BAREKENG: Journal of Mathematics and Its Applications
March 2024 Volume 18 Issue 1 Page 0537–0552
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss1pp0537-0552

EXAMINING RISK FACTORS OF ANEMIA IN PREGNANCY USING HYBRID LOGISTIC REGRESSION MODEL AND ROUGH SET THEORY

Izzati Rahmi[1*], Riswan Efendi[2], Nor Azah Samat[3], Hazmira Yozza[4], Mahdhivan Syafwan[5]

[1,2,3]*Mathematics Department, Faculty of Science and Mathematics Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia*
[1,4,5]*Department of Mathematics and Data Science, Faculty of Mathematics and Natural Science, Andalas University, Limau Manis, Padang, 25175, Indonesia*
*Corresponding author's e-mail: * **izzatirahmihg@sci.unand.ac.id***

Article History:
Received: 2nd November 2023
Revised: 28th January 2024
Accepted: 20th February 2024

Keywords:
Anemia in pregnancy;
Logistic Regression;
Rough Set Theory;
Inconsistent Sample;
Hybrid Model.

ABSTRACT

Anemia in pregnancy is a potential danger to the mother and child. Therefore, the risk of anemia in pregnant women requires serious attention from all relevant parties. Considering the numerous negative effects caused by anemia in pregnant women, efforts must be made to prevent and treat anemia in pregnant women by understanding the factors that influence it. This study assesses the risk factors for anemia in pregnant women at Tegal Rejo Community Health Center, Yogyakarta Province. In this paper, a new integrated classification approach with binary logistic regression (LR) analysis and Rough Set Theory (RST) is proposed, in order to examine factors on the incidence of anemia in pregnancy. The proposed model is called the Logistic Regression and Reduction Rough Set (LR3S). In LR3S model, the RST technique is used to detect inconsistent sample and removing inconsistent sample that have probability less than 0.5 before doing LR modelling. To evaluate the development of the resulting model, a comparison was made between the performance of Original Logistic Regression (OLR), LR model after removing outlier namely as Remove Outlier Original Logistic Regression (RO2LR), and LR3S. Using a number of model performance metrics, it is found that LR3S has the best performance for the three models used. Using LR3S model, it is found that CED status, educational level, parity and gestational are significant variable impact on the incidence of anemia.

How to cite this article:

I. Rahmi,R. Efendi, N. A. Samat, H. Yozza and M. Syafwan., "EXAMINING RISK FACTORS OF ANEMIA IN PREGNANCY USING HYBRID LOGISTIC REGRESSION MODEL AND ROUGH SET THEORY," *BAREKENG: J. Math. & App.*, vol. 18, iss. 1, pp. 0537-0552, March, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*
Research Article · Open Access

1. INTRODUCTION

Anemia in pregnancy is a national problem that has a significant impact on the quality of human resources and reflects the value of the socioeconomic welfare of the community [1]. The 2018 national basic

health research recorded that 48.9% of pregnant women in Indonesia were anemic. It means that the prevalence of anemia in pregnant women in Indonesia is high. With an anemic prevalence limit of 40%, this number categorizes the problem of anemia during pregnancy as a severe public health problem [2].

Anemia during pregnancy is a potential danger to the mother and child. Therefore, the risk of anemia in pregnant women requires serious attention from all relevant parties. Previous studies have found that anemia during pregnancy has certain adverse effects of on the mother and fetus, including increasing the risk of low birth weight babies [3]–[5], premature birth [4], [6]–[8], fetal death [6]–[9], postnatal infant death [4], [10] and maternal death [11], [12].

Considering the numerous negative effects caused by anemia in pregnant women, efforts must be made to prevent and treat anemia in pregnant women by understanding the factors that influence it. Several studies in the literature suggest that gestational age [13]–[15], age of the pregnant woman [13], [16], parity [14]–[16] employment status [17], chronic energy deficiency status [18] and education level [17], [19], are variables that impact the occurrence of anemia in pregnant women.

When analyzing the factors influencing the incidence of anemia, one statistical method that is commonly employed is logistic regression (LR) [15], [20]–[22]. However, several prior studies have revealed weaknesses in the application of LR analysis, such as the existence of outliers in the data. These outliers can unduly influence the results of the analysis and lead to incorrect inferences [23]–[25]. In addition, LR cannot establish a causal relationship between anemia and the identified independent variables [26], [27].

The study of factors affecting anemia incidence is a categorical response analysis, also known as a classification problem. The rough set theory (RST) is an alternative method to analyze categorical responses. Pawlak et al. developed RST in the early 1970s [28]. RST has received increased attention as a method of data analysis in several research disciplines, including the health sector [29]–[31]. RST is an effective tool for managing inconsistent samples and extraneous attributes, which can have a substantial impact on the performance of classification models, such as logistic regression models.

Since LR and RST are two methods commonly used in classification problems, some prior researches discussed integrating and comparing LR and RST for classification tasks. Most of these studies on integrating RST and LR focused on data reduction via attribute selection to anticipate the possibility that there are attributes that are not significant or irrelevant to the dependent variable. They employed data reduction to increase the classification problem's accuracy [32]–[36].

Meanwhile, Rasyidah et al. (2022) demonstrated that employing inconsistent sample reduction using RST and then integrating it with the ordinary linear regression might increase the performance of the model [37]. Since RST and LR are two analyses for classification problem, this study will propose a hybrid analysis that integrates LR with the reduction of sample inconsistencies with RST. This proposed method is supported by Gludice et al. (2017) research which showed that the rough set model performs well with reduced samples and under uncertain conditions [38]. However, no study has been conducted that incorporates data reduction by eliminating inconsistent data using RST for further analysis using LR.

This study aims to examine the risk factors for anemia in pregnant women using a new hybrid logistic regression model and Rough Set theory. Furthermore, the proposed model is called Logistic Regression and Reduction Rough Set (LR3S). To show the advantages of the LR3S model, this model will be compared with two other models commonly used in previous studies, namely the OLR (Original Logistic Regression) model and RO2LR (Removing Outlier Original Logistic Regression) model using several model performance indicators.

2. RESEARCH METHODS

2.1 Data Sources and Research Attributes

The data used in this study were obtained from Padmi's (2018) research on factors influencing the incidence of anemia in pregnant women at Tegal Rejo Health Centre, Yogyakarta **[39].** In the rough set theory, independent variables are denoted as condition variables and dependent variable is denoted as decision variable. The decision variable in this research is the incidence of anemia. Five conditional attributes are thought to affect the incidence of anemia as presented in **Table 1**, namely gestational age (X1), age of pregnant women (X2), parity (X3), employment status (X4), chronic energy deficiency (CED) status (X5), and education level (X6).

Table 1. Research Attributes/Variables

Type of Attribute/variable	Name of Attribute/ variable	Category
Condition Attributes / Independent Variables	Gestational age (X1)	(1) At risk (2) Non-risk
	Age of pregnant women(X2)	(1) At risk (2) Non-risk
	Parity (X3)	(1) At risk (2) Non risk
	Employment (X4)	(1) Do not work (2) Work
	CED Status (X5)	(1) CED (2) Not CED
	Education level (X6)	(1) Less educated (2) Educated
Decision Attributes / Dependent Variables	The incidence of anemia (Y)	(1) Yes (2) No

2.2 Fundamental Concepts of Rough Set Theory

The Rough Set Theory (RST) is a useful technique for dealing with vagueness and imprecision in intelligent data analysis and data mining, primarily in the context of data classification **[40].** RST and its applications have received a lot of attention during the last 20 years **[41].** Using RST, one may estimate the decision rules for classifying objects that are presented in a table called the decision table. The process of determining such decision rules is referred to as "reduct discovery". Using the rough set algorithm, reduct, as a decision table pattern, may create a classifier to categorize new items.

a. Decision Table

The decision table is a table in which each row represents the research objects and columns represent the research's conditional and decision attributes. Another way to define a decision table is

$$IS = (U, A) = (U, At = C \cup D, \{V_a | a \in At\}, I_a | a \in At) \quad (1)$$

where U is a finite non-empty set of n objects $\{x_1, x_2, \dots, x_n\}$; At is a finite non-empty set of attributes that is composed of a set of condition attributes C, which describe the objects, and a decision attribute D, which defines the class of the object; $V_a$is a non-empty set of values $a \in At$; $I_a : U \rightarrow V_a$ is a function that maps objects from U to exactly one value in V_a **[42].**

The decision table is considered to be consistent if every pair of objects that have the same conditional value in C also has the same decision value in D. On the other hand, when one pair of object has the same conditional value in C but a different decision value in D, then the decision table is considered to be

inconsistent. In general, when the decision table is constructed based on measurement or observation, the data is likely to experience inconsistencies **[43].**

b. Indiscernibility relation

The central concept in RST is the indiscernibility relation. This concept is considered as a relation between two objects or more, where all the values are identical in relation to a subset of considered condition attributes. Suppose *IS* is a decision table and A is a subset of condition attributes, $A \subseteq C$. The indiscernibility relation $IND(A) \subseteq U \times U$ is defined as

$$IND(A) = \{(x, y) \in U \times U | \forall a \in A, I_a(x) = I_a(y)\}. \quad (2)$$

The indiscernibility relation of the set of all condition attributes is called the equivalent class, which is then used to obtain a discernibility matrix **[44].**

c. Approximations

Another key term in RST is approximation. This concept is related to the meaning of the topological operations of approximations **[43].** Lower and upper approximations are defined based on the definitions of an indiscernibility relation and an equivalence class. The following explanations present and describe the various forms of approximations that are employed in RST.

Let $S = (U, A)$ be an information system and let $B \subseteq A$, and $X \subseteq U$. Now, the approximate of a set X can be made using the information contained in the set of attributes B by constructing the *B-lower* and *B-upper* approximations of X, denoted by $\underline{B}$X and $\overline{B}$X respectively, where:

$$\underline{B}X = \{x \in U \mid [x]_B \subseteq X\}\, x \mid \quad (3)$$

and

$$\overline{B}X = \{x \in U \mid [x]_B \cap X \neq \phi\}. \quad (4)$$

Similar to the general case, based on the knowledge in B, the objects in $\underline{B}X$ can certainly be categorized as members of X; on the other hand, the objects in $\overline{B}X$ can only be categorized as possible members of X. The set $BN_B(X) = \overline{B}X - \underline{B}X$ is known as the *B-boundary region of X* and thus comprises those objects that cannot be definitively classified into X based on the knowledge in B. The set $U - \overline{B}X$ is known as the *B outside region of X* and consists of those objects that are certainly classified as not belonging to X (based on the knowledge in B). When the boundary region is non-empty, a set is said to be *rough.* On the contrary, when the boundary region is empty, a set is said to be *crisp.* **Figure 1** depicts a graphical representation of the definition of set approximations explained above.

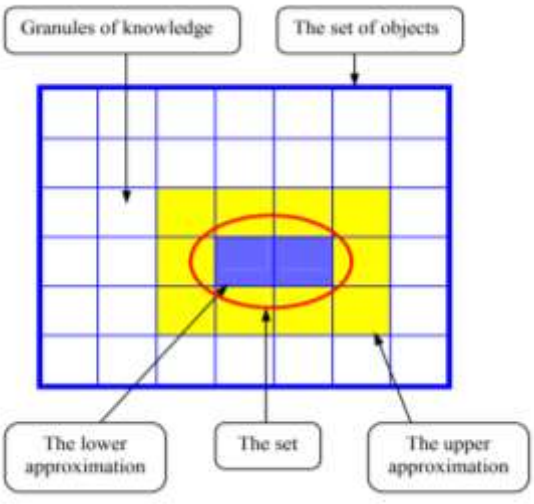

Figure 1. Illustration of approximation set

Figure 1 illustrates the basic concept of an RST. The squares in the figure represent equivalence classes while the ellipse represents the target set X. Equivalence classes are the smallest granularity in the information system because objects in the same equivalence class are indiscernible from one another.

Obviously, based on the equivalence classes (the squares), we cannot exactly define the ellipse. To solve this problem, RST defines a pair of approximations, namely the lower and the upper approximation **[43].** The lower approximation contains all equivalence classes that are completely contained by the ellipse, whereas the upper approximation includes all equivalence classes in the lower approximation as well as those that are partially contained by the ellipse. The figure indicates that a rough set has all the information based on the known attributes.

d. Discernibility Matrix dan Discernibility Function Formation

Two objects are considered discernible if their values differ in at least one attribute. The discernibility matrix $M(x, y)$ is a matrix whose elements consist of a set of attributes that distinguish object x from object y. The discernibility matrix $M(x, y)$ is defined as

$$M(x,y) = \{c \in C | [I_c(x) \neq I_c(y)] \wedge [I_D(x) \neq I_D(y)]\} \quad (5)$$

for $I_D(x) = I_D(y)$, then $M(x, y) = \emptyset$. The discernibility matrix is a symmetric matrix that is, $M(x, y) = M(y, x)$ **[46].**

The discernibility function is a boolean function that is formed from each column of the discernibility matrix. This function, denoted by f_{IS}, is defined as

$$f_{IS} = \wedge\{\vee(M(x,y)) \mid \forall x, y \in U, M(x,y) = \emptyset\}. \quad (6)$$

e. Formation of a Reduct by Simplifying the Discernibility Function

The next stage in RST is the formation of reductases. Reducts are the results of simplification of the discernibility functions of each column of the discernibility matrix using the rules of Boolean algebra. Simplifying the discernibility function means finding another equivalent function with a smaller number of terms or operations.

f. Formation of decision rules

Decision rules are expressed as statements of the form "*if f then g*" denoted as $f \rightarrow g$. The f part represents the value of the condition attribute, while the g part represents the value of the decision attribute. The decision rules are derived from the resulting reduct by examining the table of equivalent classes constructed **[46].**

If the decision table has inconsistencies in its data, the subsequent decision rules will have inconsistencies as well. To address these issues, a quality measure is employed to select decision rules that have inconsistencies. Quality measures are classified as follows: support, strength, accuracy, and coverage **[47].** In this research, strength is employed as a reference to choose which decision to make. Assume IS is a decision table, support is the number of objects that match the decision and condition decision attributes of the decision rule, and *card(U)* is the total number of objects. The strength of the decision rule is formulated

$$Strength = \frac{support}{card(U)} \times 100\% \quad (7)$$

2.3 Fundamental Concepts of Logistic Regression

The logistic regression (LR) is a statistical method used to estimate the probability of a binary outcome, such as the absence or presence of a disease or a specific event. In LR, the probability of the outcome is referred to as the dependent variable (response), and all factors that influence it are referred to as the independent variables (predictors), also known as risk factors **[48]**. For a model to fit the data well, it is expected that the independent variables are uncorrelated with one another and significantly related to the response. Moreover, it is also assumed that data elements of a model are also uncorrelated. The purpose of LR is to estimate the true parameter(s) of the model's underlying probability density function based on the response as adjusted by its predictors **[49]**.

Logistic Regression Model

Logistic regression is a class of regression employed when the dependent variable is dichotomous, whose values can be categorized as occurrence ($Y = 1$) and non-occurrence ($Y = 0$). The independent variables in LR can be of any type. Suppose x is a certain event. The probability of occurrence is denoted by $\pi(x) = P((Y = 1)|x)$. Thus, $1 - \pi(x) = P((Y = 0)|x) = 1 - P((Y = 1)|x)$ represents non-occurrence. The predicted probabilities are modelled as a natural logarithm (ln) of the odds ratio, and expressed as

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (8)$$

and,

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k} \quad (9)$$

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}} \quad (10)$$

where β_0 is the intercept, while β_1, β_2, ..., and β_k denote the regression coefficients related to x_1, x_2,..., x_k, respectively **[23].**

The logistic regression model in Equation (10) explicitly relates the probability of $Y = 1$ to the predictor variables. The LR attempts to estimate the $k + 1$ unknown parameters in Equation (7) using a maximum likelihood estimation. This method involves determining a set of parameters that maximize the probability of the observed data. The regression coefficients represent the degree of the relationship between each independent variable and the outcome. Each coefficient reflects the amount of change in the response variable that would be expected if the predictor variable changed by one unit. The LR aims to correctly predict the outcome category for an individual case using the best model. To achieve this purpose, a model is built that includes all predictor variables that may be used to predict the response variable. The LR calculates the probability of success over the probability of failure. Finally, the results of the analysis are presented as an odds ratio.

Response of a new observation is determined based on the following discriminant rules:

$$y = \begin{cases} 1, & \pi(x) \geq 0.5; \\ 0, & \pi(x) < 0.5 \end{cases} \quad (11)$$

Parameter Testing

There are two types of logistic regression parameter testing, namely the overall test and the partial test. The overall fit of a model indicates the strength of the relationship between the dependent variable and all of the independent variables, taken together. The fit of the two models with and without the independent variables can be compared to evaluate it. If the model with k independent variables shows improvement over the null model (the model with no dependent variables), it is considered to be a better match for the data. A likelihood ratio test can be used to assess the overall fit of the model with k coefficients. This test verifies the following hypothesis

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
H_1: There is at least one $\beta_j \neq 0$; $j = 1, 2, \ldots, k$

The test statistic used is the G test statistic or the likelihood ratio test

$$G = \chi^2 = -2\log\left(\frac{likelihood\ of\ the\ null\ model}{likelihood\ of\ the\ given\ model}\right) \quad (12)$$

When the overall model fit statistic's p-value is less than the test's significance level, α (p-value < α), the null hypothesis is rejected, indicating that there is evidence that at least one independent variable significantly influences the outcome prediction.

If the overall model works well, the next question is how important each of the independent variables is. Statistical tests of significance can be applied to each variable's coefficients. For each coefficient, the null

hypothesis that the coefficient is zero is tested against the alternative that the coefficient is not zero using a statistic Wald test:

$$W_j = \frac{\beta_i^2}{SE_{\hat{\beta}_j}^2} \tag{13}$$

Each Wald statistic is compared with a χ^2 critical value with 1 degree of freedom.

Interpreting the Odds Ratio in LR Model

The odds ratio (OR) is a comparative measure of two odds relative to different events. For two events *A* and *B*, the corresponding odds of *A* occurring relative to *B* occurring is

$$OR\ \{A\ vs\ B\} = \frac{odds\{A\}}{odds\ \{B\}} = \frac{P_A/_{1-P_A}}{P_B/_{1-P_B}} \tag{14}$$

where P_A and P_B denote the probabilities of conditions *A* and *B* happening respectively.

When presenting the LR's results, odds is usually used instead of the outcome's probability. Probability and odds are directly related; that is, the odds of an event is calculated by dividing the probability of an event occurring by the probability that it will not occur. When the value of an independent variable, X_i, increases by one unit while the values of other variables remain constant, the odds value of the dependent variable increases by a factor of exp(β_i). This factor is usually called the odds ratio (OR). It reflects the relative amount by which the odds of the dependent variable increase (OR > 1) or decrease (OR < 1) as a result of one unit increase in the value of the corresponding independent variable. The OR ranges from zero to positive infinity.

Outlier Detection in LR

Consider $\hat{\pi}_i$ represents the estimated values of actual probabilities, $\epsilon_i = y_i - \hat{\pi}_i$ may be defined as the deviation between $\hat{\pi}_i$ and y_i, and the ordinary residuals can be defined as follows

$$\epsilon_i = \begin{cases} 1 - \hat{\pi}_i, & y_i = 1 \\ -\hat{\pi}_i, & y_i = 0 \end{cases} \tag{15}$$

Because ϵ_i has a probability of π_i and follows the Bernoulli distribution, the errors distribution is binomial and its variance depends on the conditional mean as

$$V(Y|X) = \hat{\pi}_i(1 - \hat{\pi}_i) \tag{16}$$

where the independent variable values are unique for every observation. When it is not unique, the errors distribution is binomial and its variance can be expressed as follow

$$V(Y|X) = m_i\hat{\pi}_i(1 - \hat{\pi}_i) \tag{17}$$

where m_i is the number of observations that have the same values of X_i as observation.

As an alternative to the ordinary residuals, Hosmer and Lemeshow**,** proposed person residuals by dividing them by $\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}$. For the i^{th} covariate pattern, the Pearson residual specified is provided by

$$r_i = \frac{y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{18}$$

where *i=1, 2, …, n.* Since r_i^2 represents y_i contribution to the Pearson chi square goodness of fit Hosmer and Lemeshow statistic, it has a relationship to the Pearson chi square test statistic **[23].** Regretfully, the chi-square test statistics do not approximate the chi-square distribution in the absence of replicates when dealing with binary data. Since the $\hat{\varepsilon}_i = y_i - \hat{\pi}_i \approx (1 - h_{ii})\ y_i$, there is a serious problem when calculating the variance of $\hat{\varepsilon}_i$, hence the variance of the residual is given by

$$V(\hat{\varepsilon}_i) = (1 - h_{ii})\hat{\pi}_i\ (1 - \hat{\pi}_i). \tag{19}$$

It is clear that $V\ (\hat{\varepsilon}_i)$ lacks unit variance, and as the result, the variance of Pearson residuals is not constant.

By dividing r_i by the standard deviation, the Studentized Pearson residuals were proposed. This can be approximately expressed as $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}$, where $H = \hat{V}^{1/2}X\left(X'\hat{V}X\right)^{-1}X'\hat{V}^{1/2}$ and h_{ii} is i^{th} diagonal element of Pregibon leverage H, also known as the hat matrix. The Studentized Pearson residuals defined as

$$s_{pri} = \frac{r_i}{\sqrt{1-h_{ii}}} \tag{20}$$

and the i^{th} observation associated $|s_{pri}| > 2$ are generally identified as outlier **[49], [50].**

2.4 A Hybrid Classification Approach with Logistic Regression Analysis and RST

In this paper, a new integrated classification approach with binary logistic regression analysis and RST is proposed. Furthermore, the proposed model is called the Logistic Regression and Reduction Rough Set (LR3S). To evaluate the development of the resulting model, a comparison was made of the performance of the model produced using the original model (before data reduction), LR model after data reduction with the rough set, and modelling data that has reduced using RST and removed outlier.

According to the research objectives, this research will be conducted by considering some steps as follows

Step 1. Construct Information Table

Step 2. Construct Original Logistic Regression (OLR) Model

Step 3. Detect and remove outliers in OLR Model

Step 4. Construct RO2LR (Remove Outlier Original Logistic Regression) Model

Step 5. Detect and Remove Inconsistent Sample base on RST

Step 6. Construct LR3S (Logistic Regression Reduction Rough Set) Model

Step 7. Compare performance OLR, RO2LR and LR3S model

The overall working of this research is presented in **Figure 2.**

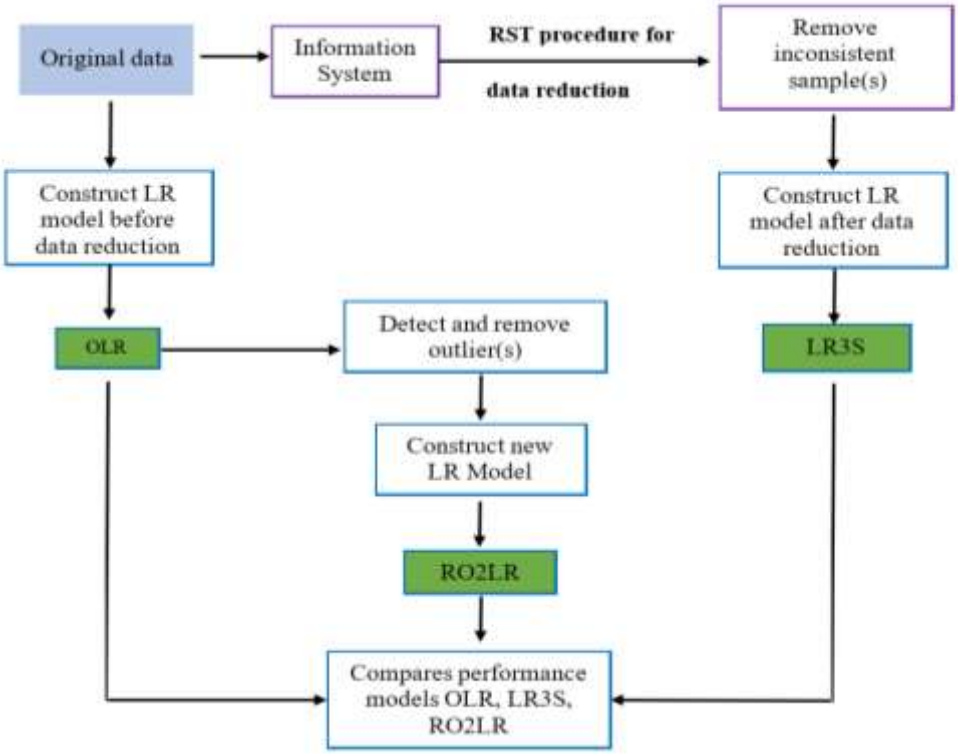

Figure 2. Research Framework

3. RESULTS AND DISCUSSION

3.1 Result

In this section, the hybrid LR and RST model classification for anemia data will be applied step by step as it is described in sub-chapter 2.4.

Step 1: Construct Information Table

The first step is to construct an information table using the utilized data *d*. In this case, the information table consists of 172 rows and 7 columns as can be seen in **Table 2.**

Table 2. Anemia Data Information Table

Pregnant Women	X1	X2	X3	X4	X5	X6	Y
P1	2	1	1	2	2	2	1
P2	1	2	2	1	2	2	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
P 172	1	2	2	2	2	2	2

Step 2: Construct Original Logistic Regression (OLR) Model

Using logistic regression to analyze the initial data (OLR Model) is the second step in the process. **Table 3** displays the outcomes of the data analysis.

Table 3. The Results of Logistic Regression Analysis on Original Data

	β	S.E.	Wald	df	Sig.	Exp(β)
Gestational age (X_1)	0.968	0.388	6.235	1	0.013	2.634
Age of pregnant women (X2)	0.455	0.523	0.755	1	0.385	1.576
Parity (X_3)	1.315	0.832	2.500	1	0.114	3.726
Employment status (X_4)	0.070	0.338	0.043	1	0.835	1.073
CED status (X_5)	1.262	0.410	9.484	1	0.002	3.531
Education level (X_6).	0.248	0.421	0.347	1	0.556	1.281
Constant	-7.363	1.834	16.115	1	0.000	0.001

Step 3: Detect and remove outliers in OLR

At this stage, outlier detection is carried out on the ORL model using the Studentized Pearson residuals as defined in **Equation (20)** criteria. There were 3 observations detected as outliers, as can be seen in **Table 4** and these observations were then discarded and new data was obtained for next modeling of LR.

Table 4. Detection Outlier in Logistic Regression Model

Case	Selected Status[a]	Observed Anemia Status	Predicted	Predicted Group	Temporary Variable Resid	ZResid	SResid
109	S	2**	.111	1	.889	2.824	2.173
137	S	2**	.151	1	.849	2.374	2.009
171	S	2**	.144	1	.856	2.439	2.036

Step 4: Construct RO2LR Model

Next, a logistic model was constructed on data that no longer contained outliers (RO2L Model). **Table 5** presents the findings of the analysis.

Table 5. Results of Logistic Regression Analysis of Anemia Data After Removing Outliers

	β	S.E.	Wald	df	Sig.	Exp(β)
Gestational age (X_1)	0.979	0.405	5.828	1	0.016	2.661
Age of pregnant women (X2)	0.715	0.544	1.730	1	0.188	2.044
Parity (X_3)	2.323	1.180	3.874	1	0.049	10.208
Employment status (X_4)	0.095	0.349	0.074	1	0.786	1.099
CED status (X_5)	1.562	0.438	12.747	1	0.000	4.770
Education level (X_6)	0.665	0.451	2.171	1	0.141	1.944
Constant	-11.238	2.620	18.400	1	0.000	0.000

Step 5: Detect and Remove Inconsistent Sample base on RST

The RST technique is then used to detect sample inconsistency. There were 48 observations that were inconsistent at this point. In order to create a new LR model, inconsistent samples with a probability of less than 0.5 are removed from the data set.

Step 6: Construct LR3S Model

The data that has been reduced in step 5 will now be subjected to data analysis using logistic regression (LR3S Model). **Table 6** displays the outcomes of the data analysis.

Table 6 Results of Logistic Regression Analysis of Anemia Data After Removing Inconsistent Samples

	β	S.E.	Wald	df	Sig.	Exp(β)
Gestational age (X_1)	2.021	0.760	7.067	1	0.008	7.546
Age of pregnant women (X2)	1.007	0.789	1.629	1	0.202	2.737
Parity (X_3)	4.502	1.469	9.392	1	0.002	90.203
Employment status (X_4)	-0.793	0.621	1.632	1	0.201	0.452
CED status (X_5)	4.487	0.855	27.563	1	0.000	88.873
Education level (X_6).	3.003	0.747	16.160	1	0.000	20.141
Constant	-24.767	4.548	29.654	1	0.000	0.000

Step 7: Compare performance OLR, RO2LR and LR3S

After the four LR models were obtained, a comparison of the performance of these models was carried out based on confusion matrix. **Table 7** displays the terms true positive (TP), true negative (TN), False positive (FP), and false negative (FN) that represent four different combinations of the predicted value and actual value on the confusion matrix. The LR model performance, including accuracy, sensitivity, specificity, and F-size, may be computed using based on confusion matrix.

Table 7. General Confusion Matrix

Predicted Values	Actual Values	
	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

The confusion matrix for the OLR model is presented in **Table 8.**

Table 8. Matrix Confusion ORL Model

Predicted Values	Actual Values		Percentage Correct
	Anemia	No Anemia	
Anemia	49	37	57.0
No Anemia	21	65	75.6
Overall Percentage			66.3

With regard to the confusion matrix, accuracy can be defined as the ratio of diagonal elements to total matrix elements. Referring to **Table 8.** Accuracy for the OLR model is

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN}\ x\ 100\% = \frac{49+65}{49+37+21+65}\ x\ 100\% = 66.3\%.$$

According to this accuracy, 66.3% of the 172 observations can be classified correctly using the OLR model

Another performance model based on confusion matrix is precision. Precision of a model can be defined as the degree of model reliability when the given prediction is “positive”. Precision for the OLR model is

$$precision = \frac{TP}{TP+FP}\ x\ 100\% = \frac{49}{49+37} = 57.0\%$$

This precision shows that in the OLR model of 86 observations suspected of being anemic, only 57% of them actually that in really experienced anemia.

Sensitivity is another confusion matrix-based performance model. The model's sensitivity is defined as its ability to recognize the data that has been classified as "true positive". Sensitivity describes the percentage of data that a model correctly classified to be labeled "positive" out of all the data that are actually labeled "positive". The precision for the OLR model can be calculated using the formula below

$$Sensitivity = \frac{TP}{TP+FN}\ x\ 100\% = \frac{49}{49+21} = 70.0\%$$

According to the computed sensitivity value, 70% of the 70 observations with anemia are predicted to have anemia in the ORL model.

The F1-score performance model is also employed in this study. F1-score illustrates the average comparison of the weighted precision and recall. When the number of false positive (FP) and false negative (FN) data in the dataset differs significantly, the F1-score should be used as the reference for classification performance. For the OLR model, the F1-score is

$$F1 - Score = 2\ \frac{precision\ x\ sensitivity}{precision+sensitivity}\ x\ 100\% = 2\ \frac{0.651\ x\ 0.7}{0.651+0.7} x\ 100\% = 67.5\%$$

Furthermore, the confusion matrix for RO2LR and LR3S models is shown in **Table 9** dan **Tabel 10**.

Table 9. Matrix Confusion RO2LR Model

Predicted Values	Actual Values		Percentage Correct
	Anemia	No Anemia	
Anemia	56	30	65.1
No Anemia	21	62	75.6
Overall Percentage			69.8

Table 10. Matrix Confusion LR3S Model

Predicted Values	Actual Values		Percentage Correct
	Anemia	No Anemia	
Anemia	51	3	94.4
No Anemia	4	66	94.3
Overall Percentage			94.4

Similar to the OLR model, each model's performance can be determined using its corresponding confusion matrix. **Table 11** displays the performance of every model that was utilized.

Table 11. Performances Models Logistic Regression

Model	Accuracy	Precision	Sensitivity	F1-Score
OLR	66.3	57.0	70	67.5
RO2LR	69.8	65.1	72.7	68.7
LR3S	94.4	94.4	92.7	93.5

Figure 3 shows a comparison of the three models' performances in terms of accuracy, precision, sensitivity, and F1-score values.

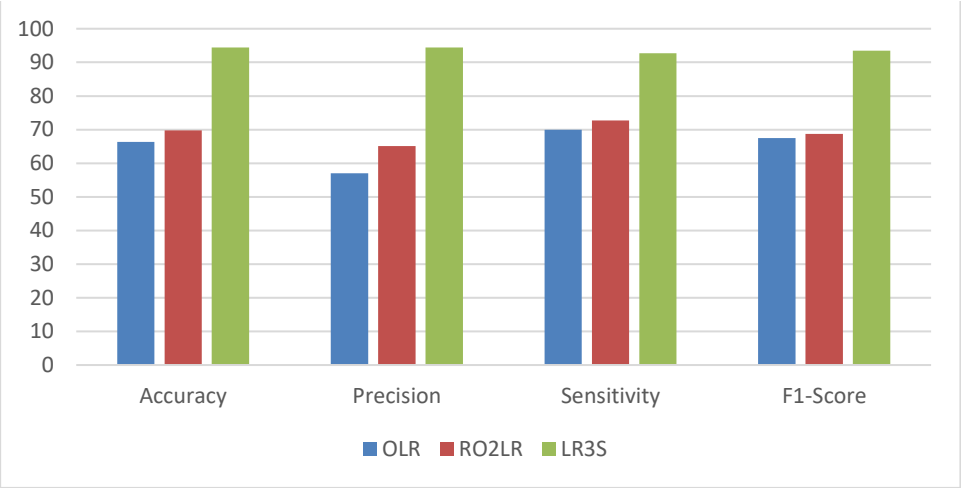

Figure 3. Comparison of the performance of the OLR, RO2LR and LR3S models

3.2 Discussion

In order to examine the factors influencing the incidence of anemia in pregnancy, a classic and modified logistic regression model were employed in this study. There are 3 models used in this research, namely OLR (Original Logistic Regression), RO2LR (Remove Outlier Original Logistic Regression) and LR3S (Logistic Regression Reduction Rough Set).

Table 11 and **Figure 3** demonstrate that the LR model performs better when data is reduced by Remove Outlier or inconsistent samples. The increase in model performance is very significant by removing inconsistent observations (LR3S model), while eliminating outliers (RO2LR model) only provides a slight improvement. In other words, in this case, the LR3S model is the most significant modification option for the LR model in improving model performance.

Next, a comparative analysis will be carried out on variables that significantly influence the incidence of anemia along with their respective odds values as presented in **Table 12**.

Table 12. List of Variables that Significantly Affect the Incidence of Anemia

Model	Significant Variable	P-value	Odd Ratio/ Exp (β)
OLR	CED status (X_5)	0.002	3.531
	Gestational age (X_1)	0.013	2.634
RO2LR	CED status (X_5)	0.000	4.770
	Gestational age (X_1)	0.016	2.661
	Parity (X_3)	0.049	10.208
LR3S	CED status (X_5)	0.000	88.873
	Education level (X_6).	0.000	20.141
	Parity (X_3)	0.002	90.203
	Gestational age (X_1)	0.008	7.546

Table 12 shows that there are 2, 3, and 4 significant variables for the OLR, RO2LR, and LR3S models, respectively. The LR model exhibits an increase in the number of significant variables upon data reduction.

For the three models used, the incidence of anemia is consistently significant influenced by two variables: gestational age and CED status, where CED status is always the variable with the highest significance. Additionally, parity is a major variable in the RO2LR and LR3S models, and educational level is a significant variable in the LR3S model, outside from CED status and gestational age. Apart from that, none of the three models utilized indicated that the age of pregnant women or employment status had a significant impact on the incidence of anemia.

Furthermore, it can be seen that for variables that significantly influence the incidence of anemia, each has an odds ratio value > 1. This means that the odd value for pregnant women who do not experience anemia will increase for pregnant women who are not at risk in terms of CED status, gestational age, parity, age of pregnant women. Apart from that, this condition also applies to mothers who are educated. This is in line

with previous studies, where it can also be said that the risk of anemia will increase in pregnant women who are at risk of CED status **[18],** gestational age **[13]–[15],** parity **[14]–[16],** and have low education **[17], [19].**

4. CONCLUSIONS

This research presents a new approach to enhance the logistic regression (LR) model's performance for anemia data by using Rough Set Theory (RST) strategy to eliminating inconsistent sample. The hybrid model namely as Logistic Regression Reduction Rough Set (LR3S). Using a number of model performance metrics, this model will be contrasted with two other models that have been often utilized in earlier studies: the RO2LR (Remove Outlier Original Logistic Model) and the OLR (Original Logistic Regression) model. Among the three models used, LR3S has the best performance based on all the indicators used. Additionally, the number of independent variables that significantly affect the incidence of anemia can be increased by using LR3S. This of course has an influence on the preparation of programs/policies to reduce the incidence of anemia in pregnancy. Using LR3S model, it is found that CED status, educational level, parity and gestational are significant variables impact on the incidence of anemia at Tegal Rejo Community Health Center, Yogyakarta Province.

Since this research is preliminary, simulation experiments must to be conducted to confirm the findings. However, according on the outcomes of experiments on this anemia data, there is a tendency that cleansing inconsistent sample using RST can improve the performance of the LR model. Additionally, it aids researchers conducting survey research including a large number of ambiguous variables or factors. If a relationship or association between variables is found to be unsatisfactory, they do not need to gather further information or conduct new surveys. With the proposed LR3R model this problem can be solved systematically and effectively. To conclude, in this proposed model researchers are advised to consider removing inconsistent samples that have a probability of less than 0.5 before doing LR modeling to get better model performance.

ACKNOWLEDGMENT

The authors would like to thank the Faculty of Mathematics and Natural Sciences, Andalas University, for supporting this research.

REFERENCES

[1] L. P. Sari, S. Sarwinanti, and S. N. Djannah, "Hubungan Status Gizi Dengan Kejadian Anemia Pada Ibu Hamil Di Puskesmas Kotagede Ii Yogyakarta," *Jurnal Cakrawala Promkes*, vol. 2, no. 1, p. 24, Mar. 2020, doi: 10.12928/promkes.v2i1.1576.

[2] Kementerian Kesehatan RI, "Profil Kesehatan Indonesia Tahun 2020," Kementerian Kesehatan RI. Accessed: Oct. 20, 2023. [Online]. Available: https://www. kemkes.go.id/id/profil-kesehatan-indonesia-2020

[3] A. D. Purwanto, "Risk Factors Correlated with Incidence of Low Birth Weight Cases," *Jurnal Berkala Epidemiologi*, vol. 4, no. 3, p. 349, Jan. 2017, doi: 10.20473/jbe.v4i3.2016.349-359.

[4] R. Anwar, K. Razzaq, and N. Noor, "Impact of Maternal Anemia on Perinatal Outcome," *Pakistan Armed Forces Medical Journal*, vol. 69, no. 2, pp. 397–402, 2019, Accessed: Nov. 17, 2023. [Online]. Available: https://pafmj.org/PAFMJ/article/view/2762

[5] N. N. Al-Hajjiah and M. A. Almkhadree, "The effect of maternal anemia on the anthropometric measurements in full-term neonates," *Asian Journal of Pharmaceutical and Clinical Research*, vol. 11, no. 4, pp. 422–424, Apr. 2018, doi: 10.22159/ajpcr.2018.v11i4.25579.

[6] M. Nair *et al.*, "Association between maternal anaemia and pregnancy outcomes: a cohort study in Assam, India", doi: 10.1136/bmjgh-2015.

[7] T. Vural, E. Toz, A. Ozcan, A. Biler, A. Ileri, and A. H. Inan, "Can anemia predict perinatal outcomes in different stages of pregnancy?," *Pak J Med Sci*, vol. 32, no. 6, pp. 1354–1359, Nov. 2016, doi: 10.12669/pjms.326.11199.

[8] F. Heydarpour *et al.*, "Maternal anemia in various trimesters and related pregnancy outcomes: Results from a large cohort study in Iran," *Iran J Pediatr*, vol. 29, no. 1, Feb. 2019, doi: 10.5812/ijp.69741.

[9] A. Patel, A. A. Prakash, P. K. Das, S. Gupta, Y. V. Pusdekar, and P. L. Hibberd, "Maternal anemia and underweight as determinants of pregnancy outcomes: Cohort study in eastern rural Maharashtra, India," *BMJ Open*, vol. 8, no. 8, Aug. 2018, doi: 10.1136/bmjopen-2018-021623.

[10] S. Parks *et al.*, "Maternal anaemia and maternal, fetal, and neonatal outcomes in a prospective cohort study in India and Pakistan," *BJOG*, vol. 126, no. 6, pp. 737–743, May 2019, doi: 10.1111/1471-0528.15585.

[11] B. A. Haider, I. Olofin, M. Wang, D. Spiegelman, M. Ezzati, and W. W. Fawzi, "Anaemia, prenatal iron use, and risk of adverse pregnancy outcomes: Systematic review and meta-analysis," *BMJ (Online)*, vol. 347, no. 7916. BMJ Publishing Group, Jul. 13, 2013. doi: 10.1136/bmj.f3443.
[12] B. Sukrat *et al.*, "Hemoglobin concentration and pregnancy outcomes: A systematic review and meta-analysis," *BioMed Research International*, vol. 2013. 2013. doi: 10.1155/2013/769057.
[13] W. F. Balcha *et al.*, "Factors associated with anemia among pregnant women attended antenatal care: a health facility-based cross-sectional study," 2023, doi: 10.1097/MS9.0000000000000608.
[14] I. Hidayati and E. N. Andyarini, "The Relationship Between The Number of Parities and Pregnancy Age with Maternal Anemia," *Journal of Health Science and Prevention*, vol. 2, no. 1, pp. 42–47, Apr. 2018, doi: 10.29080/jhsp.v2i1.113.
[15] B. Sabina Azhar, M. S. Islam, and M. R. Karim, "Prevalence of anemia and associated risk factors among pregnant women attending antenatal care in Bangladesh: A cross-sectional study," *Prim Health Care Res Dev*, vol. 22, Nov. 2021, doi: 10.1017/S146342362100061X.
[16] S. M. Davidson, G. Mangalik, and R. I. Riswandha, "Factors Affecting the Incidence of Anemia in Pregnant Women at Ampel and Gladagsari Public Health Center Boyolali Regency in 2019," *PLACENTUM Jurnal Ilmiah Kesehatan dan Aplikasinya*, vol. 10, no. 2, p. 2022, 2022.
[17] G. Obai, P. Odongo, and R. Wanyama, "Prevalence of anaemia and associated risk factors among pregnant women attending antenatal care in Gulu and Hoima Regional Hospitals in Uganda: A cross sectional study," *BMC Pregnancy Childbirth*, vol. 16, no. 1, Apr. 2016, doi: 10.1186/s12884-016-0865-4.
[18] I. Farahdiba, "Hubungan Kekurangan Energi Kronis (Kek) Dengan Kejadian Anemia Pada Ibu Hamil Primigravida Di Puskesmas Jongaya Makassar Tahun 2021," *Jurnal Kesehatan Delima Pelamonia*, vol. 5, no. 1, pp. 24–29, Sep. 2021, doi: 10.37337/jkdp.v5i1.213.
[19] G. Stephen, M. Mgongo, T. Hussein Hashim, J. Katanga, B. Stray-Pedersen, and S. E. Msuya, "Anaemia in Pregnancy: Prevalence, Risk Factors, and Adverse Perinatal Outcomes in Northern Tanzania," *Anemia*, vol. 2018, 2018, doi: 10.1155/2018/1846280.
[20] D. A. Suhardi and I. Fadila, "Penerapan Regresi Logistik Biner Untuk Mengukur Resiko Anemia Dengan Status Gizi Ibu Hamil," *Jurnal Matematika, Sains dan Teknologi* , vol. 17, no. 1, pp. 50–59, 2016.
[21] A. S. Aji, Y. Yusrawati, S. G. Malik, and N. I. Lipoeto, "Prevalence of anemia and factors associated with pregnant women in West Sumatra, Indonesia: Findings from VDPM Cohort Study," *Jurnal Gizi dan Dietetik Indonesia (Indonesian Journal of Nutrition and Dietetics)*, vol. 7, no. 3, p. 97, Jun. 2020, doi: 10.21927/ijnd.2019.7(3).97-106.
[22] M. O. Osman, T. Y. Nour, H. M. Bashir, A. K. Roble, A. M. Nur, and A. O. Abdilahi, "Risk factors for anemia among pregnant women attending the antenatal care unit in selected jigjiga public health facilities, somali region, east ethiopia 2019: Unmatched case–control study," *J Multidiscip Healthc*, vol. 13, pp. 769–777, 2020, doi: 10.2147/JMDH.S260398.
[23] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, Second. New York: John Willey and Sons, 2000.
[24] A. H. M. R. Imon and A. S. Hadi, "Identification of multiple outliers in logistic regression," in *Communications in Statistics - Theory and Methods*, Jan. 2008, pp. 1697–1709. doi: 10.1080/03610920701826161.
[25] Z. Zhang, "Residuals and regression diagnostics: Focusing on logistic regression," *Ann Transl Med*, vol. 4, no. 10, May 2016, doi: 10.21037/atm.2016.03.36.
[26] G. A. Tesema *et al.*, "Prevalence and determinants of severity levels of anemia among children aged 6-59 months in sub-Saharan Africa: A multilevel ordinal logistic regression analysis," *PLoS One*, vol. 16, no. 4 April, Apr. 2021, doi: 10.1371/journal.pone.0249978
[27] A. Yusuf *et al.*, "Factors influencing childhood anaemia in Bangladesh: A two level logistic regression analysis," *BMC Pediatr*, vol. 19, no. 1, Jun. 2019, doi: 10.1186/s12887-019-1581-9.
[28] Z. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, Oct. 1982, doi: 10.1007/BF01001956.
[29] A. Burney and Z. Abbas, "Applications of Rough Sets in Health Sciences and Disease Diagnosis," *Recent Researches in Applied Computer Science*, vol. 8, no. 3, pp. 153–161, 2015.
[30] I. Yekkala and S. Dixit, "Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection," *International Journal of Big Data and Analytics in Healthcare*, vol. 3, no. 1, pp. 1–12, Jan. 2018, doi: 10.4018/ijbdah.2018010101.
[31] I. Rahmi, Y. Wulandari, H. Yozza, and M. Syafwan, "Classification of Toddler's Nutritional Status Using The Rough Set Algorithm," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 3, pp. 1483–1494, Sep. 2023, doi: 10.30598/barekengvol17iss3pp1483-1494.
[32] K. Kim, P. Pant, and E. Y. Yamashita, "Hit-and-run crashes: Use of rough set analysis with logistic regression to capture critical attributes and determinants," *Transp Res Rec*, no. 2083, pp. 114–121, 2008, doi: 10.3141/2083-13.
[33] D. Liu, T. Li, and D. Liang, "Incorporating logistic regression to decision-theoretic rough sets for classifications," *International Journal of Approximate Reasoning*, vol. 55, no. 1 PART 2, pp. 197–210, 2014, doi: 10.1016/j.ijar.2013.02.013.
[34] B. Kan-Kilinç and Y. Yazirli, "Performance of the hybrid approach using three machine learning algorithms," *Pakistan Journal of Statistics and Operation Research*, vol. 16, no. 2, pp. 217–224, 2020, doi: 10.18187/PJSOR.V16I2.3069.
[35] K. M. Kaka-Khan, H. Mahmud, and A. A. Ali, "Rough Set-Based Feature Selection for Predicting Diabetes Using Logistic Regression with Stochastic Gradient Decent Algorithm," *UHD Journal of Science and Technology*, vol. 6, no. 2, pp. 85–93, Oct. 2022, doi: 10.21928/uhdjst.v6n2y2022.pp85-93.
[36] X. Jia, Y. Rao, L. Shang, and T. Li, "Similarity-based attribute reduction in rough set theory: a clustering perspective," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 5, pp. 1047–1060, May 2020, doi: 10.1007/s13042-019-00959-w.
[37] Rasyidah, R. Efendi, N. M. Nawi, M. M. Deris, and S. M. A. Burney, "Cleansing of inconsistent sample in linear regression model based on rough sets theory," *Systems and Soft Computing*, vol. 5, Dec. 2023, doi: 10.1016/j.sasc.2022.200046.
[38] V. Del Giudice, P. De Paola, and G. B. Cantisani, "Rough set theory for real estate appraisals: An application to directional district of Naples," *Buildings*, vol. 7, no. 1, Jul. 2017, doi: 10.3390/buildings7010012.
[39] D. R. K. N. Padmi and N. Setyawati, "Faktor-Faktor yang Mempengaruhi Kejadian Anemia pada Ibu Hamil di Puskesmas Tegalrejo Tahun 2017," Skripsi, Politeknik Kesehatan Kementerian Kesehatan, Kota Yogyakarta, 2018.

[40] Z. Pawlak, "Rough set theory and its applications to data analysis," *Cybern Syst*, vol. 29, no. 7, pp. 661–688, 1998, doi: 10.1080/019697298125470.
[41] H. Cao, "The utilization of rough set theory and data reduction based on artificial intelligence in recommendation system," *Soft Computing*, vol. 25, no. 3. Springer Science and Business Media Deutschland GmbH, pp. 2153–2164, Feb. 01, 2021. doi: 10.1007/s00500-020-05286-9.
[42] J. Komorowski, L. Polkowski, and A. Skowron, "Rough Sets: A Tutorial. In: Rough fuzzy hybridization: A new trend in decision-making," in *Rough fuzzy hybridization: A new trend in decision-making*, Singapore: Springer-Verlag, 1999, pp. 3–98.
[43] X. Li, "Attribute Selection Methods in Rough Set Theory," San Jose State University, San Jose, CA, USA, 2014. doi: 10.31979/etd.2gh8-udmy.
[44] S. M. Abbas, K. A. Alam, and S. Shamshirband, "A soft-rough set based approach for handling contextual sparsity in context-aware video recommender systems," *Mathematics*, vol. 7, no. 8, Aug. 2019, doi: 10.3390/math7080740.
[45] A. Skowron and C. Rauszer, "The Discernibility Matrices and Functions in Information Systems," in *Intelligent Decision Support*, Dordrecht: Springer Netherlands, 1992, pp. 331–362. doi: 10.1007/978-94-015-7975-9_21.
[46] R. Andersson, "Implementation of a Rough Knowledge Base System Supporting Quantitative Measures," Master's thesis, Linköping Universiity, 2004.
[47] M. Awad and R. Khanna, *Efficient Learning Machines Theories Concepts and Applications for Engineers and System Designers*. Berkeley: Apress, 2015.
[48] A. Pal, "Logistic regression: A simple primer," *Cancer Research, Statistics, and Treatment*, vol. 4, no. 3, pp. 551–554, Jul. 2021, doi: 10.4103/crst.crst_164_21.
[49] J. M. Hilbe, *Practical Guide to Logistic Regression*. CRC Press, 2015.
[50] D. Pregibon, "Logistic Regression Diagnostics," *The Annals of Statistics*, vol. 9, no. 4, Jul. 1981, doi: 10.1214/aos/1176345513.