A Heuristic Feature Selection in Logistic Regression Modeling with Newton Raphson and Gradient Descent Algorithm

Samingun Handoyo[1], Nandia Pradianti[2], Waego Hadi Nugroho[3], Yusnita Julyarni Akri[4]
Department of Statistics, Brawijaya University, Malang, Indonesia[1]
Department of EECS-IGP, National Yang Ming Chiao Tung University, Hsinchu, Taiwan[1]
Department of Statistics, Brawijaya University, Malang, Indonesia[2, 3]
Department of Midwifery, Tribuana Tunggadewi University, Malang, Indonesia[4]

Abstract*—Binary choices, such as success or failure, acceptance or rejection, high or low, heavy or light, and so on, can always be used to express decision-making. Based on the known predictor feature values, a classification model can be used to predict an unknown categorical value. The logistic regression model is a commonly used classification approach in a variety of scientific domains. The goal of this research is to create a logistic regression model with a heuristic approach for selecting input characteristics and to compare the Newton Raphson and gradient descent (GD) algorithms for estimating parameters. Among predictor traits, there were four that met the criterion for being both dependent on the target and independent of one another. Also, optional features In Malang, Indonesia, researchers used the Chi-square test to find four significant characteristics that increase the incidence of pregnant women developing preeclampsia: age (X1), parity (X2), history of hypertension (X3) and salty food consumption (X6). In the above work author proposed, the logistic regression model developed using the gradient descent approach had a lower risk of error than the logistic regression model generated using the Newton Raphson algorithm. The model with the gradient descent approach has a precision of 98.54 percent and an F1 score of 97.64 percent, while the model with the Newton Raphson algorithm has a precision of 86.34 percent and an F1 score of 72.55 percent.

***Keywords*—Classification model; feature selection; gradient descent; logistic regression; Newton Raphson**

I. Introduction

In modern statistical modeling, there is a simple point of view in developing a statistical model, namely by observing the presence of a target feature in the data set. A descriptive model is developed if there is no target feature. On the other hand, if there is a target feature, a predictive model is developed. The clustering method is the most popular descriptive model. Marji et al [1] discussed topics related to fuzzy subtractive clustering, and Handoyo et al [2] discussed the performance of the optimal clustering method with a hybrid between subtractive fuzzy and c-mean fuzzy clustering. Another type of descriptive modeling is the method used as an assessment tool to generate a ranking of objects based on their features [3]. Predictive modelling is divided into 2 types based on the measuring scale of the target feature. The regression model is built when the target feature is continuous (interval or ratio), while the classification model is built when the target feature is discrete (nominal or ordinal). In statistics, regression modeling is more emphasized to explore the causality relationship between the target and predictor features [4-5], but in the machine learning community, regression modeling is oriented to capture all existing patterns in a data set in order to obtain a model that is able to predict the unknown value of target feature with high accuracy [6]. Some examples of regression modeling for predictive purposes include Handoyo et al [7] have developed a model to predict the regional minimum wage, while Handoyo and Chen [8] have developed a model to predict daily soybean prices in Indonesia.

The application of the classification method gets more serious attention because a decision-making will be more measurable and easy to execute in the form of discrete choices, each continuous scale will also be simpler when it is transformed into a discrete scale [9]. Several researchers have compared the performance of classification models, including Widodo and Handoyo [10] compared logistic regression and Support Vector Machine, Nugroho et al [11] compared logistic regression and Learning Vector Quantization, and Handoyo et al [12] varied the threshold values to obtain the best performing logistic regression and linear discriminant models. A model involving only predictor features that have a significant contribution to the variability of the target feature is an ideal model for researchers [13-14]. Thus, feature selection is an important stage in model development. In general, the feature selection method is divided into 2 approaches, namely the wrapped and the filter approach. Wrapped approach features selection is computationally expensive because it involves the model with all of the possible feature combinations [15]. Feature selection with the filter approach method is more heuristic in nature, namely by evaluating both the dependency between predictor and target features, as well as independency among predictor features [16-17]. Chi-square test can be used for the above evaluation purposes if both features are categorical features [18].

Parameter estimation has an important role in producing the best model. In statistics, estimate parameters can be obtained by minimizing the sum of squared errors (SSE) known as the least square (LS) method [19] or by maximizing the log-likelihood function known as maximum likelihood estimation

[20-21]. The LS method is generally used for estimating parameters in linear systems, while the maximum likelihood estimation (MLE) method is used for estimating parameters in nonlinear systems. The complexity of the nonlinear model has also prompted researchers to lead using optimization methods such as Particle Swarm Optimization [22-23]. Newton Raphson algorithm works based on maximizing the likelihood function which is considered as an equation that is solved to find the equation root as the estimated parameters [24-25]. On the other hand, the gradient descent algorithm finds the estimated parameters by reducing the score function gradient and leads to be 0 which means the optimal solution has been reached [26-27].

In the field of public health, there are many problems that must be handled and controlled properly, one of which is the case of preeclampsia because it is the main cause of maternal death [28]. The immune system plays an important role in promoting the occurrence and development of preeclampsia. Wang et al [29] identified significant immune of the related genes for predicting the occurrence of preeclampsia. Women with preeclampsia are more likely to develop acute kidney injury, placental abruption, and postpartum hemorrhage syndrome before they give birth [30]. Reddy et al [31] evaluated the related application of a broader definition of hypertension and the most appropriate definition of end-organ dysfunction because there is still controversy over the definition that has been used so far.

Based on the description above, this study aims to obtain predictor features that are independent and have a significant effect on preeclampsia by using the Chi-square test, also to compare the performance of fitting the logistic regression model obtained using Newton Raphson algorithm and gradient descent by popular criteria used as classification model performance measure.

The paper consists of five sections. The remaining sections include Section 2 which described the proposed method in detail, namely the feature selection method with a filter approach using the Chi-square test, the cost function in predictive modeling, and both learning algorithms i.e. Newton Rapson and gradient descent. The presentation of empirical data, both of response and predictor features are given in Section 3, while in Section 4, the logistic regression model and its performance are discussed, both the model generated by the Newton Raphson and algorithm gradient descent. Conclusions and recommendations for further research are given in Section 5.

II. Proposed Method

A good model is simple and has high performance. One of the characteristics of the simple model is that it involves a small number of predictor features. Model parameters estimate are carried out in the training process using an optimization technique such as Maximum likelihood. When the log-likelihood function is non-linear in its parameters, a numerical iteration algorithm can be used to obtain an estimator of the model parameters. In this section, we will discuss the test of dependencies for feature selection, the score function in maximum likelihood, the Newton Raphson and gradient descent algorithm.

A. Chi Square Test for Feature Selection

In machine learning, the predictor and the response features are expected to have a relationship (dependency) while between two predictor features do not have a relationship [32]. The chi-square test is useful for evaluating the correlation between two categorical features. The chi-square (χ2) statistical test has the null hypothesis i.e. two categorical features are independent versus the alternative hypothesis i.e. two categorical features are dependent [33]. The null hypothesis is rejected when the $P\left(\chi^2_{df} > \chi^2 \text{ statistic}\right)$ is less than 0.05 (the p-value is less than 0.05) and otherwise the null hypothesis not able be rejected.

The main idea of the chi-square test is to compare the observed values in the data with the theoretically expected values and test whether the values are related to each other. The contingency table associated with both categorical features is created to support the calculation of the chi-square value. The formula of chi square statistic is the following [34]:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{i,j}-E_{i,j}\right)^2}{E_{i,j}} \tag{1}$$

Where χ^2 is Chi square statistic, $O_{i,j}$ is the observed value and $E_{i,j}$ is the expected value of two nominal variables. The Chi square statistic has a degree of freedom (df) of $(r-1)(c-1)$. The $E_{i,j}$ value can be calculated by formula:

$$E_{i,j} = \frac{\sum_{k=1}^{c} O_{i,j}\sum_{k=1}^{r} O_{k,j}}{N} \tag{2}$$

Where $\sum_{k=1}^{c} O_{i,j}$ is the sum of the i_{th}column, $\sum_{k=1}^{r} O_{k,j}$ is the sum of the k_{th}column, and N is the total instance.

When the evaluation of dependency between predictor and response feature, the expected decision is to reject the null hypothesis and the associated predictor feature is kept as the member of predictor variable. In other side, when the evaluation of dependency between 2 predictor features, the expected decision is to accept the null hypothesis that means both categorical features are kept as the member of predictor features.

B. Score Function in Maximum Likelihood Estimation

The goal of a predictive model is to make the correct prediction of the target value for a previously unseen data item. A score function is a function of the difference between the real answer $y^{(i)}$ and the predicted value $\hat{f}\left(x^{(i)};\theta\right)$ [35]. Consider the n instances hawing the response feature $y^{(i)}$ and predictor feature $x^{(i)} = [x_1, x_2 \quad \cdots \quad x_p]$ for $i = 1,2,3,\ldots,n$. Assume $y^{(i)} = \theta^{\mathrm{T}}x^{(i)} + \varepsilon^{(i)}$ is a regression model structure having as many as p unknown parameters. The $\varepsilon^{(i)}$ is a random noise (error) which is the un-modeled effect. By assuming $\varepsilon^{(i)} \sim NIID(0,\sigma^2)$, the probability density function of $\varepsilon^{(i)}$ can be stated such as the equation (3) following [36].

$$P\left(\varepsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\left(\varepsilon^{(i)}\right)^2}{2\sigma^2}\right) \tag{3}$$

The posterior probability with the unknown parameter θ is

$$P\left(y^{(i)}\middle|x^{(i)};\theta\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\left(y^{(i)}-\theta^{\mathrm{T}}x^{(i)}\right)^2}{2\sigma^2}\right) \quad (4)$$

The equation (4) means that $y^{(i)}|x^{(i)};\theta\sim N\left(\theta^{\mathrm{T}}x^{(i)},\sigma^2\right)$ and it also is called the likelihood function. The following is the likelihood function of n instances:

$$\mathcal{L}(\theta)=P(\vec{y}|X;\theta)$$

$$=\prod_{i=1}^{n}P\left(y^{(i)}\middle|x^{(i)};\theta\right)$$

$$=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\left(y^{(i)}-\theta^{\mathrm{T}}x^{(i)}\right)^2}{2\sigma^2}\right)$$

$$\ell(\theta)=n\ln\frac{1}{\sqrt{2\pi}\sigma}+\sum_{i=1}^{n}\frac{-\left(y^{(i)}-\theta^{\mathrm{T}}x^{(i)}\right)^2}{2\sigma^2} \quad (5)$$

The log likelihood is

$$\ell(\theta)=\log\mathcal{L}(\theta)\approx\ln\mathcal{L}(\theta)$$

$$=\ln\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\left(y^{(i)}-\theta^{\mathrm{T}}x^{(i)}\right)^2}{2\sigma^2}\right)$$

$$=\sum_{i=1}^{n}\left[\ln\frac{1}{\sqrt{2\pi}\sigma}+\ln\exp\left(\frac{-\left(y^{(i)}-\theta^{\mathrm{T}}x^{(i)}\right)^2}{2\sigma^2}\right)\right]$$

$$\ell(\theta)=n\ln\frac{1}{\sqrt{2\pi}\sigma}+\sum_{i=1}^{n}\frac{-\left(y^{(i)}-\theta^{\mathrm{T}}x^{(i)}\right)^2}{2\sigma^2} \quad (5)$$

Maximum Likelihood Estimation method is how to choose θ to maximize $\ell(\theta)$ in the equation (5) by the first derivative with respect to θ and set its to 0 [37]. All term in equation (5) involving the θ parameter is only the second part numerator i.e. the sum square of error which must be minimized to get the $\ell(\theta)$ maximum. In the other word, to obtain the optimum parameter θ through MLE is equivalence to minimize the equation (6) also called as the score function of regression model which is the negative of log likelihood $\ell(\theta)$.

Minimize

$$J(\theta)=\frac{1}{2}\sum_{i=1}^{n}\left(y^{(i)}-\theta^{\mathrm{T}}x^{(i)}\right)^2 \quad (6)$$

Where $J(\theta)$ is called as a loss or cost function of a regression model.

A binary classification model has the response feature of $y\epsilon\{0,1\}$. In the logistic regression case, the classifier model structure is a sigmoid function which has a primary task to separate both classes or as a boundary curve between 2 classes. Suppose the sigmoid formula of an instance is stated in the following:

$$h_\theta(x)=g(\theta^{\mathrm{T}}x)=\frac{1}{1+e^{-\theta^{\mathrm{T}}x}} \quad (7)$$

It is expected that $h_\theta(X)\epsilon[0,1]$ with $P(y=1|X;\theta)=h_\theta(X)$, and $P(y=0|X;\theta)=1-h_\theta(X)$. The posterior probability of a binary classification follows a binomial distribution as the following:

$$P(y|X;\theta)=h_\theta(X)^y\left(1-h_\theta(X)\right)^{1-y}$$

The n instances likelihood function is expressed as the following:

$$\mathcal{L}(\theta)=P(\vec{y}|X;\theta)$$

$$=\prod_{i=1}^{n}P\left(y^{(i)}\middle|X^{(i)},\theta\right)$$

$$=\prod_{i=1}^{n}h_\theta(X)^{y^{(i)}}\left(1-h_\theta(X)\right)^{1-y^{(i)}}$$

The log likelihood function for binary classification is

$$\ell(\theta)=\log\mathcal{L}(\theta)$$

$$\sum_{i=1}^{n}\left[y^{(i)}\log h_\theta\left(X^{(i)}\right)+\left(1-y^{(i)}\right)\log\left(1-h_\theta\left(X^{(i)}\right)\right)\right] \quad (8)$$

The score function of a binary classification model is the negative of $\ell(\theta)$ which has the popular name called as cross entropy loss function as the following [38].

$$J(\theta)=-\sum_{i=1}^{n}\left[y^{(i)}\log h_\theta\left(X^{(i)}\right)+\left(1-y^{(i)}\right)\log\left(1-h_\theta\left(X^{(i)}\right)\right)\right] \quad (9)$$

Machine learning model is trained by minimizing loss function to yield the estimate parameter θ.

C. Newton Raphson and Gradient Descent Algorithm

A way to obtain the estimate parameter θ is by maximizing the log likelihood function $\ell(\theta)$ through the first derivative with respect to θ and to be set 0. Because the $\ell'(\theta)$ has non linear form, the analytic (close form) solution can not be obtained. A numerical approach through the iterative method can be used to handle the problem. Newton's method was originally intended to find the roots of an equation by determining the value of the function to be 0 (to find the root of $f(\theta)=0$) [39]. Consider that the gradient (slope) of a line equation is defined as the following:

$$f'\left(\theta^{(0)}\right)=\frac{height}{base}=\frac{f\left(\theta^{(0)}\right)}{\Delta}, \text{ so } \Delta=\frac{f\left(\theta^{(0)}\right)}{f'\left(\theta^{(0)}\right)}$$

and the other hand $\Delta=\theta^{(0)}-\theta^{(1)}$ (i.e. base which is difference between 2 of x-coordinate values). For a stage t, a new x-coordinate can be expressed as the following.

$\theta^{(t+1)}:=\theta^{(t)}-\frac{f\left(\theta^{(0)}\right)}{f'\left(\theta^{(0)}\right)}$, for $f(\theta)=\ell'(\theta)$ then it is obtained

$$\theta^{(t+1)}:=\theta^{(t)}-\frac{\ell'(\theta)}{\ell''(\theta)}$$

$$\theta^{(t+1)}:=\theta^{(t)}+H^{-1}\nabla_\theta\ell(\theta) \quad (10)$$

Where Hessian H is defined as $H_{ij}=\frac{\partial^2\ell(\theta)}{\partial\theta_i\partial\theta_j}$ and $\nabla_\theta\ell(\theta)=\ell'(\theta)$. The equation (10) is the iterative formula of Newton Raphson algorithm [40]. The stopping criteria can be used either a iteration number or a threshold value desired by user. So the solution of the Newton Raphson is a value that maximize the log likelihood function $\ell(\theta)$.

In the machine learning approach, a gradient descent (GD) is an algorithm that minimizes the cost function $J(\theta)$ such as stated in equation (9). The parameters that minimize $J(\theta)$ are obtained using a search algorithm that starts with a "initial guess" value by repeatedly changing it to make $J(\theta)$ smaller until it is expected to converge to a value. Here is the formula of the GD algorithm which starts with an initial value, and is repeatedly updated [41].

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \qquad (11)$$

The GD algorithm can be implemented when the partial derivative on the right-hand side of equation (9) has been known. Suppose there is 1 instance (x, y), so the summation term in the definition of $J(\theta)$ on the equation (8) can be negligible.

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{\partial}{\partial \theta_j}\left(-\ell(\theta)\right)$$

$$\frac{\partial}{\partial \theta_j} J(\theta) = -\left(y \frac{1}{g(\theta^{\mathrm{T}} x)} - (1-y)\frac{1}{1-g(\theta^{\mathrm{T}} x)}\right)\frac{\partial}{\partial \theta_j} g(\theta^{\mathrm{T}} x)$$

$$= -\left(y \frac{1}{g(\theta^{\mathrm{T}} x)} - (1-y)\frac{1}{1-g(\theta^{\mathrm{T}} x)}\right) g(\theta^{\mathrm{T}} x)\big(1 - g(\theta^{\mathrm{T}} x)\big)\frac{\partial}{\partial \theta_j}\theta^{\mathrm{T}} x$$

$$= -\Big(y\big(1 - g(\theta^{\mathrm{T}} x)\big) - (1-y) g(\theta^{\mathrm{T}} x)\Big) x_j$$

$$= -\big(y - g(\theta^{\mathrm{T}} x)\big) x_j$$

So, it is found that the first derivative of the loss function classification is

$$\frac{\partial}{\partial \theta_j} J(\theta) = -\big(y - h_\theta(x)\big) x_j \qquad (12)$$

The gradient descent iterative formula is

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \qquad (13)$$

By substituting the equation (12) into the equation (13), It leads to the updating parameter final formula of the GD algorithm as the following:

$$\theta_j = \theta_j + \alpha \sum_{i=1}^{n}\big(y^{(i)} - h_\theta(X^{(i)})\big) X_j^{(i)} \qquad (14)$$

Where α is a learning rate determined together with a stopping criteria value such as a threshold or iteration number before the training model is started.

III. Describing Data

The data used in this study are the secondary data as many as 205 instances obtained from the Center of Child Development Studies at the Wira Husada Nusantara Midwifery Academy Malang in 2021. The data set consist of a response feature, namely preeclampsia status, and 7 predictor features, namely the factors affecting preeclampsia include age, parity, history of hypertension, pregnancy interval, household harmony, consumption of salty foods, consumption of fruits, and vegetables. The description of features in the data set is stated in Table I.

TABLE I. Class Label Distribution in the Dataset

Feature name	Class label	Label distribution
Preeclampsia (Y)	[No, Yes]	[140, 65]
Age (X1)	[No risk, Risk]	[133, 72]
Parity (X2)	[No risk, Risk]	[133, 72]
History of Hypertension (X3)	[No, Yes]	[135, 70]
Pregnancy Interval (X4)	[No risk, Risk]	[153, 52]
Household Harmony (X5)	[Yes, No]	[145, 60]
Salty Food Consumption (X6)	[No, Yes]	[116, 89]
Fruits and Vegetables Consumption (X7)	[Yes, No]	[141, 64]

All features in the data set are categorical consisting of 2 class labels, namely [No or No risk, Yes or Risk] except for X5 and X7 features which have class labels [Yes, No]. The class label in the first order is worth 0, while the class label in the second-order is worth 1. In the target feature y, the proportion of class 0 is 68% and the proportion of class 1 is 32%. The distribution of class labels on the predictor features is very similar to the distribution of class labels on the target features, except that the X6 feature has a distribution of class labels of 58% and 42% for class 0 and class 1. Imbalance class on the target feature should receive serious attention in building a classification model. Fortunately, in this data set, both the target and predictor features have a distribution of class labels that are classified as balanced.

IV. Result and Discussion

This section initially discusses feature selection by evaluating the dependencies between target and predictor features. The predictor features that have significant dependencies are preserved as the final candidate features that are evaluated for their independence. The final predictor features are selected from the final candidate features that are independent of each other. The classification model parameters associated with the final predictor feature are estimated using the Newton Rapson and Gradient descent algorithms. The performance of the two models is evaluated using several measures that are popularly used in classification.

A. Heuristic Feature Selection

Dependencies between two categorical features can be evaluated using Chi-square statistic which is calculated based on the contingency table formed from these two features. The contingency table between the target feature (Preeclampsia) and the Parity feature is presented in Table II.

The values in the cells of the contingency table are the observed values between the two categories (combination of 2 labels) derived from the two features. The observation values are compared with the expected values calculated using formula (2). Then the Chi-square statistic was calculated using formula (1). Table III presents the Chi-square statistic and associated p-value of the dependency measure between target and predictor feature.

All p-values in Table III are less than 0.05 (level of significance) which means that all predictor features have a significant dependence on the target feature. The evaluation between predictor features was based on the Chi-square statistic and the corresponding p-values which are presented in Table IV and Table V, respectively.

TABLE II. THE CONTINGENCY TABLE BETWEEN PARITY (X1) AND PREECLAMPSIA (Y)

Parity	**Preeclampsia**		
	No	***Yes***	***Total***
No Risk	108	25	133
Risk	32	40	72
Total	140	65	205

TABLE III. THE CHI SQUARE STATISTIC AND P VALUE OF DEPENDENCY BETWEEN PREDICTOR AND RESPONSE

Predictor	**Chi square**	**P value**
X1	136.59	0
X2	29.15	0
X3	166.76	0
X4	10.76	0.00104
X5	57.47	0
X6	98.53	0
X7	16.95	4.00E-05

TABLE IV. THE CHI SQUARE STATISTIC OF DEPENDENCY AMONG 2 PREDICTORS

Feature	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**
X1	205.0	1.842	2.168	10.73	37.06	2.316	13.23
X2	1.842	205.0	1.977	21.35	12.35	2.742	13.23
X3	2.168	1.977	205.0	12.01	79.30	2.386	17.47
X4	10.73	21.35	12.01	205.0	17.27	13.68	4.000
X5	37.06	12.35	79.30	17.27	205.0	34.44	13.94
X6	2.316	2.742	2.386	13.68	34.44	205.0	13.79
X7	13.23	13.23	17.47	4.000	13.94	13.79	205.0

TABLE V. THE P VALUE OF DEPENDENCY AMONG 2 PREDICTORS

Feature	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**
X1	0	0.117	0.092	0.001	0	0.082	0.000
X2	0.117	0	0.107	0	0.000	0.061	0.000
X3	0.092	0.107	0	0.001	0	0.078	0
X4	0.001	0	0.001	0	0	0.000	0.046
X5	0	0.000	0	0	0	0	0.000
X6	0.082	0.061	0.078	0.000	0	0	0.000
X7	0.000	0.000	0	0.0456	0.000	0.000	0

The independent features are obtained by using the grid search method. The first time the X1 feature is used as a search base i.e. to look for a p-value greater than 0.05 (significant level) in the X1 row, and the results show that the p-values of the X2, X3, and X6 features are greater than 0.05 that means features X1 are independent to features X2, X3, and X6. Next, feature X2 as the basis for searching and do checking whether the p-value of X3 and X6 in row X2 is greater than 0.05, lastly, feature X3 as the basis for searching and do checking whether the p-value of X6 in row X3 is greater than 0.05. The p-values in Table V which are greater than the significant level are marked with different colours. Thus the predictor features that have a significant dependence on the target feature and are also significantly independent of each other are features X1, X2, X3, and X6. These four features are finally used as predictor features of the logistic regression model to be built.

B. Model with the Newton Raphson Algorithm

The Newton Raphson algorithm is widely implemented in various statistical data analysis software, including R and SAS, which are statistical computing software that is popular among the statistician community. By setting the number of iterations = 1000 and the threshold value = 0.0001, the parameter estimators of the logistic regression model are presented in Table VI.

Based on the parameter estimator values in the second column of Table VI, the logistic regression model, namely the posterior probability of an instance as a member of class 0 is expressed in equation (15) as follows:

$$\pi(x) = \frac{\exp\begin{pmatrix}-13.1080+4.3990X_{1(1)}+5.2480X_{2(1)}+\\+7.9540X_{3(1)}+4.6360X_{6(1)}\end{pmatrix}}{1+\exp\begin{pmatrix}-13.1080+4.3990X_{1(1)}+5.2480X_{2(1)}+\\+7.9540X_{3(1)}+4.6360X_{6(1)}\end{pmatrix}} \qquad (15)$$

If the coefficient is positive, it means that it contributes to support for class 0, on the other hand, a coefficient that is negative means that it contributes to support for class 1. All of coefficients except the intercept support for class 0 where the feature X3 has the highest contribution to support for class 0.

The ability of the model to predict the instances used to build the logistic regression model is determined based on the confusion matrix, which is a matrix whose elements state the number of instances that were predicted correctly or the number of instances that were predicted incorrectly by the logistic regression model in equation (15). The Table VII presents the confusion matrix of model in equation (15).

TABLE VI. THE ESTIMATE MODEL PARAMETERS RESULTED BY THE NEWTON RAPHSON AND GRADIENT DESCENT

Feature	$\hat{\beta}_j$ **of Newton Raphson**	$\hat{\beta}_j$ **of Gradient descent**
Intercept	-13.11	-10.02
X1	4.399	3.760
X2	5.248	3.688
X3	7.954	6.046
X6	4.636	3.575

TABLE VII. THE CONFUSION MATRIX WITH NEWTON RAPHSON ALGORITHM

Actual Class	Predicted Class	
	Class 0	*Class 1*
Class 0	140	0
Class 1	28	37

Based on Table VII, it can be seen that there is no instance of the class 0 which is predicted to be wrong. However, there are the 28 instances of the 65 instances of the class 1 which are predicted to be wrong. This logistic regression classification model with Newton Raphson algorithm turned out to produce a model that was only able to detect the sensitivity of the model in that the risk of misclassifying people with preeclampsia was very high, which was above 40%. The model performance is presented in Table VIII.

TABLE VIII. PERFORMANCE OF MODEL WITH NEWTON RAPHSON AND GRADIENT DESCENT ALGORITHM

Performance	Newton Raphson	Gradient descent
Accuracy	0.8634	0.9854
Precision	0.5692	0.9538
Recall	1	1
F1 Score	0.7255	0.9764

The model's accuracy performance is 86.34% meaning that the model is able to predict instances according to their actual class of 86.34%. While the performance of the F1 score of 72.55% means that the model is able to correctly predict the occurrence of preeclampsia cases by 72.35%.

C. *Model with Gradient Descent Algorithm*

As described in section 2, the gradient descent algorithm works based on the minimization of the cost function. In this research, the stochastic gradient descent method is applied by setting the learning rate hyper-parameter value = 0.015, and the number of iterations = 1000. After the training process is complete, the results of the cost function graph in Fig. 1, and the parameter estimator in the last column of Table VI.

Fig. 1 is the learning curve of the logistic regression model shows the curve of cross-entropy loss in which starting from the 200th iteration there is only a fairly small change and the curve tends to slope after the 800th iteration. This curve also illustrates that the selection of a learning rate of 0.015 is the right value, namely in the initial iterations. , the curve does not experience a very sharp decrease (occurs when the learning rate value is too large) or the curve decreases very slowly (occurs when the learning rate is too small).

Based on the estimated parameter values which are in the last column of Table VI, the logistic regression model obtained with GD algorithm is as follows.

$$\pi(x) = \frac{\exp\begin{pmatrix}-10.0160+3.7602X_{1(1)}+3.6878X_{2(1)}+\\+6.0457X_{3(1)}+3.5749X_{6(1)}\end{pmatrix}}{1+\exp\begin{pmatrix}-10.0160+3.7602X_{1(1)}+3.6878X_{2(1)}+\\+6.0457X_{3(1)}+3.5749X_{6(1)}\end{pmatrix}} \tag{16}$$

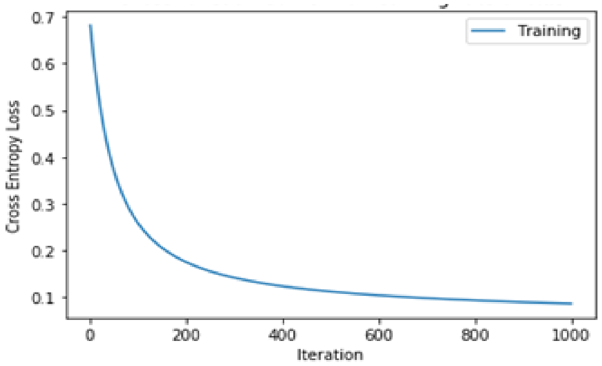

Fig. 1. The Learning Curve of the Logistic Regression Model.

In this logistic regression model, all of the coefficients except the intercept support for the class 0 where the X3 feature has the highest contribution to support for the class 0. Although the coefficients generated by the GD algorithm have a similar pattern to the coefficients generated by the Newton Rapson algorithm, the two models have different performances. The confusion matrix and performance measures of the logistic regression model with the GD algorithm are presented in Table VIII and Table IX.

TABLE IX. THE CONFUSION MATRIX WITH THE GRADIENT DESCENT ALGORITHM

Actual Class	Predicted Class	
	Class 0	Class 1
Class 0	140	0
Class 1	3	62

Table IX shows that only 3 instances of the 65 instances from the class 1 are predicted to be wrong and also all of instances from the class 0 are predicted to be correct. The Gradient descent method produces a logistic regression classification model that is able to detect the sensitivity of the model, namely the risk of misclassification of patients with preeclampsia case is very low, which is less than 5%.

The last column of Table VIII shows very clearly that the logistic regression classification model with gradient descent algorithm has superior performance than the one with Newton Raphson algorithm. It has the model's accuracy performance is 98.54% and the performance of the F1 score of 97.64%.

V. CONCLUSION

Feature selection using Chi-square test on factors that influence the incidence of pregnant women experiencing preeclampsia in Malang, Indonesia, obtained 4 significant features, namely consisting of age (X1), parity (X2), history of hypertension (X3), and consumption of salty foods (X6). The logistic regression model with the gradient descent algorithm has a lower risk of error in predicting cases of preeclampsia than the logistic regression model generated with the Newton Raphson algorithm. The model with the gradient descent algorithm has an accuracy performance of 98.54% and an F1 score of 97.64%, while the model with the Newton Raphson algorithm has an accuracy performance of 86.34% and an F1 score of 72.55%.

The dataset used in this study is too simple, which only consists of 7 predictor features, all of which are of binary categorical type. The comparison of the two algorithms will be more interesting if a dataset with a large number of predictor features is used and also involves both categorical and numeric features. Furthermore, the feature selection method used, not only involves the Chi-square test but also involves analysis of variance (F test) and also the Spearman correlation test.

References

[1] Marji, Handoyo S, Purwanto I N and Anizar M Y, "The Effect of Attribute Diversity in the Covariance Matrix on the Magnitude of the Radius Parameter in Fuzzy Subtractive Clustering" Journal of Theoretical and Applied Information Technology, vol. 96, no.12, pp. 3717-3728, 2018.

[2] Handoyo S, Widodo A, Nugroho W H and Purwanto I N, "The Implementation of a Hybrid Fuzzy Clustering Public Health Facility Data" International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no.6, pp. 3549-3554, 2019.

[3] Purwanto I N, Widodo A and Handoyo S, "System For Selection Starting Lineup of A Football Players by Using Analytical Hierarchy Process" Journal of Theoretical & Applied Information Technology, vol. 97, no. 1, pp. 19-31, 2018.

[4] Utami H N, Candra and Handoyo S, "The Effect of Self Efficacy And Hope on Occupational Health Behavior in East Java of Indonesia" International Journal of Scientific & Technology Research, vol. 9, no.2, pp. 3571-3575, 2020.

[5] Kusdarwati H and Handoyo S, "Modeling Treshold Liner in Transfer Function to Overcome Non Normality of the Errors" IOP Conf. Series on The 9th Basic Science International Conferences, vol. 546, no. 5, pp. 052039, 2019.

[6] Kusdarwati H and Handoyo S, "System for Prediction of Non Stationary Time Series based on the Wavelet Radial Bases Function Neural Network Model" Int J Elec & Comp Eng (IJECE), vol. 8, no. 4, pp. 2327-2337, 2018.

[7] Handoyo S, Marji, Purwanto I N and Jie F, "The Fuzzy Inference System with Rule Bases Generated by using the Fuzzy C-Means to Predict Regional Minimum Wage in Indonesia" International J. of Opers. and Quant. Management (IJOQM), vol. 24, no. 4, pp. 277-292, 2018.

[8] Handoyo S and Chen Y-P, "The Developing of Fuzzy System for Multiple Time Series Forecasting with Generated Rule Bases and Optimized Consequence Part" International Journal of Engineering Trends and Technology, vol. 68, no. 12, pp. 118-122, 2020.

[9] Handoyo S and Kusdarwati H, "Implementation of Fuzzy Inference System for Classification of Dengue Fever on the villages in Malang" IOP Conf. Series on The 9th Basic Science International Conferences, vol. 546, no. 5, pp. 052026, 2019.

[10] Widodo A and Handoyo S, "The Classification Performance Using Logistic Regression And Support Vector Machine (Svm)" Journal of Theoretical & Applied Information Technology, vol. 95, no. 19, pp. 5184-5193, 2017.

[11] Nugroho W H, Handoyo S and Akri Y J, "An Influence of Measurement Scale of Predictor Variable on Logistic Regression Modeling and Learning Vector Quantization Modeling for Object Classification" Int J Elec & Comp Eng (IJECE), vol. 8, no. 1, pp. 333-343, 2018.

[12] Handoyo S, Chen Y-P, Irianto G and Widodo A, "The Varying Threshold Values of Logistic Regression and Linear Discriminant for Classifying Fraudulent Firm" Mathematics and Statistics, vol. 9, no. 2, pp. 135 – 143, 2021.

[13] Stalin S, Roy V, Shukla P K, Zaguia A, Khan M M, Shukla P K, & Jain A, "A machine learning-based big EEG data artifact detection and wavelet-based removal: an empirical approach" Mathematical Problems in Engineering, vol. 2021, pp. 2942808, 2021.

[14] Shukla, P. K., Roy, V., Shukla, P. K., Chaturvedi, A. K., Saxena, A. K., Maheshwari, M., & Pal, P. R. "An Advanced EEG Motion Artifacts Eradication Algorithm" The Computer Journal, 2021.

[15] Zhao Z, Li J, Fan C, Du Y, Zhou M, Zhang X, Zhao H, "Robust phase unwrapping algorithm for interferometric applications based on Zernike polynomial fitting and Wrapped Kalman Filter" Optics and Lasers in Engineering, vol. 152, pp. 106952, 2022.

[16] Alkan Ö, Abar H, "Determination of factors influencing tobacco consumption in Turkey using categorical data analyses" Archives of environmental & occupational health, vol. 75, no.1, pp. 27-35, 2020.

[17] Spiga O, Cicaloni V, Fiorini C, Trezza A, Visibelli A, Millucci L, Santucci A, "Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease" Orphanet journal of rare diseases, vol. 15, no. 1, pp. 1-10, 2020.

[18] Thaseen I S, Kumar C, Ahmad A, "Integrated intrusion detection model using chi-square feature selection and ensemble of classifiers" Arabian Journal for Science and Engineering, vol. 44, no. 4, pp. 3357-3368, 2019.

[19] Handoyo S and Marji, "The Fuzzy Inference System with Least Square Optimization for Time Series Forecasting" Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), vol. 7, no. 3, pp. 1015-1026, 2018.

[20] Lio W and Liu B, "Uncertain maximum likelihood estimation with application to uncertain regression analysis" Soft Computing, vol. 24, no. 13, pp. 9351-9360, 2020.

[21] Orellana R, Bittner G, Carvajal R, Agüero J C, "Maximum Likelihood estimation for non-minimum-phase noise transfer function with Gaussian mixture noise distribution" Automatica, vol. 135, pp. 109937, 2021.

[22] Handoyo S, Efendi A, Jie F, Widodo A, "Implementation of particle swarm optimization (PSO) algorithm for estimating parameter of arma model via maximum likelihood method" Far East Journal of Mathematical Sciences, vol. 102, no. 7, pp. 1337-1363, 2017.

[23] Efendi A, Handoyo S, Prasojo A P S, and Marji, "The Implementation of The Optimal Rule Bases Generated By Hybrid Fuzzy C-Mean And Particle Swarm Optimization" Journal of Theoretical & Applied Information Technology, vol. 97, no. 16, pp. 4453-4453, 2019.

[24] Liu Z, Zhang X, Su M, Sun Y, Han H, Wang P, "Convergence analysis of newton-raphson method in feasible power-flow for DC network" IEEE Transactions on Power Systems, vol. 35, no. 5, pp. 4100-4103, 2020.

[25] Feng Z, Ma N, Li W, Narasaki K, Lu F, "Efficient analysis of welding thermal conduction using the Newton–Raphson method, implicit method, and their combination" The International Journal of Advanced Manufacturing Technology, vol. 111, no. 7, pp. 1929-1940, 2020.

[26] Farajtabar M, Azizan N, Mott A, Li A, "Orthogonal gradient descent for continual learning" International Conference on Artificial Intelligence and Statistics, pp. 3762-3773, 2020.

[27] Fearnley J, Goldberg P W, Hollender A, Savani R, "The complexity of gradient descent: CLS= PPAD∩ PLS" Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, pp. 46-59, 2021.

[28] Pribadi A, "Zero mother mortality preeclampsia program: Opportunity for a rapid acceleration in the decline of maternal mortality rate in Indonesia" International Journal of Women's Health and Reproduction Sciences, vol. 9, no. 3, pp. 160-163, 2021.

[29] Wang Y, Li Z, Song G and Wang J, "Potential of Immune-Related Genes as Biomarkers for Diagnosis and Subtype Classification of Preeclampsia" Frontiers in genetics, vol. 11, pp. 1481, 2020.

[30] Novotny S, Lee-Plenty N, Wallace K, Kassahun-Yimer W, Jayaram A, Bofill J A and Martin J N, "Acute kidney injury associated with preeclampsia or hemolysis, elevated liver enzymes and low platelets syndrome Pregnancy hypertension, vol. 19, pp. 94-99, 2020.

[31] Reddy M, Fenn S, Rolnik D L, Mol B W, da Silva Costa F, Wallace E M and Palmer K R, "The impact of the definition of preeclampsia on disease diagnosis and outcomes: a retrospective cohort study" American Journal of Obstetrics and Gynecology, vol. 224, no. 2, pp. 217-e1, 2021.

[32] Shukla P K, Bhatele M, Chaturvedi A K, Sharma P, Rizvi M A, Pathak Y, "A Novel Machine Learning Model to Predict the Staying Time of International Migrants" International Journal on Artificial Intelligence Tools, vol. 30, no. 02, pp. 2150002, 2021.

[33] Franke T M, Ho T, Christie C A, "The chi-square test: Often used and more often misinterpreted" American Journal of Evaluation, vol. 33, no. 3, pp. 448-458, 2012.

[34] Adekpedjou A, De Mel W A, Zamba G K, “Data dependent cells chi-square test with recurrent events” Scandinavian Journal of Statistics, vol. 42, no. 4, pp. 1045-1064, 2015.

[35] Du S, Lee J, Li H, Wang L and Zhai X, “Gradient descent finds global minima of deep neural networks” In International Conference on Machine Learning, pp. 1675-1685, 2019.

[36] Luo Z, Li W, Gan Y, Mendu K and Shah S P, “Maximum likelihood estimation for nanoindentation on sodium aluminosilicate hydrate gel of geopolymer under different silica modulus and curing conditions” Composites Part B: Engineering, vol. 198, pp. 108185, 2020.

[37] Liu Y, Liu B, “Estimating unknown parameters in uncertain differential equation by maximum likelihood estimation” Soft Computing, pp. 1-8, 2022.

[38] Tripathi D, Edla D R, Bablani A, Shukla A K, Reddy B R, “Experimental analysis of machine learning methods for credit score classification” Progress in Artificial Intelligence, vol. 10, no. 3, pp. 217-243, 2021.

[39] Ypma T J, “Historical development of the Newton–Raphson method” SIAM review, vol. 3, no. 4, pp. 531-551, 1995.

[40] Gnetchejo P J, Essiane S N, Dadjé A, Ele P, “A combination of Newton-Raphson method and heuristics algorithms for parameter estimation in photovoltaic modules” Heliyon, vol. 7, no. 4, pp. e06673, 2021.

[41] Fehrman B, Gess B, Jentzen A, “Convergence rates for the stochastic gradient descent method for non-convex objective functions” Journal of Machine Learning Research, vol. *21*, pp. 136, 2020.