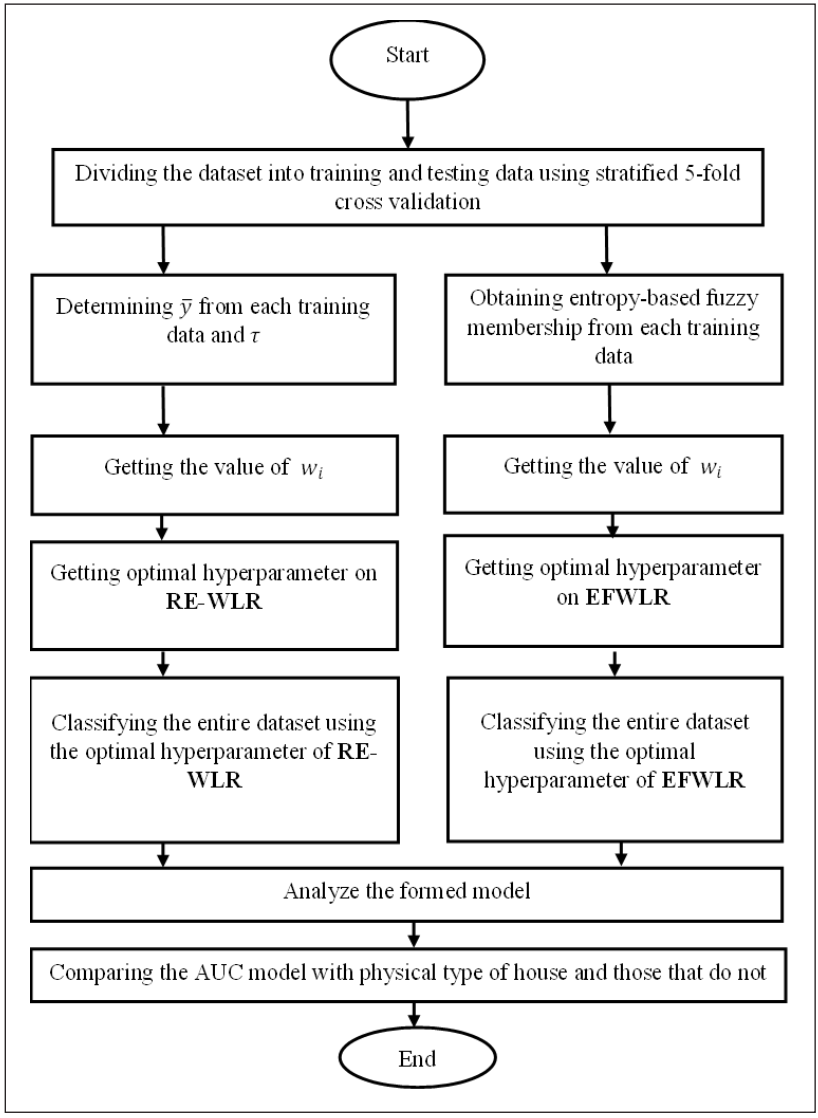
$$AUC_{cv} = \frac{1}{5}\sum_{k=1}^{5} AUC_k \tag{17}$$

5. The study used the optimal hyperparameter of RE-WLR and EFWLR from Procedure 4 to make the model from all datasets.
6. By interpreting significant variables, particularly the physical type of housing variable, the created models were examined.
7. The optimal AUC of the dataset that contained the physical type of house in Procedure 4 was obtained. The next procedure was to determine the optimal AUC of the dataset that did not contain the physical type of house using the following steps: 1) removing variables that were included in the physical type

of house, and 2) using Procedures 1 until 4 to obtain optimal AUC. Finally, the study compared the AUC obtained by the model that had a physical type of house and those that did not have it.

Figure 2

Analysing the Effect of Physical Type of House Procedure



RESULTS AND DISCUSSION

The Optimal Hyperparameter of RE-WLR

Table 3 describes the Stratified 5-fold Cross-Validation's AUC on each hyperparameter, λ. The optimal hyperparameter was indicated by numbers in bold. The highest AUC value was obtained when λ equals 10^{-3}. The table also shows that the AUC value continues to decrease when λ is more than 10^{-3}.

Table 3

Stratified 5-fold Cross-Validation's AUC on Each RE-WLR Hyperparameter

λ	AUC
10^{-5}	0.9134
10^{-4}	0.9135
10^{-3}	**0.9136**
10^{-2}	0.9135
10^{-1}	0.9127
1	0.9010
10^{1}	0.8659
10^{2}	0.8263
10^{3}	0.8070
10^{4}	0.8055
10^{5}	0.8100

The Optimal Hyperparameter of EFWLR

EFWLR had more hyperparameters to be tuned than RE-WLR. Table 4 shows the Stratified 5-fold Cross-Validation's AUC on each hyperparameter, λ and k. The optimal hyperparameter was indicated by numbers in bold. The highest AUC value was obtained when λ equalled 10^{-3} and k equalled 3. λ trended where every k had something in common. It showed a pattern up to a specific λ, and then it would decline till the experiment ended. For instance, when k was equal to 5, AUC increased until λ equalled 10^{-2}, after which it fell until λ was equal to 10^{4}. When λ was equal to 10^{5} in all k, AUC marginally climbed once more.

Table 4

Stratified 5-fold Cross-Validation's AUC on each EFWLR Hyperparameter

	k				
	3	**5**	**7**	**9**	**11**
10^{-5}	0.9147	0.9135	0.9139	0.9103	0.9117
10^{-4}	0.9147	0.9135	0.9139	0.9104	0.9118
10^{-3}	**0.9148**	0.9135	0.9139	0.9109	0.9121
10^{-2}	0.9144	0.9139	0.9130	0.9131	0.9146
10^{-1}	0.9119	0.9111	0.9107	0.9104	0.9116
1	0.8991	0.8982	0.8979	0.8982	0.8990
10^{1}	0.8612	0.8617	0.8619	0.8626	0.8624
10^{2}	0.8224	0.8236	0.8237	0.8236	0.8237
10^{3}	0.8046	0.8054	0.8061	0.8067	0.8067
10^{4}	0.8056	0.8057	0.8055	0.8055	0.8059
10^{5}	0.8103	0.8101	0.8101	0.8101	0.8102

Formed Model

The optimal lambda on RE-WLR and EFWLR had the same value, which was 10^{-3}. The estimation results with optimal hyperparameters are presented in Tables 5 and 6. According to RE-WLR, four factors—the number of household members, the head of family's education level, the availability of drinking water, and the floor area per capita—significantly (α=0.05) affected the likelihood that a household would be classified as poor. The significant variables are marked in bold in Table 5. According to the p-value, the roof type was significant, but the data, both descriptive and modelling, showed a different direction from the previous study. The lower the quality of roof type, the greater the tendency for households to be categorised as poor (Triasmoro et al., 2018b; Zainal et al., 2012). Therefore, this study considered that this variable was insignificant. The only variable from the physical type of a house factor significantly affecting the likelihood that a household would be classified as poor was floor area per capita. In contrast, the roof type, wall type, and floor type did not significantly affect it.

Table 5

RE-WLR Full Model

Variables	β	p-value	Odd Ratio
Intercept	**-6.6762**	**0.0003**	**-**
X1	0.5373	0.1686	1.7114
X2	**0.5078**	**0.0000**	**1.6616**
X3	**1.0936**	**0.0008**	**2.9850**
X4	**1.1645**	**0.0018**	**3.2043**
X5	3.1798	0.0596	24.0424
X6	**-0.1936**	**0.0000**	**1.2136**
X7	-1.0058	0.0061	2.7340
X8	0.3250	0.6353	1.3840
X9	0.7374	0.1015	2.0904

Gold asset ownership became significant in EFWLR. Therefore, five variables significantly affected the probability of a household being categorised as poor on EFWLR, namely the number of household members, head of household's education level, drinking water facility, gold asset ownership, and floor area per capita. The significant variables are marked in bold and presented in Table 6. For the same reason as RE-WLR, this study made roof type an insignificant variable. Compared to households without gold holdings, those that did have a chance to be classified as non-poor roughly 31 times more frequently.

Table 6

EFWLR Full Model

Variables	β	p-value	Odd Ratio
Intercept	**-7.1271**	**0.0001**	**-**
X1	0.5396	0.2119	1.7154
X2	**0.5250**	**0.0000**	**1.6904**
X3	**1.2056**	**0.0008**	**3.3389**
X4	**1.1502**	**0.0049**	**3.1589**

(continued)

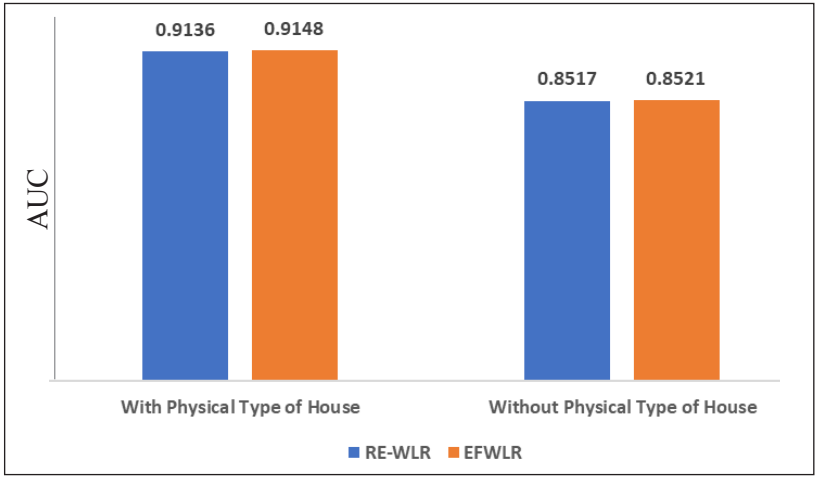
Variables	β	p-value	Odd Ratio
X5	**3.4380**	**0.0406**	**31.1258**
X6	**-0.2006**	**0.0000**	**1.2221**
X7	-1.0156	0.0122	2.7610
X8	0.3875	0.5990	1.4733
X9	0.7070	0.1501	2.0280

Comparison of AUC with and without the Physical Type of House in the Model

The effect of the physical type of house on AUC is shown in Figure 3. When the variables of the physical type of house were removed from the model, the AUC value of the RE-WLR model was reduced from 0.9136 to 0.8517 (-6.78%). Under the same treatment, the EFWLR model's AUC also decreased from 0.9148 to 0.8521 (-6.85%).

Figure 3

Comparison of the AUC Model with and Without the Physical Type of House



CONCLUSION

Due to the high percentage of urban households that have decent housing physical conditions in urban areas, especially in Surabaya, it

is feared that this factor will no longer be able to distinguish between households that are poor or not, and there will be households that are classified as poor but have a good physical type of house. This study aimed to examine whether it is still possible to categorise poor households in urban areas using the physical type of houses. Two methods, namely Rare Event Weighted Logistic Regression (RE-WLR) and Entropy-based Fuzzy Weighted Logistic Regression (EFWLR), were used to answer this objective.

First, the accuracy of EFWLR was slightly higher than the accuracy of RE-WLR. From the significance test, the statistically significant variables obtained by RE-WLR were the number of household members, head of household's education level, drinking water facility, and floor area per capita, while the significant variables obtained from EFWLR were the number of household members, head of household's education level, drinking water facility, gold asset ownership, and floor area per capita. Floor area per capita was the only significant variable for the physical type of house. The other three variables, namely roof type, wall type, and floor type, did not significantly affect the probability of a household being categorised as poor. Second, the elimination of the physical type of house from the model would reduce the AUC value by 6.78 percent on RE-WLR, and 6.85 percent on EFWLR.

Based on variable significance, in the first outcome, only floor area per capita could still be used to categorise poor households. Nevertheless, depending on accuracy (the second outcome), the use of the physical type of house can increase the accuracy of the model. Therefore, it is still possible to categorise poor households in urban areas, especially Surabaya, using the physical type of house variables.

ACKNOWLEDGMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Alatas, V., Banerjee, A., Hanna, R., & Olken, B. A. (2012). Targeting the poor: Evidence from a field experiment in Indonesia.

American Economic Review, *102*(4), 1206–1240. https://doi.org/10.1257/aer.102.4.1206.Targeting

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A review. *International Journal of Advances in Soft Computing and Its Applications*, *7*(3), 176–204.

BPS. (2016). *Penghitungan dan analisis kemiskinan makro Indonesia 2016*. Badan Pusat Statistik.

BPS. (2020). *Data dan Informasi Kemiskinan Kabupaten/Kota 2020*. Badan Pusat Statistik.

BPS. (2021a). *Jumlah penduduk hasil proyeksi menurut provinsi dan jenis kelamin (ribu jiwa), 2018–2020*. Badan Pusat Statistik. https://www.bps.go.id/indicator/12/1886/1/jumlah-penduduk-hasil-proyeksi-menurut-provinsi-dan-jenis-kelamin.html

BPS. (2021b). *Master file desa provinsi Jawa Timur 2020*. Badan Pusat Statistik.

BPS. (2021c). *Persentase penduduk miskin (P0) menurut provinsi dan daerah 2020-2021*. Badan Pusat Statistik. https://www.bps.go.id/indicator/23/192/1/persentase-penduduk-miskin-p0-menurut-provinsi-dan-daerah.html

De Donder, P., & Hindriks, J. (1998). The political economy of targeting. *Public Choice*, *95*(1–2), 177–200. https://doi.org/10.1023/A:1005023531490

Fan, Q., Wang, Z., Li, D., Gao, D., & Zha, H. (2017). Entropy-based fuzzy support vector machine for imbalanced datasets. *Knowledge-Based Systems*, *115*, 87–99. https://doi.org/10.1016/j.knosys.2016.09.032

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer Science+Business Media. https://doi.org/10.2200/S00899ED1V01Y201902MAS024

Gebremedhin, S., Bekele, T., & Retta, N. (2018). Inclusion and exclusion errors in the targeted supplementary feeding programme of Ethiopia. *Maternal and Child Nutrition*, *14*(4), 1–7. https://doi.org/10.1111/mcn.12627

Harumeka, A., Purnami, S. W., & Rahayu, S. P. (2021, November). Entropy-based fuzzy weighted logistic regression for classifying imbalanced data. In *International Conference on Soft Computing in Data Science* (pp. 312–327). Springer Singapore.

Hoshmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression, second edition*. John Wiley & Sons , Inc.

Jamaluddin, A. H., & Mahat, N. I. (2021). Validation assessments on resampling method in imbalanced binary classification for linear discriminant analysis. *Journal of Information and Communication Technology*, *20*(1), 83–102. https://doi.org/https://doi.org/10.32890/jict.20.1.2021.6358

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Journal of Statistical Software*, *8*, 137–163. https://doi.org/10.18637/jss.v008.i02

Komarek, P., & Moore, A. W. (2005, November). Making logistic regression a core data mining tool with TR-IRLS. In *Proceedings - IEEE International Conference on Data Mining, ICDM* (pp. 4–7). IEEE. https://doi.org/10.1109/ICDM.2005.90

Krisliani, P., & Setyari, N. P. W. (2020). Determinan Tingkat kemiskinan di Kabupaten/kota provinsi Bali. *E-Jurnal EP Unud*, *10*(6), 2545–2573.

Lilik Andrietya, A., Pujiati, A., & Setyadharma, A. (2020). Determinants of poverty in Central Java Province 2013–2018. *Journal of Economic Education*, *9*(1), 81–88. https://doi.org/10.15294/jeec.v9i1.38671

Lin, C. F., & Wang, S. De. (2002). Fuzzy support vector machines. *IEEE Transactions on Neural Networks*, *13*(2), 464–471. https://doi.org/10.1109/72.991432

Maalouf, M., & Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, *59*, 142–148. https://doi.org/10.1016/j.knosys.2014.01.012

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, *128*(584 PART B), 2145–2166. https://doi.org/10.1256/003590002320603584

Mathiassen, A., & Roll-Hansen, D. (2007). *Predicting poverty for Mozambique 2000 to 2005: How robust are the models*. Statistics Norway/Division for Development Cooperation.

Purwa, T. (2019). Perbandingan metode regresi logistik dan random forest untuk klasifikasi data imbalanced (studi kasus: klasifikasi rumah tangga miskin di Kabupaten Karangasem, Bali Tahun 2017). *Jurnal Matematika, Statistika Dan Komputasi*, *16*(1), 58. https://doi.org/10.20956/jmsk.v16i1.6494

Rahayu, S. P., Zain, J. M., Embong, A., Juwari, & Purnami, S. W. (2012). Logistic regression methods with truncated newton method. *Procedia Engineering*, *50*, 827–836. https://doi.org/10.1016/j.proeng.2012.10.091

Sakinah, N., & Pudjianto, H. (2018). Determinants of poverty in East Java Metropolitan area in 2010–2016. *Eko-Regional Jurnal Pengembangan Ekonomi Wilayah*, *13*(2), 32–40. https://doi.org/10.20884/1.erjpe.2018.13.2.1171

Santosa, B. (2018). *Kajian simulasi over sampling k-tetangga terdekat pada regresi logistik terboboti dan penerapannya untuk klasifikasi rumah tangga miskin di provinsi daerah istimewa Yogyakarta* [Doctoral dissertation, IPB Bogor Agricultural University].

Sri Kusumawati, A. (2019). The effectiveness of targeting social transfer programs in Indonesia. *Jurnal Perencanaan Pembangunan: The Indonesian Journal of Development Planning*, *3*(3), 282–297. https://doi.org/10.36574/jpp.v3i3.90

The World Bank. (2001). *World Development Report 2000/2001*. Oxford University Press, Inc.

Triasmoro, S. P., Ratnasari, V., & Rumiati, A. T. (2018a, March). Comparison performance between rare event weighted logistic regression and truncated regularised prior correction on modelling imbalanced welfare classification in Bali. In *2018 International Conference on Information and Communications Technology, ICOIACT 2018* (pp. 108–113). IEEE. https://doi.org/10.1109/ICOIACT.2018.8350751

Triasmoro, S. P., Ratnasari, V., & Rumiati, A. T. (2018b). *Perbandingan metode rare event weighted logistic regression dan truncated regularized prior correction* [Doctoral dissertation, Institut Teknologi Sepuluh Nopember].

UNDP, & OPHI. (2020). *Charting pathways out of multidimensional poverty : Achieving the SDGs*. https://hdr.undp.org/system/files/documents/2020mpireportenpdf.pdf

United Nations. (2021). *The sustainable development goals report 2021*. United Nations. https://unstats.un.org/sdgs/report/2021/The-Sustainable-Development-Goals-Report-2021.pdf

Utariyanto, G. R., Sukiyono, K., & Widiono, S. (2020). Determinant factors of the household poverty probability: Study on household around the Taman Nasional Kerinci Sebelat (TNKS) Lebong Regency. *AGRITROPICA : Journal of Agricultural Sciences*, *3*(1), 29–37. https://doi.org/10.31186/j.agritropica.3.1.29-37

Zainal, N. R., Kaur, G., Ahmad, N. 'Aisah, & Khalili, J. M. (2012). Housing conditions and quality of life of the urban poor in Malaysia. *Procedia - Social and Behavioral Sciences*, *50*(July 2012), 827–838. https://doi.org/10.1016/j.sbspro.2012.08.085