


RESEARCH ARTICLE | MARCH 17 2025

Classification infant mortality rates (IMR) using logistic regression ensemble (LORENS)

Muhamad Sabit Munawar; Ria D. L. N. Karisma ; Dwi I. Kharisma

AIP Conf. Proc. 3302, 050001 (2025)

<https://doi.org/10.1063/5.0262606>



Articles You May Be Interested In

Application of nonparametric truncated spline regression on infant mortality rate in Kalimantan

AIP Conf. Proc. (May 2023)

Modeling the maternal mortality rate in Indonesia using geographically weighted Poisson Regression approach

AIP Conf. Proc. (December 2023)

Modeling bivariate Poisson regression for maternal and infant mortality in Central Java

AIP Conf. Proc. (February 2021)

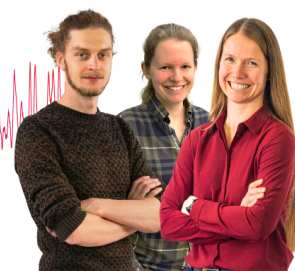
Webinar From Noise to Knowledge

May 13th – Register now



Zurich
Instruments

Universität
Konstanz



Classification Infant Mortality Rates (IMR) Using Logistic Regression Ensemble (LORENS)

Muhamad Sabit Munawar,^{1, a)} Ria D. L. N. Karisma,^{1, b)} and Dwi I. Kharisma^{2, c)}

¹⁾*Department of Mathematics, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang City 65144, Indonesia*

²⁾*Badan Pusat Statistik, Kupang City, Indonesia*

^{a)}*Electronic mail: muhamad.sabit.munawar@gmail.com*

^{b)}*Corresponding author: riadhea@uin-malang.ac.id*

^{c)}*Electronic mail: dwi.iwan@bps.go.id*

Abstract. Infant Mortality Rate (IMR) is the number of cases of infant mortality less than one-year-old divided by 1000 births. IMR is one of the most important indicators in determining public health problems. The number of IMR cases is influenced by several factors, such as healthcare services, the mother's age at childbirth, and many more. Logistic Regression Ensemble (LORENS) is a classification method using ensemble techniques developed based on the Logistic Regression (LR) method. The advantage of LORENS is freedom from data dimension assumptions and the determination of classification classes using an optimal threshold. The method used to evaluate the goodness of the LORENS model is cross-validation. The purpose of this study is to obtain LR models and accuracy in the classification of infant mortality rates in Indonesia using LORENS. The best accuracy is achieved based on a training data to testing data ratio of 85%: 15% with two partitions and ten ensembles. The classification results of IMR using LORENS resulted in 20 LR models, with the breastfeeding variable being the most influential variable on infant mortality. The classification accuracy results are shown in the calculation of LORENS, which is 79.47%. The accuracy level of the LORENS model against IMR has a good value. Since, the cross-validation using LORENS has accuracy of the LORENS method using cross-validation, it showed an accuracy value of 78.87%.

Keywords: classification, infant mortality rate, logistic regression, logistic regression ensemble

INTRODUCTION

The purpose of government programs is to build welfare for the people. Prosperity in a region can be achieved by holding development. Improvements in community sectors such as the education system, economy, technology, infrastructure, and health are efforts in carrying out development. One of the most important components in determining the level of well-being of people is their level of health. Public health is a benchmark in development because the better the public health level, the productivity will definitely increase as well.

According to [1], in the study of the Government Work Plan (RKP) of the State Budget (APBN) DPR RI, the infant mortality rate (IMR) is the number of cases of infant deaths with an age of less than one year per 1000 births. IMR is one of the most important indicators in determining the level of public health problems. The high level of IMR in a region illustrates the lack of health services in the region. The size of IMR is influenced by many factors, including health services, parents' education level, mother's age during childbirth, mother's occupation, and many more. IMR of the world in 2021 was 28 deaths per 1000 live births, lower than 2020 at 29 deaths per 1000 live births. Meanwhile, the IMR in Indonesia in 2021 is 18.9, lower than the IMR level in previous years [2]. Indonesia still has to try to achieve the Sustainable Development Goals (SDGs) target in 2030, namely 12 cases of IMR.

Classification is a multivariate method that relates to partitioning training samples and allocating new observations into certain classes or categories. The goal is to obtain an optimal discriminant function that can separate observations belonging to different classes or obtain rules for categorizing each new observation. The methods that can be used in classification are Logistic Regression (LR), discriminant analysis, artificial neural networks (ANN), support vector machines (SVM), and others. Among several classification methods, Logistic Regression (LR) is the most popular because it can be represented clearly and concisely [3]. According to [4], LR is a method of describing the relationship between a response variable (dependent) that has two or more categories and one or more predictor variables (independent) on a category or interval scale. However logistic regression models tend to over-fit the learning sample when the number p of features, or input variables, largely exceeds the number n of samples. This is referred to as the small n large p setting, commonly found in biomedical problems such as gene selection from microarray data [5].

The Logistic Regression Ensemble (LORENS) is a classification method that uses the ensemble technique developed by [6]. This classification method is a development of basic statistics Logistic Regression (LR). The LORENS

method developed the Classifications by Ensembles from Random Partition (CERP) algorithm by dividing it into several subspaces based on the partitioning process of its variants. They are then combined again in Logistic Regression (LR) models on each partition into one function [7].

The advantage of LORENS when compared to LR is that it is free from data dimension assumptions because the predictors in LORENS are partitioned randomly. In addition, LORENS uses probability thresholds in each of the optimal classes. High-dimensional data using LR will produce problems because it uses a probability threshold of 0.5. It is a problem because it is considered unfair if the probability of each class is stated to be 0.5.

The purpose of this study is to obtain LR models and determine the accuracy results of LORENS on IMR in Indonesia. Previously, the application of the LORENS model had been carried out by [3] entitled Logistic Regression Ensemble (LORENS) Applied to Drug Discovery in 2020 which obtained an accuracy of 69.41%. In addition [8] have also done in the Enzyme classification on DUD-E database Using Logistic Regression Ensemble (LORENS) in 2018 where obtained 88.95% for manoanime oxidase B (aofb) enzyme, 92.1% for carbonic anhydrase II (cah2) enzyme, and 100% for heat stock protein HSP 90-alpha (hs90a) enzyme.

LITERATURE REVIEW

Logistic Regression

Logistic regression (LR) is a statistical analysis method to analyze the best model based on the results of the relationship that occurs between predictor variables (independent) and response variable variables (dependent). Response variable variables have two or more categories, while predictor variables are category scales or intervals with one or more variables [9]. LR is a nonlinear regression used to determine the relationship between predictor and response variables with nonlinear properties. Y-spread abnormalities and nonconstant response diversity cannot be explained using ordinary linear regression models [10].

The response variable (y) in binary logistic regression has two categories: "failed" and "succeeded". The response variable category "failed" is denoted with $y = 0$ and the category "success" is denoted with $y = 1$. That is, the response variable y can be assumed to follow the Bernoulli distribution in every single observation. The probability function in each observation is

$$f(y_i) = (\pi_j(x_i))^{y_i} (1 - \pi_j(x_i))^{1-y_i}; y_i = 0, 1, \quad (1)$$

with $\pi_j(x_i)$ is probability of j -th event at the time of i -th data. If value of $y_i = 1$, so $f(y_i) = \pi_j(x_i)$ and if value of $y_i = 0$, so $f(y_i) = 1 - \pi_j(x_i)$.

Based on the probability function in (1), the Logistic Regression (LR) function is obtained as follows:

$$f(z) = \frac{e^z}{1 + e^z},$$

with z is $\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ and p is many predictor variables.

For any given z value, the value of $f(z)$ lies between 0 and 1. Because z is the value that lies between $-\infty$ and ∞ . This shows that the LR model actually illustrates the risk or probability of an object. Where known LR models are as follows.

$$\pi_j(x_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}. \quad (2)$$

Note that the relationship between predictor variables and response variables in LR is a non-linear function. Furthermore form (2) can be transformed into linear form or in logistic regression called logit transformation from the following $\pi_j(x_i)$.

$$\text{Logit} [\pi_j(x_i)] = \ln \left(\frac{\pi_j(x_i)}{1 - \pi_j(x_i)} \right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}. \quad (3)$$

Meanwhile, parameter estimation in LR can use the Maximum Likelihood Estimation (MLE) method. This method provides an estimated value of β by maximizing the likelihood function [9]. The MLE method requires that the

data be based on a certain distribution. In LR, each observation is based on the Bernoulli distribution; the following equation can determine the likelihood function.

$$\begin{aligned}
L(\beta, \mathbf{X}) &= \prod_{j,i=1}^n [\pi_j(x_i)]^{y_i} [1 - \pi_j(x_i)]^{1-y_i} \\
&= \left\{ \prod_{j,i=1}^n \pi_j(x_i)^{y_i} \right\} \left\{ \prod_{j,i=1}^n (1 - \pi_j(x_i))^{1-y_i} \right\} \\
&= \left\{ \prod_{j,i=1}^n (1 - \pi_j(x_i)) \right\} \left\{ \prod_{j,i=1}^n \pi_j(x_i)^{y_i} \right\} \left\{ \prod_{j,i=1}^n (1 - \pi_j(x_i))^{-y_i} \right\} \\
&= \left\{ \prod_{j,i=1}^n (1 - \pi_j(x_i)) \right\} \left\{ \prod_{j,i=1}^n e^{\left(\ln \left(\frac{\pi_j(x_i)}{1 - \pi_j(x_i)} \right)^{y_i} \right)} \right\} \\
&= \left\{ \prod_{j,i=1}^n (1 - \pi_j(x_i)) \right\} \left\{ e^{\left(\sum_{i=1}^n y_i \ln \left(\frac{\pi_j(x_i)}{1 - \pi_j(x_i)} \right) \right)} \right\} \\
&= \left\{ \prod_{j,i=1}^n \left(\frac{1}{1 + e^{\sum_{p=0}^m \beta_p x_{ip}}} \right) \right\} \left\{ e^{\left(\sum_{p=0}^m \beta_p \left(\sum_{i=1}^n y_i x_{ip} \right) \right)} \right\}
\end{aligned}$$

MLE is obtained by maximizing the logarithm of the likelihood function in equation (3) and obtaining the following results.

$$\begin{aligned}
\ln L(\beta, \mathbf{X}) &= \ln \left[\left\{ \prod_{j,i=1}^n \left(\frac{1}{1 + e^{\sum_{p=0}^m \beta_p x_{ip}}} \right) \right\} \left\{ e^{\left(\sum_{p=0}^m \beta_p \left(\sum_{i=1}^n y_i x_{ip} \right) \right)} \right\} \right] \\
&= \left\{ \ln \left(e^{\left(\sum_{p=0}^m \beta_p \left(\sum_{i=1}^n y_i x_{ip} \right) \right)} \right) \right\} \left\{ \ln \left(\prod_{j,i=1}^n \left(1 + e^{\sum_{p=0}^m \beta_p x_{ip}} \right)^{-1} \right) \right\} \\
&= \sum_{p=0}^m \beta_p \left(\sum_{i=1}^n y_i x_{ip} \right) - \sum_{j,i=1}^n \ln \left(1 + e^{\sum_{p=0}^m \beta_p x_{ip}} \right).
\end{aligned}$$

The β value is obtained by creating its first child

$$\begin{aligned}
\frac{\partial \ln L(\beta, \mathbf{X})}{\partial \beta} &= \frac{\partial}{\partial \beta} \left[\sum_{p=0}^m \beta_p \left(\sum_{i=1}^n y_i x_{ip} \right) - \sum_{j,i=1}^n \ln \left(1 + e^{\sum_{p=0}^m \beta_p x_{ip}} \right) \right] \\
&= \sum_{j,i=1}^n y_i x_{ij} - \sum_{j,i=1}^n x_{ij} \left(\frac{e^{\sum_{p=0}^m \beta_p x_{ip}}}{1 + e^{\sum_{p=0}^m \beta_p x_{ip}}} \right) \\
&= \sum_{j,i=1}^n y_i x_{ij} - \sum_{j,i=1}^n x_{ij} \pi_j(x_i) \\
&= \sum_{j,i=1}^n x_{ij} (y_i - \pi_j(x_i)) \\
&= \mathbf{X}^T (\mathbf{y} - \hat{\boldsymbol{\pi}}),
\end{aligned} \tag{4}$$

with \mathbf{y} is vector observation on a response variable that is sized $n \times 1$, while \mathbf{X} is a matrix of predictor variables of size $n \times (p+1)$.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & a_{1p} \\ 1 & x_{21} & x_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

If value of $\frac{\partial \ln L(\beta, \mathbf{X})}{\partial \beta} = 0$ and $\hat{\mathbf{y}} = \hat{\boldsymbol{\pi}}$, So we get equation (5) as follows

$$\mathbf{X}^T (\mathbf{y} - \hat{\mathbf{y}}) = 0, \tag{5}$$

then the results of estimating the parameters of $\hat{\beta}$ are obtained as follows:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{z}$$

with \mathbf{z} is an $n \times 1$ vector and \mathbf{W} is a weight vector, with \mathbf{z} as follows.

$$\mathbf{z} = \text{logit}[\hat{\boldsymbol{\pi}}(\mathbf{x})] + \frac{y - \hat{\boldsymbol{\pi}}(\mathbf{x})}{\hat{\boldsymbol{\pi}}(\mathbf{x}) [1 - \hat{\boldsymbol{\pi}}(\mathbf{x})]}.$$

The covariance matrix for $\hat{\beta}$ is shown in the following equation.

$$\text{Var}(\hat{\beta}) = (\mathbf{X}^T \text{diag}[\hat{\boldsymbol{\pi}}(1 - \hat{\boldsymbol{\pi}})] \mathbf{X})^{-1}.$$

Logistic Regression CERP

Logistic Regression Classification by Ensembles from Random Partitions (LR CERP) is a classifier using an ensemble where LR is the basis of the classifier. The space of predictor variables in the LR CERP algorithm is partitioned randomly into several mutually exclusive subspaces. For example, a predictor space Θ is partitioned into mutually exclusive subspaces $(\theta_1, \theta_2, \dots, \theta_n)$ of the same size. Therefore in each subspace it can be assumed that there is no bias [11].

The performance of LR CERP corresponds to the number of predictor variables in each partition. Before determining the number of variables in a partition, it is necessary to determine the most optimal number of partitions. The amount of data influences the determination itself. Data with more data (n) than the number of predictor variables (p) can be partitioned using the following equation:

$$K = \frac{p}{q}, \quad (6)$$

where q is an integer number that is less than n . As for data that has a smaller number of data (n) than many predictor variables (p), the most optimal partition can be obtained using the following equation.

$$K = \frac{6 \times p}{n}.$$

The explanation of LR CERP can be depicted in Figure 1.

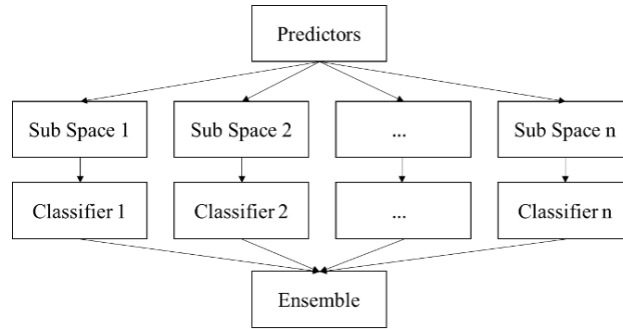


FIGURE 1: LR CERP Concept Chart

A classification model will be built for each subspace using Logistic Regression (LR), but LR models are weak to variable selection [12]. LR CERP is used to improve accuracy where the classification results in each subspace formed will be combined. Combining several LR models performed in LR CERP by taking the average predictive value generated by each ensemble is used to improve accuracy. The predicted values generated from all classification results in a subspace are averaged and classified as 1 or 0 based on a threshold [13].

Logistic Regression Ensemble (LORENS)

One method to improve the logistics method is the ensemble method. Classification analysis based on poor models can be corrected using the ensemble method. One of the logistic regression developments with the ensemble concept is the Logistic Regression Ensemble (LORENS) [14]. Logistic Regression Ensemble (LORENS) is a classification method developed by Lim et al. in 2010. The basis of LORENS classification is to use the LR model developed based on the LR CERP algorithm. When the LR model results are combined in LORENS to get a strong classification compared to other complex aggregation methods. LORENS forms multiple ensembles by repeating the LR CERP procedure multiple times [15].

Like LR CERP, LORENS partitions the data into k subspaces determined by randomization and the same distribution. It can be assumed that variable selection is not biased in each subspace. Each subspace uses the LR model without variable selection. Then from the random variable partitioning, it is expected to obtain almost the same probability of error from k classifications to improve accuracy. LORENS can improve accuracy in each ensemble generated.

The increase in accuracy in one ensemble is obtained from the combination of the predictive value of each LR model in each partition. The repetition of the LR CERP procedure in LORENS produces an average combination of the most values with almost the same accuracy. LORENS uses an average result with a little value that is superior to the most value [16].

Some ensembles generated by LORENS have different partitions according to the most value of each ensemble. Based on these values, one general accuracy is obtained, which has been improved. Generally, this improved accuracy will be obtained if the number of ensembles built is more than 10. The classification process in LR is generally based on a probability threshold. The threshold commonly used in classification is 0.5, but the classification accuracy will only be good if the proportion of class 0 and class 1 is not equal. To equalize sensitivity and specificity, LORENS uses an optimal threshold. The threshold is obtained using the following formula [17]:

$$Threshold = \frac{\bar{y} + 0.5}{2}, \quad (7)$$

where \bar{y} is the proportion of positive responses contained in the data.

The classification process in the LORENS method requires the following steps [18]:

1. Form a logit model based on the training data;
2. Ensuring the testing data into the logit model so that the predicted value is obtained;
3. Classify observations on testing data; if the probability value obtained is smaller than the threshold value, then the observation is in the negative class. Otherwise, if the probability obtained is greater than the threshold value, then the observation is in a positive class;
4. Compare the results of the actual class with the classification prediction class;
5. Group the comparison results into True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) groups.

True Positive (TP) is the total positive class infant deaths predicted in the positive class, True Negative (TN) is the total negative class infant deaths predicted in the negative class, False Positive (FP) is the total negative class infant deaths predicted to be positive, and False Negative (FN) is the total positive class infant deaths predicted to be negative.

TABLE I: Confusion Matrix of Actual Class and Predicted Class.

		Actual Class	
		$p(+)$	$n(-)$
Predict Class	$p(+)$	TP	FP
	$n(-)$	FN	TN

Prediction accuracy in classification can be obtained by dividing the number of correct predictions by the total number of predictions. The following is the formula used to obtain classification accuracy in the form of accuracy [19].

$$Accuracy = \frac{TP + TN}{(TP + FP + TN + FN)}. \quad (8)$$

LORENS has another advantage free of assumption about the dimension of data as LORENS does the partition randomly. LORENS has been proven to have a good performance also to model a very large dataset number of observation [20]. The explanation of LORENS can be depicted in Figure 2.

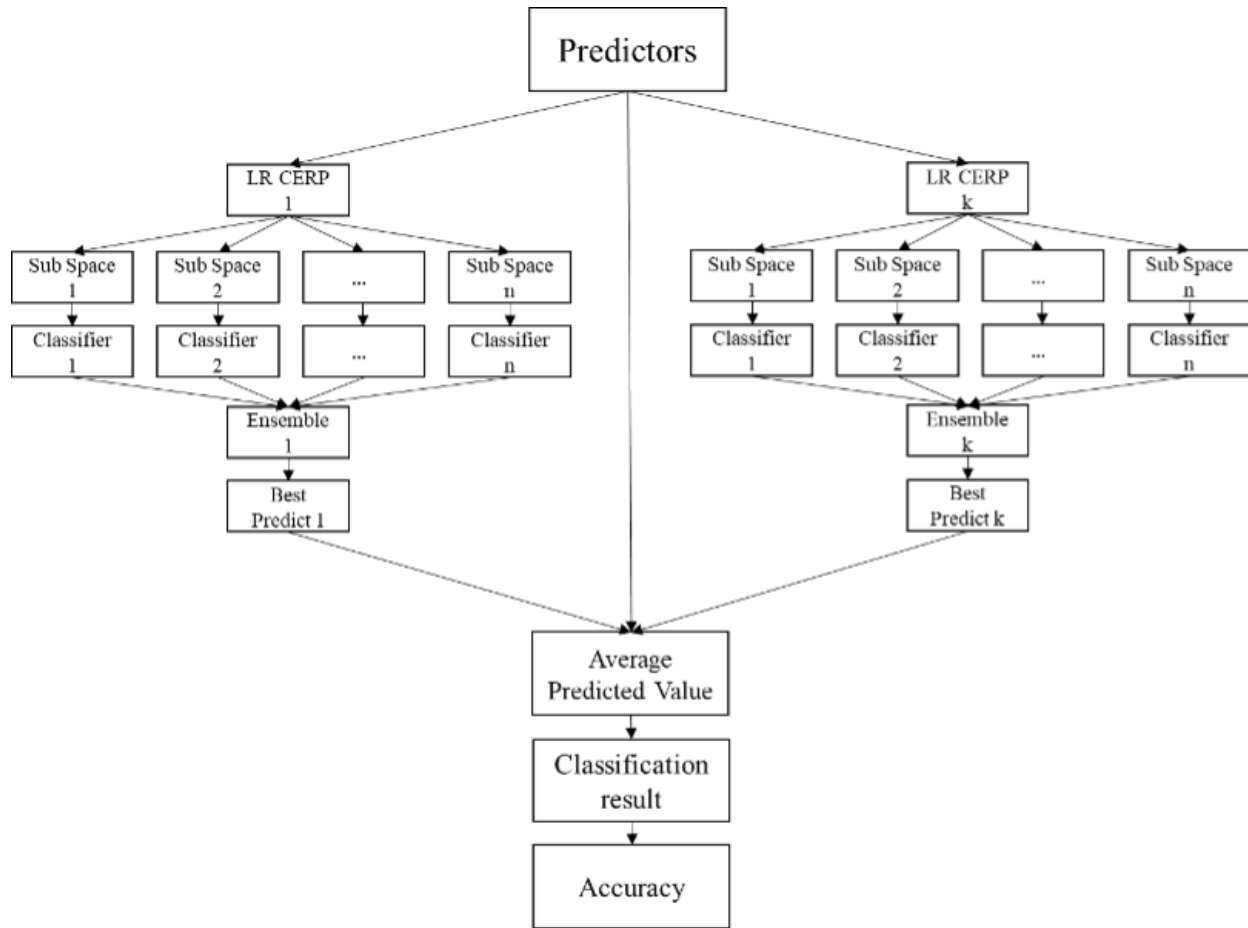


FIGURE 2: LORENS Concept Chart

Evaluation of Model Performance

The method used to evaluate model performance is cross-validation. Cross-validation is a data resampling method to assess the generalization ability of predictive models and to prevent overfitting [21]. The cross-validation method is a method to evaluate model performance by dividing data into k folds or partitions that have an equal number. Cross-validation also divides the data into testing and training data based on these partitions. All partitions are used as both testing data and training data. The use of data as testing data or training data is done during each turn. The turn of each data can be done because of the iterative procedure of cross-validation. Suppose one partition is used as testing data in cross-validation, and the remaining $k - 1$ partitions are used as training data. Furthermore, it is repeated until all partitions have been used as testing data. This procedure is called k fold cross-validation [22].

Infant Mortality

The infant mortality rate (IMR) is the ratio of infant deaths to live births in a given year [23]. According to [1] the Government Work Plan (RKP) of the State Budget (APBN) DPR RI, infant mortality is a death that occurs in a baby after birth until the baby is less than one year old. Infant mortality rate (IMR) is the number of infant deaths per 1000 births. Determination of the level of public health problems can be seen from the IMR level. IMR is one of the indicators launched by the national health system in realizing the success of health development, even used as a central indicator of health success in Indonesia.

The causes of infant mortality are divided into two types: exogenous and endogenous. Exogenous infant mortality

is death in infants caused by factors related to the external environment, such as the level of knowledge of a mother or even malnutrition due to economic factors. In contrast, endogenous is death in infants caused by factors innate to the baby from birth, such as premature birth or congenital abnormalities [24].

IMR of the world in 2021 was 28 deaths per 1000 live births, lower than 2020 at 29 deaths per 1000 live births. Meanwhile, the IMR in Indonesia in 2021 is 18.9, lower than the IMR level in previous years [2]. It shows that IMR both in Indonesia and on a world scale is declining, meaning that Public Health services have succeeded in reducing IMR cases. This effort is one of the efforts to be able to achieve the SDGs target by 2030. The Sustainable Development Goals (SDGs) or also known as the Global Goals are a call to end poverty, protect the planet, and ensure that by 2030 all people enjoy peace and prosperity [25]. The SDGs target related to IMR is 12 deaths per 1000 deaths by 2030.

Factors affecting infant mortality include lack of maternal awareness of health, mothers rarely checking their wombs to midwives, pregnancy at a young age, lack of nutritional intake for mothers and their babies, pregnancy at an old age, food consumed less hygienic, and many more. In addition, the mother's condition during pregnancy also affects the health of her womb, such as psychological, physical, social, and cultural factors.

METHOD

Data Source

The data used in this study is secondary data about infant mortality rate (IMR) data in Indonesia obtained from the documentary results of the BPS census [26], IMR data consists of 9495 live babies and 5505 dead babies.

Study's Variables

This study uses the infant mortality variable as a response variable in the form of categorical data for live infants and dead babies. In addition, it also uses infant mortality factors consisting of 13 factors as predictor variables. Where birth baby weight as x_1 , Age of mother at childbirth as x_2 , Mother's Education as x_3 , Mother's job as x_4 , Birth interval as x_5 , Antenatal care as x_6 , Place of delivery as x_7 , Baby gender as x_8 , Breastfeeding as x_9 , Type of birth as x_{10} , Birth order as x_{11} , Residence as x_{12} , and Family wealth index as x_{13} .

Analysis Steps

The process of classifying infant mortality rates using the logistic regression ensemble (LORENS) method is carried out with the following

1. Group Training data 85% and Testing Data 15%;
2. Determine the number of l partitions using equation (6) and m ensembles following LORENS concept;
3. Base on 2nd step, Partition the variable predictor in k subspace;
4. Build LR models on partitions and training data following equation (2);
5. Calculates the predicted value of each prediction and then averages it;
6. Find the best predicted value from each ensemble following LORENS concept;
7. Calculate the most optimal threshold value using equation (7);
8. Classification of predictions in 0 and 1 based on threshold values;
9. Calculate accuracy using confusion matrix using equation (8);
10. Evaluate model performance using k fold cross-validation where the data is divided into k folds.

RESULT AND DISCUSSION

LORENS Analysis

LR Models on Training Data

The LORENS analysis was carried out with a comparison of the best training and testing data, which is 85%: 15%, besides that the optimal ensemble in this study is ten with two partitions. The LR model formed in 10 ensembles is 2 LR models per partition. Furthermore, the study used the LR model formed from 2 partitions is 20 LR models. This model is built based on the LR model coefficient in each partition and the LR model coefficient in each variable.

TABLE II: LR Model Coefficients for Each Partition.

Intercept	Ensemble									
	1	2	3	4	5	6	7	8	9	10
1st Partition	-21.90	-21.36	-4.32	-20.35	-20.35	-1.95	-20.65	-20.29	-18.41	-3.66
2nd Partition	-0.24	-0.50	-18.80	-3.24	-3.24	-20.05	-1.56	-2.64	-3.70	-18.47

TABLE III: LR Model Coefficients for Each Variables.

Variable	Ensemble									
	1	2	3	4	5	6	7	8	9	10
x_1	1.19	1.19	1.13	1.27	1.27	1.05	1.02	1.17	1.17	1.24
x_2	0.02	0.01	0.06	0.06	0.06	0.03	0.03	0.06	0.03	0.02
x_3	0.18	0.11	0.21	0.24	0.24	0.13	0.11	0.00	0.29	0.24
x_4	-0.72	-0.44	-0.57	-0.66	-0.66	-0.32	-0.34	-0.62	-0.61	-0.65
x_5	-0.26	-0.22	-0.69	-0.67	-0.67	-0.20	-0.21	-0.65	-0.68	-0.66
x_6	0.31	0.21	0.35	0.26	0.26	0.21	0.19	0.34	0.35	0.29
x_7	-0.17	-0.05	-0.01	0.00	0.00	0.09	0.10	0.00	-0.15	-0.14
x_8	0.62	0.69	0.59	0.68	0.68	0.58	0.58	0.44	0.48	0.67
x_9	18.45	18.37	19.32	19.37	19.37	18.43	18.43	19.36	19.28	19.38
x_{10}	1.89	1.93	1.16	1.82	1.82	1.60	1.54	0.99	1.53	1.55
x_{11}	0.30	0.30	0.25	0.32	0.32	0.21	0.21	0.28	0.26	0.30
x_{12}	0.33	-0.03	0.14	0.08	0.08	0.06	0.16	0.25	0.16	0.11
x_{13}	0.05	0.08	0.10	0.05	0.05	0.08	0.12	0.07	0.07	0.09

^a Bold number: 1st partition; Otherwise: 2nd partition

Table 2 and 3 show the coefficients of the LR model formed from each partition space in each ensemble based on the parameter estimates in equation (4). These values show the influence that each variable has on infant mortality. In addition, the table also shows the division of partitions in each ensemble that is done randomly. In the first ensemble it is known that the first partition consists of variables $x_1, x_2, x_7, x_8, x_9, x_{11}, x_{12}$ and the rest is the second partition. This LR model was built by substituting the values of $\beta_0, \beta_1, \dots, \beta_p$ in equation (2). The following is the LR model formed in the first ensemble.

$$\pi_1(x_1) = \frac{e^{-21.9+1.19x_1+0.02x_2-0.17x_7+0.62x_8+18.45x_9+0.30x_{11}+0.33x_{12}}}{1 + e^{-21.9+1.19x_1+0.02x_2-0.17x_7+0.62x_8+18.45x_9+0.30x_{11}+0.33x_{12}}} \quad (9)$$

$$\pi_2(x_1) = \frac{e^{-0.24+0.18x_3-0.72x_4-0.26x_5+0.31x_6+1.89x_{10}+0.05x_{13}}}{1 + e^{-0.24+0.18x_3-0.72x_4-0.26x_5+0.31x_6+1.89x_{10}+0.05x_{13}}} \quad (10)$$

The first LR model, denoted $\pi_1(x_1)$, was formed from the predictor variables in the first partition. The second LR model, denoted $\pi_2(x_1)$, was formed from the predictor variables in the second partition. Based on the LR model

obtained and equation (3), the logit transformation in the first ensemble is as follows

$$\begin{aligned} \text{Logit} [\pi_1(x_1)] &= \ln \left(\frac{\pi_1(x_1)}{1-\pi_1(x_1)} \right) \\ &= -21.9 + 1.19x_1 + 0.02x_2 - 0.17x_7 + 0.62x_8 + 18.45x_9 + 0.30x_{11} + 0.33x_{12} \end{aligned} \quad (11)$$

$$\begin{aligned} \text{Logit} [\pi_2(x_1)] &= \ln \left(\frac{\pi_2(x_1)}{1-\pi_2(x_1)} \right) \\ &= -0.24 + 0.18x_3 - 0.72x_4 - 0.26x_5 + 0.31x_6 + 1.89x_{10} + 0.05x_{13} \end{aligned} \quad (12)$$

The Logit models in Equations (11) and (12) show the impact of each predictor variable on the response variable of infant mortality. A variable with a positive β value indicates that it has a higher impact on infant mortality. Otherwise, the value of β indicates that the variable had a lower impact on infant mortality. As in the first ensemble, the second ensemble is treated similarly to obtain the LR model. The LR model formed in each partition and ensemble was used to calculate the probability value of each pair of testing data to predict the class of testing data. Based on Table 3 and logit model equations (9) and (10), it can be seen that the variable that has the highest impact on infant mortality is x_9 or factor breastfeeding.

Classification Prediction Class

The prediction value in this study was obtained by substituting the testing data in the LR. The prediction value in this study was obtained by substituting the testing data in the LR. The predicted value was obtained from each ensemble's average probability of each partition. Based on this step, the prediction results for each ensemble are obtained as follows. The final prediction result was obtained by voting by choosing the most predictive decision from all ensembles. If the final probability value exceeds the optimal threshold, the testing data are included in class 1, which is a dead baby. In contrast, if the final probability value is less than the optimal threshold value, the testing data enters class 0, a live baby. The optimal threshold value obtained is 0.4334118 based on equation (7). The following is the average probability value and classification class decision.

TABLE IV: Prediction and Classification Decision Results.

Testing Data	Mean Predict	Comparison Threshold	Decision
1	0.850	>0.4334	1
10	0.252	<0.4334	0
11	0.380	>0.4334	0
16	0.623	<0.4334	1
17	0.249	<0.4334	0
21	0.254	>0.4334	0
28	0.196	<0.4334	0
33	0.322	<0.4334	0
44	0.193	<0.4334	0
46	0.264	<0.4334	0
54	0.361	>0.4334	0
55	0.419	<0.4334	0
58	0.158	<0.4334	0
78	0.492	<0.4334	1
⋮	⋮	⋮	⋮
14995	0.535	<0.4334	1

Table 4 shows that the first average probability value of the testing data was 0.85. Thus, value is higher than the optimal threshold, meaning that the classification decision is to enter class 1. In the second testing dataset, the average probability value was 0.252, meaning that the second average probability value fell into class 0. The average probability value in the last testing dataset exceeds the threshold value of 0.535. Therefore, all classification decisions fall into class 1.

Accuracy of Classification Results

Based on the classification class decisions in table 8, it can be included in the confusion matrix tabulation calculation, namely the comparison of actual class and predict class. Then table 5 shows accuracy classification using equation (8).

TABLE V: Confusion Matrix Classification.

		Actual Class	
		$p(+)$	$n(-)$
Predict Class	$p(+)$	581	244
	$n(-)$	218	1207

Based on Table 5, the True Positive (TP) value is 581, meaning 581 cases of death in infants who are correctly predicted to die. The False Negative (FN) value was 218, meaning 218 death cases in babies predicted to live. The True Negative (TN) value is 1207, meaning 1207 live babies are correctly predicted to live. The False Positive (FP) value is 244, meaning that there are 244 cases of live babies who are predicted to die. Based on equation (8), the accuracy of the LORENS classification on the IMR is as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{581+1207}{581+244+218+1207} \times 100\% \\ &= 79.47\%. \end{aligned}$$

The accuracy of the calculation was 79.47%. This means that the accuracy of the LORENS model for the IMR has a good value. This classification can be used as a basis and knowledge of the Indonesian government and society, especially mothers and prospective mothers. In addition, mothers have more pay attention to breastfeeding their children. Furthermore, Indonesia can reduce IMR cases and become one of the countries that have succeeded in achieving the SDGs target.

Model Performance Evaluation

Cross-validation was used to evaluate the goodness of fit of the classification model used in this study. All data acted as training data and testing data according to their respective turns. The study divides the data into ten folds, each acting as training and testing data. When the first fold was used as the testing data, the second to tenth folds were used as the training data. When the second fold was used as the testing data, the first, third, and 10th fold were used as the training data. The same pattern is done until all folds act as training and testing data.

TABLE VI: Optimal Threshold Results on cross-validation.

Fold	Optimal Threshold	Fold	Optimal Threshold
1	0.43289	6	0.43233
2	0.43304	7	0.43304
3	0.43444	8	0.434
4	0.43337	9	0.43474
5	0.43311	10	0.43304

Table 6 shows the results for each optimal threshold for each divided fold. Furthermore, a performance evaluation was performed using a cross-tabulation table to determine the accuracy of the classification model. The following are the results of the classification cross-tabulation of the IMR data for two partitions.

Based on Table 7, the True Positive value is 3848, meaning 3848 cases of death in infants correctly predicted to die. The False Negative value was 1515, meaning that 1515 cases of dead babies were predicted to live. The True Negative value is 7983, meaning there are 7983 cases of live babies who are correctly predicted to live. The False

TABLE VII: Confusion Matrix of Classification Results Cross-validation.

Predict Class		Actual Class	
		$p(+)$	$n(-)$
	$p(+)$	3848	1654
	$n(-)$	1515	7983

Positive value is 1654, meaning there are 1654 cases of live babies predicted to die. Therefore, an accuracy value was obtained for the following model evaluation.

$$\begin{aligned} \text{Accuracy} &= \frac{3848+7983}{3848+1654+1515+7983} \times 100\% \\ &= 78.87\% \end{aligned}$$

The results of the accuracy calculation were 78.87%. Based on the evaluation of model performance using cross-validation, the implementation of LORENS for IMR classification has a fairly good accuracy level.

CONCLUSION

The conclusion based on LORENS in the infant mortality rate (IMR) classification is obtained from as many as 20 LR models. Based on the model, it is known that the most influential factor in Indonesia's infant mortality rate is breastfeeding. The accuracy level in classifying infant mortality rates (IMR) using LORENS is 79.47%. The former model can be classified with an accuracy level of 79.47%. LORENS classification with model performance evaluation in cross-validation used data division in 10 folds. Each fold is used as the training data and the testing data. The use of 10-fold cross-validation yielded an accuracy of 78.87%.

ACKNOWLEDGMENTS

The author would like to thanks the Badan Pusat Statistika (BPS) to provide data on infant and Dwi Iwan Kharisma for useful suggestions for this study.

REFERENCES

1. R. Alvaro, R. Christianingrum, and T. Riyono, "Analisis rkp dan pembicaraan pendahuluan apbn," Pusat Kajian Anggaran Badan Keahlian DPR RI: Jakarta (2021).
2. World Bank Group, "Mortality rate, infant (per 1,000 live births)," (2021).
3. T. D. A. Widhianingsih, H. Kuswanto, and D. D. Prastyo, "Logistic regression ensemble (lorens) applied to drug discovery," *Matematika*, 43–49 (2020).
4. D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression* (John Wiley & Sons, 2013).
5. R. Zakharov and P. Dupont, "Ensemble logistic regression for feature selection," in *Pattern Recognition in Bioinformatics: 6th IAPR International Conference, PRIB 2011, Delft, The Netherlands, November 2-4, 2011. Proceedings 6* (Springer, 2011) pp. 133–144.
6. S. Ratnawati and S. Sunendiari, "Penggunaan metode logistic regression ensemble (lorens) pada klasifikasi leukemia akut," *Prosiding Statistika* 7, 56–63 (2021).
7. A. Asfihani, *Prediksi Pembelotan Konsumen Software Antivirus "X" dengan Binary Logistic Regression dan Logistic Regression Ensembles*, Ph.D. thesis, Institut Teknologi Sepuluh Nopember (2015).
8. H. Kuswanto, J. N. Melasasi, and H. Ohwada, "Enzyme classification on dud-e database using logistic regression ensemble (lorens)," *Innovative Computing, Optimization and Its Applications: Modelling and Simulations*, 93–109 (2018).
9. D. W. Hosmer Jr and S. Lemeshow, *Applied Logistic Regression second Edition* (John Wiley & Sons, 2000).
10. A. Agresti, *Categorical data analysis, 3rd ed* (John Wiley & Sons, 2013).
11. J. N. Melasasi, *Klasifikasi Enzim Pada Database Dud-E Dengan Metode Logistic Regression Ensemble (Lorens) Dengan Metode Logistic Regression*, Ph.D. thesis, Institut Teknologi Sepuluh Nopember (2015).
12. N. Lim, *Classification by ensembles from random partitions using logistic regression models* (State University of New York at Stony Brook, 2007).
13. H. Ahn, H. Moon, M. J. Fazzari, N. Lim, J. J. Chen, and R. L. Kodell, "Classification by ensembles from random partitions of high-dimensional data," *Computational Statistics & Data Analysis* 51, 6166–6179 (2007).

14. T. L. Malau, T. Joseph, *et al.*, “Analisis metode logistik regresi ensemble untuk klasifikasi dengan pra-pemrosesan menggunakan principal component analysis,” *IJM: Indonesian Journal of Multidisciplinary* **1**, 707–722 (2023).
15. T. D. A. Widhianingsih, *Klasifikasi Data Berdimensi Tinggi dengan Metode Ensemble berbasis Regresi Logistik dalam permasalahan Drug Discovery*, Ph.D. thesis, Institut Teknologi Sepuluh Nopember (2018).
16. N. Lim, H. Ahn, H. Moon, and J. J. Chen, “Classification of high-dimensional data with ensemble of logistic regression models,” *Journal of Biopharmaceutical Statistics* **20**, 160–171 (2009).
17. H. Kuswanto, A. Asfihani, Y. Sarumaha, and H. Ohwada, “Logistic regression ensemble for predicting customer defection with very large sample size,” *Procedia Computer Science* **72**, 86–93 (2015).
18. R. W. Werdhana, *Klasifikasi Gen Yang Terkait Sindrom Alzheimer Menggunakan Metode Naïve Bayes Classifier Dan Logistic Regression Ensemble*, Ph.D. thesis, Institut Teknologi Sepuluh Nopember (2017).
19. C. Catal, “Performance evaluation metrics for software fault prediction studies,” *Acta Polytechnica Hungarica* **9**, 193–206 (2012).
20. H. Kuswanto and R. W. Werdhana, “Logistic regression ensemble to classify alzheimer gene expression,” in *2017 International Conference on Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS)* (IEEE, 2017) pp. 36–41.
21. D. Berrar *et al.*, “Cross-validation.” (2019).
22. I. H. Witten, E. Frank, and M. A. Hall, *Data mining: practical machine learning tools and techniques* (Morgan Kaufmann, 2011).
23. S. L. Murphy, K. D. Kochanek, J. Xu, and E. Arias, “Mortality in the united states, 2020,” (2021).
24. L. Azizah, “Pengujian signifikansi model geographically weighted regression (gwr) dengan statistik uji f dan uji t,” Malang: UIN Maulana Malik Ibrahim Malang: Skripsi (2013).
25. United Nations Development Programme, “The sdgs in action,” (2023).
26. Badan Pusat Statistik, “Kasus angka kematian bayi di indonesia tahun 2017,” (2017).