第 49 卷 第 4 期
2022 年 4 月

湖南大学学报（自然科学版）
Journal of Hunan University (Natural Sciences)

Vol. 49 No. 4
April 2022

Open Access Article https://doi.org/10.55463/issn.1674-2974.49.4.17

Building Multiclass Classification Model of Logistic Regression and Decision Tree Using the Chi-Square Test for Variable Selection Method

Waego H. Nugroho[1], Samingun Handoyo[1,2], Yusnita J. Akri[3], Agus D. Sulistyono[4]

[1] Department of Statistics, Brawijaya University, Malang, CO 65145, Indonesia

[2] Department of Electrical Engineering and Computer Science-IGP, National Yang Ming Chiao Tung University, Department of Electrical Engineering and Computer Science-IGP, National Chiao Tung University, Hsinchu, CO 30010, Taiwan

[3] Department of Midwifery, Tribhuwana Tunggadewi University, Malang, CO 65145, Indonesia

[4] Department of Fisheries and Marine Socio-Economics, Brawijaya University, Malang, CO 65145, Indonesia

Abstract: The growth and development of children under five (toddlers) affect their health conditions. Each region uniquely identifies the main factors influencing the toddler's health condition. The status of toddlers is generally categorized into two classes, namely normal and abnormal, so it is often found that the condition of toddler status is in the form of multi-response variables. Combining the two binary classes' response variables will form a multiclass response variable requiring different model development techniques and performance measurements. This study aims to determine the main factors that affect toddlers' health conditions in Malang, Indonesia, build multiclass logistic regression and decision tree classification models, and measure the model's performance. The Chi-square test selected predictor features as the input of multiclass logistic regression and decision tree models. From the feature selection, four main factors influence the status of toddlers' health conditions in Malang: the mother's history of diabetes before pregnancy, the father's blood pressure, psychological condition, and drinking water quality. The decision tree model performs better than the logistic regression model on the various performance measures used.

Keywords: Chi-square test, decision tree, logistic regression, multiclass classification, variable selection.

使用变量选择方法的卡方检验建立逻辑回归和决策树的多类分类模型

摘要：五岁以下儿童（幼儿）的生长发育会影响他们的健康状况。每个地区都唯一确定影响幼儿健康状况的主要因素。幼儿的状态一般分为正常和异常两类，所以经常发现幼儿状态的状态是多反应变量的形式。结合两个二元类的响应变量将形成一个多类响应变量，需要不同的模型开发技术和性能测量。本研究旨在确定影响印度尼西亚玛琅幼儿健康状况的主要因素，建立多类逻辑回归和决策树分类模型，并衡量模型的性能。卡方检验选择预测特征作为多类逻辑回归和决策树模型的输入。从特征选择来看，影响玛琅幼儿健康状况的主要有四个因素：母亲孕前糖尿病史、父亲血压、心理状况和饮用水质量。决策树模型在所使用的各种性能指标上的表现优于逻辑回归模型。

关键词: 卡方检验、决策树、逻辑回归、多类分类、变量选择。

Received: February 12, 2022 / Revised: March 1, 2022 / Accepted: March 5, 2022 / Published: April 30, 2022

About the authors: Waego H. Nugroho, Department of Statistics, Brawijaya University, Malang, Indonesia; Samingun Handoyo, Department of Statistics, Brawijaya University, Malang, Indonesia; Department of Electrical Engineering and Computer Science-IGP, National Yang Ming Chiao Tung University, Department of Electrical Engineering and Computer Science-IGP, National Chiao Tung University, Hsinchu, Taiwan; Yusnita J. Akri, Department of Midwifery, Tribhuwana Tunggadewi University, Malang, Indonesia; Agus D. Sulistyono, Department of Fisheries and Marine Socio-Economics, Brawijaya University, Malang, Indonesia

1. Introduction

Low birth weight (LBW) is considered a sensitive index of the health of a nation. LBW is caused by the multifactorial interaction of socio-demographic and biological factors. The parents' economic condition and the mother's condition during pregnancy also affect LBW [1]. Stunting or short toddlers is one of the LBW effects. A stunting baby is caused by impaired growth and development in infants due to continuous malnutrition for 1000 days of its life [2]. The complexity of the factors that affect infant growth and development conditions requires data exploration methods to select the main factors that influence the response variables, as done by Romero and Ventura [3]. Suppose both the response and predictor variables are nominal or ordinal. In that case, the dependencies between the predictor and response variables and the independence between the predictor variables can be evaluated using the Chi-square test [4].

Learning algorithm in machine learning is classified according to the existence of the target variable. Marji et al. [5] investigated the effect of the radius magnitude on fuzzy subtractive clustering, and Purwanto et al. [6] selected the starting lineup of a football club by hierarchical process analyses. If a dataset contains a response variable, then a learning algorithm that seeks to obtain a function mapping accurately the predictor variable to the response variable is known as supervised learning. The model produced by supervised learning with a response variable having an interval or ratio scale is called a regression model. Some examples of regression modeling in machine learning are the wavelet neural network in the time series regression by Kusdarwati and Handoyo [7], the fuzzy logic regression by Handoyo and Marji [8] to predict the USD -> IDR exchange rate, and the fuzzy inference system for multiple time series forecasting by Handoyo and Chen [9].

The learning algorithm associated with the dataset having the discrete (categorical) response variable will produce a classification model. Examples of classification modeling in machine learning modeling include the classification of faults in the vehicle power transmission system by Gong et al. [10]. Handoyo et al. [11] applied the fuzzy system to predict the regional minimum wages. The application of a classification model has spread widely in various aspects of life. The logistic regression [12] and decision tree classification models [13] have satisfactory performance. Both models are very popular because they are also easy to understand.

Response variables with multinomial class labels, namely more than 2 class labels, can be handled using a multiclass classification model [14]. Applications of multiclass logistic regression models include Wang et al.'s [15] diagnosing transformer fault and classification of newspaper articles by Sebők and Kacsuk [16]. On the other hand, several researchers used the decision tree method to handle the classification of various objects having multiclass labels, including Walega et al. [17], Mena and Bolte [18], Rojarath and Songpan [19]. Furthermore, a multiclass label formed from a combination of 2 nominal response variables (LBW and stunting) on the health status of a child under five years old has prompted researchers to develop a multiclass classification model involving heuristic feature selection. Thus, the study aims to obtain predictor features independent of each other and having a significant dependence on the response variable and build multiclass logistic regression and decision tree models to classify the health status of a child under five years old.

2. Literature Review

This section will discuss the methods used in this research theoretically, including the Chi-square dependency test, multiclass logistic regression, and decision tree classification.

2.1. Chi-Square Test for Dependency between Two Categorical Features

Dependency between 2 categorical variables can be evaluated using the Chi-square test, which has the null hypothesis, i.e., the two variables are not dependent, and the alternative hypothesis, i.e., the two variables depend on each other [20]. The formula for Chi-square statistic is the following:

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

where χ^2 is the 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 equal to $(r-1)(c-1)$. The $E_{i,j}$ value can be calculated by the formula:

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

where $\sum_{i=1}^{c} O_{i,j}$ 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.

The null hypothesis is rejected if $\chi^2 \geq \chi^{\alpha}_{(r-1)(c-1)}$ or $p_{-value} \leq \alpha$, where α is a significant level which is usually equal to 0.05. When a hypothesis test aims to evaluate dependency between the predictor and target variables, a decision to reject the null hypothesis is the desired result, which means the related predictor variable is kept as a member of the predictor variable set. When a hypothesis test aims to evaluate dependency between 2 predictor variables, the rejecting null hypothesis means that one among two variables must be dropped from the predictor variable set [21].

2.2. Multiclass Logistic Regression

The basis for building a classification model is determining the classification class boundary, separating the different class instances [22]. The number of separating boundaries depends on the many classes of instances that will be separated. In the case of binary classes, it just needs one decision boundary. If the number of classes is greater than two, it is called the multiclass classification, and the number of decision boundaries is k-1, where k represents the many class instances separated.

The logistic regression classification model for binary classes is called the sigmoid function and defined in the machine learning approach as follows [23]:

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

where $\sigma(a) = \mathrm{y}(w^{\mathrm{T}}\phi)$ is the associated probability of the class C_1 and $p(C_2|\phi) = 1 - p(C_1|\phi)$ is the class C_2 probability. The multiclass cases' logistic regression model is called softmax function, defined as the following:

$$p(C_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \quad (3)$$

where the activation a_k is given by $a_k = w_k^{\mathrm{T}}\phi$, and $p(C_k|\phi)$ is the probability associated with the class C_k. The parameter model can be obtained by defining a maximum likelihood function.

Suppose the training data pairs of $\{\phi_n, t_n\}$ where $t_n \epsilon \{0,0,0,1,..0\}$ is a hot vector having k dimension, $\phi_n = \phi(\mathrm{X_n})$ for $1, \ldots, N$ is the N instances having the X_n features on each instance. The likelihood function of multiclass logistic regression is defined as the following [24]:

$$p(\mathrm{T}|\mathrm{w}_1, \ldots \mathrm{w}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(C_k|\phi_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

where $y_{nk} = y_n(\phi_n)$ is a softmax function stated in Equation (3).

The optimal multiclass logistic regression model can be obtained by a numerical iteration method such as the Newton-Raphson method maximizing the log-likelihood function by considering the first partial derivative of the parameters *w* as a nonlinear equation whose roots are solved through numerical iteration [25]. The Newton-Raphson method is started by determining an initial solution which is continuously updated using Equation (4):

$$\mathrm{w}^{(new)} = \mathrm{w}^{(old)} - \mathrm{H}^{-1}\nabla\mathcal{L}(\mathrm{w}) \quad (4)$$

where

$$\nabla_{\mathrm{w}_j}\mathcal{L}(\mathrm{w}_1, \ldots \mathrm{w}_K) = \sum_{n=1}^{N}\frac{\partial E}{\partial a_{nj}}\nabla_{\mathrm{w}_j} a_{nj} = \sum_{n=1}^{N}(y_{nj} - t_{nj})\phi_n$$

It is called the gradient vector,

$$\nabla_{\mathrm{w}_k}\nabla_{\mathrm{w}_j}\mathcal{L}(\mathrm{w}_1, \ldots \mathrm{w}_K) = \sum_{n=1}^{N} y_{nk}(I_{kj} - y_{nj})\phi_n\phi_n^{\mathrm{T}}$$

It is called the Hessian matrix, and

$$\mathcal{L}(\mathrm{w}_1, \ldots \mathrm{w}_K) = \ln p(\mathrm{T}|\mathrm{w}_1, \ldots \mathrm{w}_K) = -\prod_{n=1}^{N}\prod_{k=1}^{K} t_{nk} \ln y_{nk}$$

which is the log-likelihood function. The iteration (parameter updating) is stopped when the threshold value or iteration number as a stopping condition has been fulfilled.

Updating parameters in the Newton-Raphson method involves the inverse of the Hessian matrix. A convergent solution cannot be reached if the Hessian matrix is singular or close to the singular matrix. Therefore, in machine learning to get the parameters W that maximize the log-likelihood function, the optimization problem is changed to a minimization problem by defining a cost (score) function, namely the negative log-likelihood function. The iteration method to get the parameters W that minimizes the cost function is known as the gradient descent method [26]. The formula for updating parameters W of the gradient descent method is given in Equation (5):

$$\mathrm{w}^{(new)} = \mathrm{w}^{(old)} - \mathrm{б}\nabla\ell(\mathrm{w}) \quad (5)$$

where $\ell(\mathrm{w}) = -E(\mathrm{w}_1, \ldots \mathrm{w}_K)$, and б is the learning rate, where it must be set wisely. A too large α value will lead to an overshoot against the optimal solution, and a too small α causes the updated value which will vanish to zero, and the optimal solution will never be found.

2.3. Decision Tree Classification

Tree models (including decision trees) are a heuristic approach with the basic principle of repeatedly partitioning the input space into subsets to maximize a score of overall class purity until some stopping criterion is met. Predicting the class label is done by traversing the tree down to a leaf, which indicates the predicted class label. Tree models can handle mixed variables and have high comprehensibility [27]. A Decision Tree is a tree-structured plan of a set of attributes to test to predict the output [28]. The top-down tree construction is carried out by partitioning the examples recursively by choosing one attribute each time. An internal node is a test on an attribute (feature). A branch represents the test outcome (e.g., color = red). A leaf node represents a class label or class label distribution. At each node, one attribute is chosen to split training examples into as distinct subsets as possible. A new example is classified by following a matching path to a leaf node.

An attribute to split on at an internal node is determined using a score function that measures a degree of purity on each feature and choosing the feature producing the "purest" nodes. The 0-1 score function is defined as follows [29]:

$$S(y) = \sum_{S_i}^{|D|}\sum_{x_j} 1(y_j \neq \hat{y}_j), where\ S = \{S_1, S_2, S_3, \ldots, S_k\}, i.e.\ k\ subsets$$

The illustration of constructing a decision tree is presented in Fig. 1. Suppose the number of the examples is |D| = 100, and each example has $X_i, for\ i = 1,2,3, \ldots, p$ features. The first step is to pick an attribute and the associated value that optimizes some criterion, such as information gain.

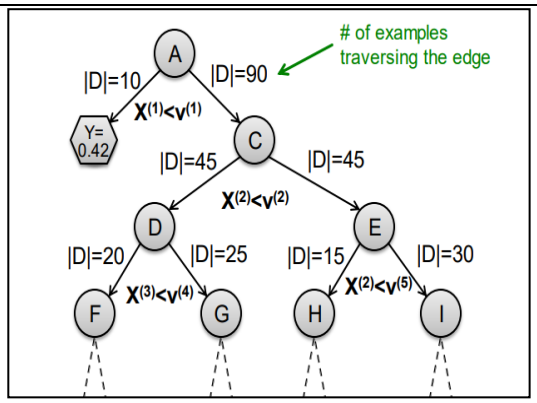

Fig. 1 The decision tree concept

Information gain (IG) is calculated using entropy**,** the smallest possible number of bits, on average per symbol, needed to transmit a stream of symbols drawn from ***X***'s distribution. The entropy of data set D, H(D) containing C classes is defined as

$H(D) = -\sum_{i=1}^{C} p_i log_2(p_i)$

A partition based on attribute A and producing subsets $D_1 \sim D_k$ and the entropy after splitting on A, H(D, A) are defined as

$H(D,A) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} H(D_i)$

Information gain (IG) is defined as

$IG(D,A) = H(D) - H(D,A)$

IG is computed for all available attributes, and the one with the highest IG is picked.

3. Data Description and Analysis Steps

The dataset used in this research was collected by the Department of Midwifery at Tribuana Tungga Dewi University, Malang, Indonesia, through a survey of some medical clinics in Malang. The dataset consists of 900 instances with 12 predictor features and one response feature. All predictor features have the domain of categorical binary values such as Yes or No, Normal or Abnormal, Good or Bad, Enough or Not Enough. In addition, the target feature (baby's health condition) has an ordinal scale that consists of four classes: Class 0 normal in weight and height, Class 1 normal in weight but abnormal in height, Class 2 abnormal in weight but normal in height, and Class 3 abnormal in weight and height. The feature names and their label distribution are presented in Table 1.

Table 1 The dataset features and their properties

No.	Symbol	Feature name	Categorical class	Distribution class
0	X1	Maternal blood pressure before pregnancy	[0 1]	[629 271]
1	X2	Diabetes maternal history before pregnancy	[0 1]	[770 130]
2	X3	Mother's psychological condition	[0 1]	[855 45]
3	X4	Father's blood pressure	[0 1]	[742 158]
4	X5	Paternal history of diabetes	[0 1]	[728 172]
5	X6	Father's congenital disease	[0 1]	[695 205]
6	X7	Family welfare education	[0 1]	[676 224]
7	X8	Father's psychological condition	[0 1]	[877 23]
8	X9	Family income	[0 1]	[688 212]
9	X10	Drinking water quality	[0 1]	[691 209]
10	X11	House floor condition	[0 1]	[734 166]
11	X12	House sanitation condition	[0 1]	[744 156]
12	Y	Baby's health condition	[0 1 2 3]	[664 3 80 153]

Class 0 in the predictor features dominates Class 1 up to 80 % of the instances. Class 0 represents the status quo situation, and Class 1 is the counterpart condition. The target feature plays a critical role in building a machine learning model. Because its scale is a category, the yielded model is called the classification model.

The sequential summary of the analysis to develop the logistic regression and decision tree is as follows:

a. Evaluating the dependency between predictor and target features and dropping the features without dependency on the target feature as the predictor feature set element;

b. Evaluating independency among the remaining predictor features and choosing the subset of those independent of each other as the final predictor features;

c. Dividing the dataset with final predictor features into the training and testing part;

d. Using the training part to develop a multiclass logistic regression model and evaluate its performance in the testing part;

e. Using the training part to develop a decision tree model and evaluate its performance in the testing part.

4. Results

The simple model is preferred over the complex one because it will be easier to explain and understand the system modeled. One of the simple model properties is that it has fewer predictor features. The feature selection method is used to obtain them. The feature selection with a data-centric approach does not involve the classifier model when a feature is decided as the selected feature for building a classification model.

4.1. Data-Centric Feature Selection

As an input of the machine learning model, the selection feature plays a critical role in predictive

modeling. It expects that all input features have significant dependence on the target feature, and features in the predictor feature set are statistically independent of each other. When the predictor and target features are categorically scaled, their dependency can be evaluated using the Chi-square test. Table 2 presents the Chi-square statistic and its corresponding p-value. The Chi-square statistic is calculated to show the dependency between each predictor feature and the target feature.

Table 2 The Chi-square statistic and the associating p-value to test dependency between the predictor and target features

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
Chi-Square	290.78	73.13	24.61	356.16	278.7	489.37	436.57	45.33	234.97	340.62	352.78	366.64
P-value	0.0	0.0	2.00E-05	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The decision result of the hypothesis test is that all predictor features are significantly dependent on the target feature, shown by all p-values being less than 0.01, which means the null hypothesis is rejected. The consequence of the decision is to pick up all of the predictor features as candidates for the input features. For determining that predictor features are independent of each other, the Chi-square statistic is calculated for the pair combination between 2 predictor features. There are 12 predictor features, so it calculates 144 Chi-square statistics and its associating p-value. Because the dependency between Features A and B is the same as between Features B and A, the Chi-square values and their corresponding p-values are given in Table 3 as follows:

Table 3 The Chi-square and p-values among of 2 predictor features

		X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
X1	**Chi-Square**	900.01	329.82	23.21	142.13	145.24	152.47	132.72	0.91	76.74	106.86	87.82	117.86
	P-Value	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3406	0.0	0.0	0.0	0.0
X2	**Chi-Square**	329.82	899.89	3.83	30.56	26.04	19.22	20.49	1.02	15.08	4.85	8.64	26.81
	P-Value	0.0	0.0	0.0503	0.0	0.0	1.00E-05	1.00E-05	0.3126	0.0001	0.0276	0.0033	0.0
X3	**Chi-Square**	23.21	3.83	900	32.13	56.95	12.64	7.61	0.68	43.98	60.93	9.22	9.64
	P-Value	0.0	0.0503	0.0	0.0	0.0	0.0004	0.0058	0.4101	0.0	0.0	0.0024	0.0019
X4	**Chi-Square**	142.13	30.56	32.13	899.92	340.43	258.83	301.48	0.28	114.3	151.46	189.04	190.85
	P-Value	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5941	0.0	0.0	0.0	0.0
X5	**Chi-Square**	145.24	26.04	56.95	340.43	900.03	187.95	217.36	3.33	72.03	150.3	146.03	125.59
	P-Value	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0679	0.0	0.0	0.0	0.0
X6	**Chi-Square**	152.47	19.22	12.64	258.83	187.95	900.09	423.72	5.75	151.47	217.7	206.89	295.66
	P-Value	0.0	1.00E-05	0.0004	0.0	0.0	0.0	0.0	0.0165	0.0	0.0	0.0	0.0

Based on Table 3, when using the significant level of 0.01, the null hypothesis cannot be rejected on the dependency between X2 and X3, X2 and X8, X2 and X10, which are on the marked raw in Table 3. The decision cannot reject the null hypothesis meaning that two predictor features are independent of each other. The result leads to picking up X2, X3, X8, and X10 as the final input features. Furthermore, the final input and target feature pairs will be used to build the logistic regression and decision tree classifiers.

4.2. Dividing the Dataset with Selected Features into the Training and Testing Parts

The dataset obtained in the feature selection that consists of 4 predictor features (X2, X2, X8, and X10) and the target feature Y is divided randomly into the training and testing parts where the training set proportion is 70%, and the remaining 30% is the testing set. The class distribution in the training and testing sets is presented in Table 4.

Table 4 The target feature distribution class label after the splitting data

Class label	Training distribution	Testing distribution
Class 0	464	200
Class 1	2	1
Class 2	56	24
Class 3	108	45

The target feature class distribution seems to be imbalanced. Class 0 dominates with around 70% of members, only 0.33% (3) instances were the members of Class 1, and around 9% (80) instances were the members of Class 2. The imbalance of class distribution will lead to some problems, but this research does not resolve the imbalanced class problem.

4.3. Modeling a Multiclass Logistic Regression Classifier

The building model in machine learning uses only the training data part. A training process is important because it determines the model's quality. The initializing parameters (weights) and tuning of parameter learning (learning rate, iteration number, and tolerated error) hold a critical role in determining model structure, which will be a classifier model when the estimated parameters have been obtained. The logistic regression model is trained using an iteratively numerical optimization, including the gradient descent algorithm. Initializing the parameter model randomly, setting the iteration number (1000), and tuning the learning rate by trial and error (obtained of 0.1) are the steps that cannot be done carelessly. The training process is to find the optimal parameters model minimizing the cross-entropy loss function. With the

gradient descent algorithm, the learning curve is presented in Fig. 2 as follows:

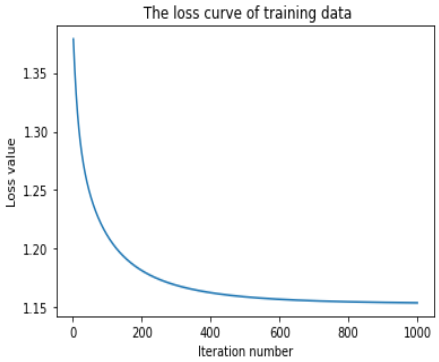

Fig. 2 The learning curve with a learning rate of 0.1

The trained model is obtained when the score (loss) function has a flat value, which means the learning process has converged, and the optimal estimated parameters (weights) are obtained. Table 5 presents the weights of the multiclass logistic regression classification model.

Table 5 The multiclass logistic regression estimated parameters

	Weight			
Label	**w1**	**w2**	**w3**	**w4**
Class 0	-1.2669	0.25433	-0.8946	-2.7639
Class 1	0.56965	0.59663	-0.1219	-0.8601
Class 2	-0.5111	0.29056	1.68964	1.2612
Class 3	1.5355	-0.651	-0.8731	2.61811

The more negative weight supports its corresponding class more, and the opposite is true. As an example, we can consider Class 0 weights. The w4 value is -2.7639, which means the feature X10 provides the highest support to Class 0, and the w2 value is 0.25433, which means the feature X3 provides the highest contribution to Class 0.

When the weights are substituted into the softmax function, the multiclass logistic regression classifier model is obtained, which can be used to calculate an instance probability on each class. An instance label can be predicted by substituting the instance feature values into the classifier model. The instance label is determined by the class having the highest probability. Some performance metrics used to evaluate the multiclass logistic regression model are presented in Table 6.

Table 6 The performance metrics of the multiclass logistic regression on the testing set

Performance metric for the testing set	**Precision**	**Recall**	**F1-score**	**Support**
Class 0	0.97	0.99	0.98	200
Class 1	0.0	0.0	0.0	1
Class 2	0.54	0.29	0.38	24
Class 3	0.75	0.87	0.80	45
Accuracy	0.90			270
Macro Avg.	0.56	0.54	0.54	270
Weighted Avg.	0.89	0.90	0.89	270

Based on Table 6, the model accuracy is 90%, with the precision, recall, and F1-score of 89%, 90%, and 89%, respectively.

4.4. Modeling a Decision Tree Classifier

Decision tree modeling has the basic principle of the divide-and-conquer method. Dividing instances into subsets based on the best splitting feature has the highest information gain or gain ratio. Some important hyper-parameters play a critical role in building the decision tree model, including the tree depth maximum and the example minimum on the leaf node. A relationship between the current and previous node is mediated by a conjunction operator. The tree model is presented in Fig. 3 (it just displays the tree having Depth 4 for convenient presentation). A path connecting the root to a leaf implies that the antecedent part is the path from the root to the previous leaf node, and the consequent part is the leaf node. The label of an instance can be predicted by a transversal tree from the root to the leaf node. The instance label is indicated by the tree leaf label traveled.

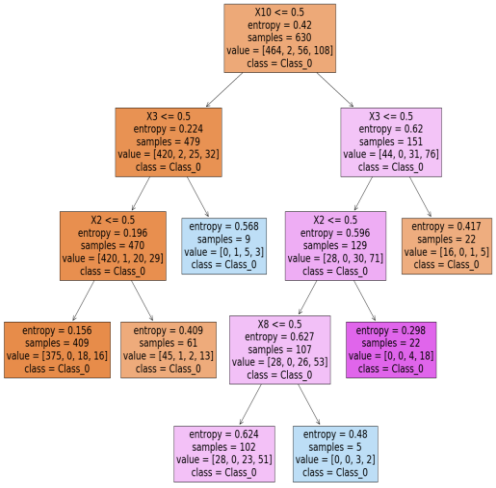

Fig. 3 The decision tree classification model

The performance of the decision tree classifier on the testing set is presented in Table 7.

Table 7 The performance metrics of the decision tree on the testing set

Performance metric for testing set	**Precision**	**Recall**	**F1-score**	**Support**
Class 0	0.97	0.98	0.97	200
Class 1	0.0	0.0	0.0	1
Class 2	0.67	0.42	0.51	24
Class 3	0.75	0.87	0.80	45
Accuracy	0.91			270
Macro Avg.	0.60	0.57	0.57	270
Weighted Avg.	0.90	0.91	0.90	270

Based on Table 7, the decision tree model accuracy is 91%, with the precision, recall, and F1-score of 90%, 91%, and 90%, respectively. The decision tree performance metrics are slightly different from the

logistic regression.

5. Discussion

The total of instances in the testing set is 270. The confusion matrices of the logistic regression and decision tree classifier models are given in Tables 8 and 9, respectively.

Table 8 The confusion matrix of the logistic regression

Actual	Predicted			
	Class 0	**Class 1**	**Class 2**	**Class 3**
Class 0	198	0	2	0
Class 1	1	0	0	0
Class 2	4	0	7	13
Class 3	2	0	4	39

Table 9 The confusion matrix of the decision tree

Actual	Predicted			
	Class 0	**Class 1**	**Class 2**	**Class 3**
Class 0	196	0	1	3
Class 1	1	0	0	0
Class 2	4	0	10	10
Class 3	2	0	4	39

By comparing both confusion matrices, the logistic regression predicts two instances false on the unseen instances of Class 0. Therefore, the logistic regression performance is better than the decision tree (it predicts four instances false) performance. On the other hand, the decision tree has better performance when predicting 10 of 24 instances in Class 2, and the logistic regression can predict 7 of 24 instances in Class 2. The difference in the performance of the predicting instances of Class 2 is 12.5% for the decision tree model. The difference in the performance of the predicting instances of Class 1 is 1% for the logistic regression model. It is enough reason to state that the decision tree model is better than the logistic regression model.

Furthermore, one advantage of the decision tree model is that the important features can be explored clearly. The important features describe the proportion of each feature to become a splitting feature in building a decision tree model. The bar chart in Fig. 4 displays the important features of the built decision tree model.

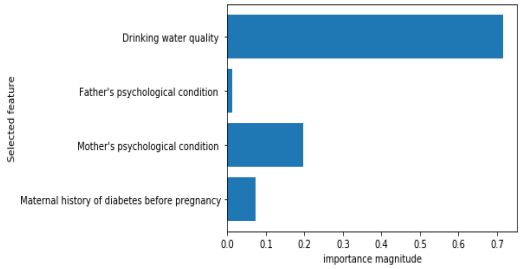

Fig. 4 The important features of the built decision tree model

The drinking water quality becomes the splitting feature of around 70%, and the mother's psychological state becomes the splitting feature of around 20%. On the other hand, the maternal history of diabetes before pregnancy and the father's psychological state become the splitting features of around 10%. The important feature information can guide future research on the study field to pay more serious attention or make a deeper exploration of the important features having high proportion value.

6. Conclusion

Feature selection using the filter approach showed that all predictor features depend on the target feature. The predictor features which are independent of each other include the features of maternal history of diabetes before pregnancy (X2), father's blood pressure (X3), father's psychological condition (X8), and drinking water quality (X10). Furthermore, they are the input features of a classification model.

The dataset class label distribution is imbalanced where Class 0 dominates, which is around 70% of the instances from Class 0, but there are only 0.33% or 3 of 900 instances from Class 1. Of course, the imbalance of class label distribution will cause a disturbance in obtaining a better classification model.

The decision tree model has a higher performance than the logistic regression model, where the performance measures, including the accuracy, precision, recall, and F1-score, are 90%, 89%, 90%, and 89%, respectively, for the logistic regression compared to 91%, 90%, 91%, and 90% for the decision tree model. Besides, the decision tree model can show the degree of important features in the splitting nodes process when building the decision tree model.

Variable selection with the Chi-square test can only be used if the predictor and response variables are categorically scaled. The response variable can be either categorical or numerical, and the predictor variables are also mixtures of categorical and numerical variables. One-way analysis of variance (F-test) was used to test the dependencies between numerical and categorical variables. The dependencies among numerical variables could be tested using the Spearman correlation test. Combining two binary response variables into a multiclass categorical variable can lead to extreme class imbalance problems and be seen as outliers. For example, in this study, there are only 3 out of 900 instances from class 1. Although the decision tree model has better performance than the logistic regression model, according to the authors, the difference in performance between the two models in this study is not significant. The author assumes that if the predictor variables are mixtures of categorical and numerical variables, or even all predictor variables are on a numerical scale, the decision tree classification model will outperform compared to the logistic regression model.

The next research will be interesting if a collected dataset consists of categorical and numerical features. The imbalanced class problem needs resolving by oversampling, undersampling, bootstrap, and outlier modeling to predict outlier class instances.

References

[1] CORMACK B. E., JIANG Y., HARDING J. E., CROWTHER C. A., and BLOOMFIELD F. H. Neonatal refeeding syndrome and clinical outcome in extremely low-birth-weight babies: secondary cohort analysis from the provide trial. *Journal of Parenteral and Enteral Nutrition*, 2021, 45(1): 65-78. https://doi.org/10.1002/jpen.1934

[2] AYELIGN A., and ZERFU T. Household, dietary and healthcare factors predicting childhood stunting in Ethiopia. *Heliyon*, 2021, 7(4): e06733. https://doi.org/10.1016/j.heliyon.2021.e06733

[3] ROMERO C., and VENTURA S. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2020, 10(3): e1355. http://dx.doi.org/10.1002/widm.1355

[4] BAHASSINE S., MADANI A., AL-SAREM M., and KISSI M. Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences*, 2020, 32(2): 225-231. https://doi.org/10.1016/J.JKSUCI.2018.05.010

[5] 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*, 2018, 96(12): 3717-3728. http://www.jatit.org/volumes/Vol96No12/11Vol96No12.pdf

[6] PURWANTO I. N., WIDODO A., and HANDOYO S. System for Selection Starting Lineup of a Football Players by Using Analytical Hierarchy Process (AHP). *Journal of Theoretical & Applied Information Technology*, 2018, 96(1): 19-31. http://www.jatit.org/volumes/Vol96No1/3Vol96No1.pdf

[7] KUSDARWATI H., and HANDOYO S. System for Prediction of Non Stationary Time Series based on the Wavelet Radial Bases Function Neural Network Model. *International Journal of Electrical and Computer Engineering*, 2018, 8(4): 2327-2337. http://doi.org/10.11591/ijece.v8i4.pp2327-2337

[8] HANDOYO S., and MARJI. The Fuzzy Inference System with Least Square Optimization for Time Series Forecasting. *Indonesian Journal of Electrical Engineering and Computer Science*, 2018, 7(3): 1015-1026. http://doi.org/10.11591/ijeecs.v11.i3.pp1015-1026

[9] 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*, 2020, 68(12): 118-122. http://doi.org/10.14445/22315381/IJETT-V68I12P220

[10] GONG C. S. A., SU C. H. S., and TSENG K. H. Implementation of machine learning for fault classification on vehicle power transmission system. *IEEE Sensors Journal*, 2020, 20(24): 15163-15176. https://doi.org/10.1109/JSEN.2020.3010291

[11] 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 Journal of Operations and Quantitative Management*, 2018, 24(4): 277-292. https://www.ijoqm.org/papers/24-4-2-p.pdf

[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*, 2021, 9(2): 135–143. https://doi.org/10.13189/MS.2021.090207

[13] MU Y., LIU X., and WANG L. A Pearson's correlation coefficient based decision tree and its parallel implementation. *Information Sciences*, 2018, 435: 40-58. http://dx.doi.org/10.1016/j.ins.2017.12.059

[14] RÁCZ A., BAJUSZ D., and HÉBERGER K. Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. *Molecules*, 2021, 26(4): 1111. https://doi.org/10.3390/molecules26041111

[15] WANG L., LITTLER T., and LIU X. Gaussian Process Multi-Class Classification for Transformer Fault Diagnosis Using Dissolved Gas Analysis. *IEEE Transactions on Dielectrics and Electrical Insulation*, 2021, 28(5): 1703-1712. https://doi.org/10.1109/TDEI.2021.009470

[16] SEBŐK M., and KACSUK Z. The multiclass classification of newspaper articles with machine learning: The hybrid binary snowball approach. *Political Analysis*, 2021, 29(2): 236-249. https://doi.org/10.1017/pan.2020.27

[17] WAŁĘGA G., and WAŁĘGA A. Over-indebted households in Poland: Classification tree analysis. *Social Indicators Research*, 2021, 153(2): 561-584. https://doi.org/10.1007/s11205-020-02505-6

[18] MENA E., and BOLTE G. Classification tree analysis for an intersectionality-informed identification of population groups with non-daily vegetable intake. *BMC Public Health*, 2021, 21(1): 2007. https://doi.org/10.1186/s12889-021-12043-6

[19] ROJARATH A., and SONGPAN W. Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems. *Applied Intelligence*, 2021, 51(7): 4908-4932. https://doi.org/10.1007/s10489-020-02106-3

[20] BAHASSINE S., MADANI A., AL-SAREM M., and KISSI M. Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences*, 2020, 32(2): 225-231. http://dx.doi.org/10.1016/j.jksuci.2018.05.010

[21] 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. *International Journal of Electrical and Computer Engineering*, 2018, 8(1): 333-343. https://doi.org/10.11591/IJECE.V8I1.PP333-343

[22] WIDODO A., and HANDOYO S. The Classification Performance Using Logistic Regression and Support Vector Machine (SVM). *Journal of Theoretical & Applied Information Technology*, 2017, 95(19): 5184-5193. http://www.jatit.org/volumes/Vol95No19/23Vol95No19.pdf

[23] CHIU I. M., ZENG W. H., CHENG C. Y., CHEN S. H., and LIN C. H. R. Using a multiclass machine learning model to predict the outcome of acute ischemic stroke requiring reperfusion therapy. *Diagnostics*, 2021, 11(1): 80. https://doi.org/10.3390/diagnostics11010080

[24] GNETCHEJO P. J., ESSIANE S. N., DADJÉ A., and ELE P. A combination of Newton-Raphson method and heuristics algorithms for parameter estimation in photovoltaic modules. *Heliyon*, 2021, 7(4): e06673.

https://doi.org/10.1016/j.heliyon.2021.e06673
[25] ABRAMOVICH F., GRINSHTEIN V., and LEVY T. Multiclass classification by sparse multinomial logistic regression. *IEEE Transactions on Information Theory*, 2021, 67(7): 4637-4646. https://doi.org/10.1109/TIT.2021.3075137
[26] TONKIN M., WOODHAMS J., BULL R., BOND J. W., and SANTTILA P. A comparison of logistic regression and classification tree analysis for behavioural case linkage. *Journal of Investigative Psychology and Offender Profiling*, 2012, 9(3): 235-258. https://doi.org/10.1002/JIP.1367
[27] BLANQUERO R., CARRIZOSA E., MOLERO-RÍO C., and MORALES D. R. Optimal randomized classification trees. *Computers & Operations Research*, 2021, 132: 105281. https://doi.org/10.1016/j.cor.2021.105281
[28] PANIGRAHI R., BORAH S., BHOI A. K., IJAZ M. F., PRAMANIK M., KUMAR Y., and HAVERI R. H. A consolidated decision tree-based intrusion detection system for binary and multiclass imbalanced datasets. *Mathematics*, 2021, 9(7): 751. https://doi.org/10.3390/math9070751
[29] FRAIWAN L., and HASSANIN O. Computer-aided identification of degenerative neuromuscular diseases based on gait dynamics and ensemble decision tree classifiers. *PLoS ONE*, 2021, 16(6): e0252380. https://doi.org/10.1371/journal.pone.0252380

参考文:

[1] CORMACK B. E., JIANG Y., HARDING J. E., CROWTHER C. A. 和 BLOOMFIELD F. H. 极低出生体重婴儿的新生儿再喂养综合征和临床结果：来自提供试验的二级队列分析。肠外和肠内营养杂志，2021，45（1）：65-78。https://doi.org/10.1002/jpen.1934
[2] AYELIGN A. 和 ZERFU T. 预测埃塞俄比亚儿童发育迟缓的家庭、饮食和保健因素。赫利昂, 2021, 7(4): e06733。https://doi.org/10.1016/j.heliyon.2021.e06733
[3] ROMERO C. 和 VENTURA S. 教育数据挖掘和学习分析：一项更新的调查。威利跨学科评论：数据挖掘和知识发现，2020，10(3)：e1355。http://dx.doi.org/10.1002/widm.1355
[4] BAHASSINE S.、MADANI A.、AL-SAREM M. 和 KISSI M. 使用改进的卡方进行阿拉伯文本分类的特征选择。沙特国王大学学报-计算机与信息科学, 2020, 32(2): 225-231. https://doi.org/10.1016/J.JKSUCI.2018.05.010
[5] MARJI、HANDOYO S.、PURWANTO I. N. 和 ANIZAR M. Y. 协方差矩阵中的属性多样性对模糊减法聚类中半径参数大小的影响。理论与应用信息技术学报, 2018, 96(12): 3717-3728. http://www.jatit.org/volumes/Vol96No12/11Vol96No12.pdf
[6] PURWANTO I. N.、WIDODO A. 和 HANDOYO S. 使用层次分析法(层次分析法)选择足球运动员首发阵容的系统。理论与应用信息技术学报，2018，96(1)：19-31. http://www.jatit.org/volumes/Vol96No1/3Vol96No1.pdf
[7] KUSDARWATI H.，和 HANDOYO S. 基于小波径向基函数神经网络模型的非平稳时间序列预测系统。国际电气与计算机工程杂志，2018，8(4)：2327-2337. http://doi.org/10.11591/ijece.v8i4.pp2327-2337
[8] HANDOYO S. 和 MARJI。用于时间序列预测的具有最小二乘优化的模糊推理系统。印度尼西亚电气工程与计算机科学杂志，2018，7(3)：1015-1026. http://doi.org/10.11591/ijeecs.v11.i3.pp1015-1026
[9] HANDOYO S. 和 CHEN Y. P. 开发具有生成规则库和优化结果部分的多时间序列预测模糊系统。国际工程趋势与技术杂志，2020，68(12)：118-122. http://doi.org/10.14445/22315381/IJETT-V68I12P220
[10] GONG C. S. A.，SU C. H. S.，和 TSENG K. H. 车辆动力传输系统故障分类机器学习的实现。IEEE传感器杂志，2020，20(24)：15163-15176。https://doi.org/10.1109/JSEN.2020.3010291
[11] HANDOYO S.、MARJI、PURWANTO I. N. 和 JIE F. 使用模糊C均值生成规则库的模糊推理系统来预测印度尼西亚的区域最低工资。国际运营与量化管理杂志，2018年，24（4）：277-292。https://www.ijoqm.org/papers/24-4-2-p.pdf
[12] HANDOYO S.、CHEN Y. P.、IRIANTO G. 和 WIDODO A. 用于分类欺诈公司的逻辑回归和线性判别的可变阈值。数学与统计学，2021，9(2)：135–143. https://doi.org/10.13189/MS.2021.090207
[13] MU Y.，LIU X.，和 WANG L. 基于皮尔逊相关系数的决策树及其并行实现. 信息科学，2018，435：40-58。http://dx.doi.org/10.1016/j.ins.2017.12.059
[14] RÁCZ A.、BAJUSZ D. 和 HÉBERGER K. QSAR/QSPR多类分类中数据集大小和训练/测试分割比率的影响。分子，2021，26(4)：1111. https://doi.org/10.3390/molecules26041111
[15] WANG L.，LITTLER T.，和 LIU X. 使用溶解气体分析进行变压器故障诊断的高斯过程多类分类。IEEE电介质和电绝缘汇刊，2021，28(5)：1703-1712。https://doi.org/10.1109/TDEI.2021.009470
[16] SEBŐK M. 和 KACSUK Z. 使用机器学习对报纸文章进行多类分类：混合二元雪球方法。政治分析，2021年，29（2）：236-249。https://doi.org/10.1017/pan.2020.27
[17] WAŁĘGA G. 和 WAŁĘGA A.

波兰过度负债家庭：分类树分析。社会指标研究，2021，153（2）：561-584。https://doi.org/10.1007/s11205-020-02505-6

[18] MENA E. 和 BOLTE G. 分类树分析，用于对非每日蔬菜摄入的人群进行交叉性知情识别。BMC公共卫生, 2021, 21(1): 2007. https://doi.org/10.1186/s12889-021-12043-6

[19] ROJARATH A. 和 SONGPAN W. 在多类分类问题的集成模型中加权投票的成本敏感概率。应用智能, 2021, 51(7): 4908-4932。https://doi.org/10.1007/s10489-020-02106-3

[20] BAHASSINE S.、MADANI A.、AL-SAREM M. 和 KISSI M. 使用改进的卡方进行阿拉伯文本分类的特征选择。沙特国王大学学报-计算机与信息科学, 2020, 32(2): 225-231. http://dx.doi.org/10.1016/j.jksuci.2018.05.010

[21] NUGROHO W. H.、HANDOYO S. 和 AKRI Y. J. 预测变量的测量尺度对对象分类的逻辑回归建模和学习向量量化建模的影响。国际电气与计算机工程杂志, 2018, 8(1): 333-343. https://doi.org/10.11591/IJECE.V8I1.PP333-343

[22] WIDODO A. 和 HANDOYO S. 使用逻辑回归和支持向量机(支持向量机)的分类性能。理论与应用信息技术学报, 2017, 95(19): 5184-5193. http://www.jatit.org/volumes/Vol95No19/23Vol95No19.pdf

[23] CHIU I. M., ZENG W. H., CHENG C. Y., CHEN S. H., 和 LIN C. H. R. 使用多类机器学习模型预测需要再灌注治疗的急性缺血性中风的结果。诊断，2021年，11(1)：80。https://doi.org/10.3390/diagnostics11010080

[24] GNETCHEJO P. J.、ESSIANE S. N.、DADJÉ A. 和 ELE P. 牛顿-拉夫森方法和启发式算法的组合，用于光伏模块中的参数估计。赫利昂, 2021, 7(4): e06673。https://doi.org/10.1016/j.heliyon.2021.e06673

[25] ABRAMOVICH F.、GRINSHTEIN V. 和 LEVY T. 通过稀疏多项逻辑回归进行的多类分类。IEEE信息论汇刊, 2021, 67(7): 4637-4646。https://doi.org/10.1109/TIT.2021.3075137

[26] TONKIN M.、WOODHAMS J.、BULL R.、BOND J.W. 和 SANTTILA P. 行为案例关联的逻辑回归和分类树分析的比较。调查心理学杂志和罪犯分析，2012，9（3）：235-258。https://doi.org/10.1002/JIP.1367

[27] BLANQUERO R.、CARRIZOSA E.、MOLERO-RÍO C. 和 MORALES D. R. 最佳随机分类树。计算机与运筹学, 2021, 132: 105281. https://doi.org/10.1016/j.cor.2021.105281

[28] PANIGRAHI R.、BORAH S.、BHOI A.K.、IJAZ M.F.、PRAMANIK M.、KUMAR Y. 和 HAVERI R.H. 一种基于综合决策树的入侵检测系统，用于二元和多类不平衡数据集。数学, 2021, 9(7): 751. https://doi.org/10.3390/math9070751

[29] FRAIWAN L. 和 HASSANIN O. 基于步态动力学和集成决策树分类器的退行性神经肌肉疾病的计算机辅助识别。公共科学图书馆一期，2021年，16(6)：e0252380。https://doi.org/10.1371/journal.pone.0252380