Journal of Theoretical and Applied Information Technology
15th August 2022. Vol.100. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

PERFORMANCE OF RIDGE LOGISTIC REGRESSION AND DECISION TREE IN THE BINARY CLASSIFICATION

MARJI[1], SAMINGUN HANDOYO[2, 3]

[1]Department of Informatics Engineering, Brawijaya University, Malang 65145, Indonesia

[2]Department of Statistics, Brawijaya University, Malang 65145, Indonesia

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

E-mail: [1]marji@ub.ac.id, [2,3]samistat@ub.ac.id

ABSTRACT

In everyday life, we are always faced with making decisions to choose the right decision between 2 choices of decision candidates. The development of high-performance binary classification models is a challenge for researchers in the modeling field. Deployment both of logistic regression and decision tree model use the dataset having predictor features which are a mixture of categorical and numerical features, both models tend to suffer an overfitting problem. This study has the aim of building a ridge logistic regression and decision tree model on a dataset that has all features of a binary categorical scale. The novelty of this study is to observe the distribution of the two classes in the dataset using the transformation of principal components and linear discriminant projections and also to explore the importance of feature that plays a role in building the decision tree model. The ridge logistic regression model has an accuracy performance of 84% which is better than the decision tree model having an accuracy performance of 81%. There are only 2 features in the dataset dominating around 80% of the feature importance.

Keywords: *Confusion Matrix, Decision Tree, Feature Importance, Machine Learning, Ridge Logistic Regression*

1. INTRODUCTION

The future of a nation depends on its young generation. A healthy and intelligent young generation can only be realized if the health of the mother during pregnancy and the condition of the baby when born is in good condition. The growth of babies under the age of 1 year is considered very influential on mental health and also the process of growth and development into the younger generation. Therefore, many studies were conducted to classify the weight status of infants under 1 year of age [1-3]. The other studies discuss Cancer incidence and survival trends among infants in the United States [4], while Zhang et al.[5] studied the relationship between maternal weight gain in pregnancy and newborn weight. The dataset in those above researches had the features mixture of a categorical and continuous scale. We will build the classification model with the dataset only containing the categorical features in both predictor and target features.

Linear classification such as logistic regression has a satisfactory performance when it was applied to a suitable dataset especially having the class boundary can be separated linearly [6-7]. The decision tree model is categorized as a nonlinear model [8]. Some decision tree models can assemble to form a more complex model called random forest [9-10]. De Caigny, et al.[11] developed the hybrid model between logistic regression and decision trees. All of the models above have a very satisfactory performance when they are applied to the dataset considered in their research, although both the logistic regression and decision tree models tend to suffer an overfitting problem when there are a large number of the predictor features involved. The main characteristic of their dataset is the large number of predictor features consisting of a mixture of categorical and numerical features. An interesting question is how the performance of both logistic regression and decision tree models when they are applied to the dataset having a few predictors features with binary categorical scale.

To address the above problem, the research has some goals including building ridge logistic regression by applying some learning algorithms, building a decision tree model by limiting the tree depth, exploring the class distribution by scatter plot and histogram, and also displaying the important features used in building decision tree model. Regularization of L2 norm and tree pruning is intended to handle an over-fitting problem. Exploring class boundaries aims to know how hard 2 classes can be separated. Displaying the important features to know how large a feature influences the target feature.

2. LITERATURE REVIEW

A model developer's desire to obtain a model with high-performance Machine learning offers many advantages (such as robustness and accuracy) compared to conventional modeling which is loaded with many assumptions and constraints. Some examples of the application of the machine learning approach include the application of heart disease identification [12], Cyber intrusion detection [13], and classification of mechanical properties of friction stir welding of copper [14].

In machine learning modeling when a dataset contains a target feature a predictive model can be built. On the other hand, if all the features in the data set have the same status, namely as attributes that describe the observed instances, the modeling task will produce a descriptive model. The clustering method is an example of descriptive modeling that is often used in the real world including Marji, et al.[15] investigated the effect of the measuring scale of features on fuzzy subtractive clustering, and Handoyo, et al.[16] applied the hybrid clustering method to classify the health facilities data set in Malang. Another type of descriptive modeling is a model for ranking a set of instances so that the instance having the highest preference for decision-makers can be selected [17].

In predictive modeling, the characteristics of the target feature play an important role. If the target feature is a categorical scale, the model yielded is called a classification model. Another hand, if the target feature is a continuous scale, the resulted model is called a regression model. In machine learning, the regression model has a specific purpose to forecast future value. The development of regression models has been widely applied in various fields including time series modeling [18-20], confirmation of a theory in behavior science [21], and the development of fuzzy inference systems for forecasting [22-24].

The decision-making process will be simpler, easier, and more measurable if the problem is formulated in the form of a binary choice. This encourages the application of classification models in various fields to support the decision-making process [25]. Widodo and Handoyo [26] compared the performance of logistic regression models and support vector machines, while Nugroho, et al.[27] evaluated the performance of Logistic Regression and Learning Vector Quantization. On the other hand, some researchers applied the classification of various instances using the decision tree method. A Classification tree analysis of the over-indebted households in Poland is conducted by Walega and Walega [28], Mena and Bolte [29] done a classification tree analysis for an intersectionality-informed identification of population groups with non-daily vegetable intake, while an ensemble model for multi-class classification problems had been applied by Rojarath and Songpan [30].

3. PROPOSED METHODS

The development of a predictive model with a machine learning approach is oriented to get a model having the ability to predict the unseen value of an instance label class with high accuracy [31]. The splitting data set into training and validation (testing) parts must be carried out in order for the model performance can be evaluated based on unseen data, so as to guarantee that the developed model has a satisfactory performance [32]. Logistic regression is a linear classification model that performs very satisfactorily for the classification of linearly separable binary class data sets [33]. While the decision tree is a nonlinear model having a state of art for classification purposes [34].

3.1 Logistic regression

Consider a random variable Y has a Bernoulli distribution, and a set of predictor features X_1, X_2, …,X_p are independent of each others. The posterior probabilities of each class are given as the following:

$p(Y = 0|X) = \frac{e^{-Z}}{1+e^{-Z}}$ *and* $p(Y = 1|X) = \frac{1}{1+e^{-Z}}$,

where $Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

The ratio between 2 posterior probabilities is called as odd ratio and the log of odd ratio is stated as the following:

$$\log(\text{odd ratio}) = \log\left(\frac{p(Y=0|X)}{p(Y=1|X)}\right)$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (1)$$

The equation (1) is a linear combination between predictor features and associated weights, but the weights (coefficients) $\beta_0, \beta_1, \ldots, \beta_P$ can not be

obtained through the least squared error directly because the values in the left-hand side of the equation (1) are not available. Instead, the problem is solved by maximizing likelihood function [35].

The likelihood function of the n instances taken independently can be defined as:

$$L(\beta) = f(x_1;\beta)\big(1-f(x_1;\beta)\big).f(x_2;\beta)\big(1-f(x_2;\beta)\big) \ldots f(x_n;\beta)\big(1-f(x_n;\beta)\big) = \prod_{i=1}^{n} f(x_i;\beta)^{y_i}\left(1-f(x_i;\beta)\right)^{1-y_i}$$

Maximum likelihood estimation (MLE) is to find β that maximizes the log likelihood function L.

$$arg\max_{\beta} \ln L(\beta) = arg\max_{\beta} ln\left(\prod_{i=1}^{n} f(x_i;\beta)^{y_i}\left(1-f(x_i;\beta)\right)^{1-y_i}\right)$$

This is equivalent to minimize the negative of log likelihood, $\ell(\beta)$. The MLE estimator can be statted as

$MLE_\beta = arg\min_{\beta} \ell(\beta)$, where

$$\ell(\beta) = -\ln L(\beta) = \sum_{i=1}^{n} y_i \ln p_i + (1-y_i)\ln(1-p_i) = -\sum_{i=1}^{n}[y_i \ln p_i + \ln(1-p_i) - y_i \ln(1-p_i)] = -\sum_{i=1}^{n}[y_i \ln p_i + (1-y_i)\ln(1-p_i)] \quad (2)$$

The equation (2) in machine learning term is called the cross entropy loss function [35]. The term p_i is the predicted probability value which is the output of logistic regression model associating with the input features X_i for $i = 1,2,\ldots,p$

The loss function of the ridge logistic regression model is defined by putting the L2 norm of the coefficients into the equation (2) as the following [36]:

$$\ell(\beta)_{ridge} = \ell(\beta) + \frac{\lambda}{2}\sum_{i=1}^{p}\beta_i \quad (3)$$

The ridge logistic regression coefficients estimator are found by minimizing of the equation (3) through the first partial derivative with respect to each coefficient and set it to 0. They are some learning algorithm to solve the minimizing problem above including Liu and Zhang [37] compared some learning algorithm namely 'liblinear', 'newton-cg', 'lbfgs', and also Gupta et al. [38] applied the learning algorithm namely 'sag', and 'saga'.

3.2 Decision tree classification

The basic principle of the decision tree model is to repeatedly partition the set of instances into subsets to maximize the overall class purity score. Prediction of the class label of an instance is done by tracing the decision tree starting from the root node to the leaf node. The decision tree model can handle mixed variables and has a high degree of accuracy [39]. Decision tree construction is top-down, which is done by partitioning the set of instances recursively by selecting an attribute that has the ability to separate the set of instances with the highest separation at each partitioning process [40]. Internal nodes play a role in testing each attribute (feature). The tree branch represents the test result. All leaf nodes contain class labels or class label distributions. At each node, an attribute is chosen to divide the training example into as many different subsets as possible. An instance with an unknown class label is then classified by continuing to trace a suitable path to arrive at the leaf node.

A feature selected as the splitting feature on an internal node is determined by using a score function that measures a degree of purity on each feature and chooses one producing the "purest" nodes. The score function used to measure the purity degree is defined as the following [41]:

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

where |D| is the instances number, and each instance has $X_i, for\ i = 1,2,3,\ldots,p$ features. The first step of constructing a decision tree is to pick up an attribute and the associated value that optimizes a criterion such as information gain. (IG) which is calculated by using entropy.

Entropy value is known as the smallest possible number of bits needed to transmit a stream of symbols drawn from ***X***'s distribution. The entropy of dataset D containing C classes is denoted by H(D) which is defined as

$$H(D) = -\sum_{i=1}^{C} p_i log_2(p_i) \quad (4)$$

Where p_i is the probability of class I, partition process is based on feature F having the highest purity degree, so that the partition produces the subsets $D_1 \sim D_k$ and the entropy after splitting on the feature F is called H(D,F) defined in the equation (5)

$$H(D,F) = \sum_{i=1}^{k}\frac{|D_i|}{|D|}H(D_i) \quad (5)$$

Computing Information Gain (IG) for all available features where the IG formula is given in the equation (6)

$$IG(D,F) = H(D) - H(D,F) \quad (6)$$

The splitting feature is a feature having the highest IG. In the next child node, calculate again the entropy, IG, choose the splitting feature and then make partition. The above process is stopped when the tree depth reached or the minimum instance number was fulfilled [42].

3.3 Data projection methods and feature importance

Principal component analysis (PCA) has proven to be a powerful multivariate exploratory tool for processing and interpreting high-dimensional data. PCA has been applied in various forensic problems

to provide direct or indirect solutions [43]. PCA is used for various applications, such as data compression, feature extraction, and data visualization. PCA linearly projects data from a high-dimensional input space to a low-dimensional feature space. The principal component is a projection vector found by PCA. A projection is composed of one or a few principal components which the principal component matrix Z is defined as the following:

$$Z = U^T(X - m) \quad (7)$$

Where U is the eigenvector obtained through singular value decomposition and m is the mean vector of the input features X.

A linear discriminant is a linear function that can have the role of a linear classification method and a data transformation function [44]. The Fisher linear discriminant (FLD) algorithm was introduced to reduce dimensions and efficiently extract relevant and significant features from high-dimensional data sets [45]. The FLD is defined as

$$y(X) = W^T X + W_0 \quad (8)$$

where W is the weight vector and W_0 is the bias. Both weight vector and bias can be computed by using the constraints ordinary least square. By transforming the input features into y, it will be obtained the new data having the dimension of C-1. In the case of a dataset having binary classes, It will be yielded 1 dimension of the transformed data.

A split (decision function) is composed of a feature (input variable) and a threshold [46]. A threshold value is a categorical or numerical value used as a decision criterion to split a dataset. For example, suppose a feature F has a categorical value domain (Yes or No), so it has 2 possible threshold values. While a numerical feature has the possible threshold values as many as N-1 where N is a number of unique observed values had by the feature. Of course, there are only a few threshold values that are representative. Consider the feature F, a degree of feature F importance is calculated by how many threshold values come from the feature F divided by the total of threshold values. Whenever a node is split on feature j, the joint impurity for the two child nodes is less than the parent node [47].

4. DATA DESCRIPTION AND RESEARCH STAGES

This research uses a dataset obtained from the centre for child development studies at the Health Polytechnic of Wira Husada Nusantara, Tlogomas Lowokwaru Malang 65145, Indonesia. The dataset consists of 9 factors (predictor features) that affect infant weight under 1 year old. Table I presents the response and predictor features used in this study.

Table 1: Response and predictor features and their class label distribution.

Feature name	Label distribution	Label name
Infant weight under 1 year old (Y)	[163, 47]	[0: Normal, 1: Abnormal]
Birth weight (X1)	[159, 51]	[0: Normal, 1: Abnormal]
Pregnancy weight gain (X2)	[166, 44]	[0: Normal, 1: Abnormal]
Pregnancy upper arm circumference (X3)	[178, 32]	[0: Normal, 1: Abnormal]
Complaints during pregnancy (X4)	[180, 30]	[0: Yes, 1: No]
Dietary habit during pregnancy (X5)	[165, 45]	[0: Good, 1: Bad]
Early initiation of breastfeeding (X6)	[170, 40]	[0: Yes, 1: No]
Exclusive breastfeeding (X7)	[150, 60]	[0: Yes, 1: No]
Immunization (X8)	[156, 54]	[0: Yes, 1: No]
Monthly family income (X9)	[178, 32]	[the the 0: Good, 1: Bad]

Table 1 shows that all of the features are categorical (ordinal scale) consisting of 2 class labels (binary classes). The response feature is the infant weight under 1-year-old which has 2 labels namely normal is class 0 or abnormal is class 1. The dataset consisting of 210 instances are divided randomly into training and testing set where as many as 140 instances are used as training set and the remaining 70 instances are used as testing set. The distribution of the training set is as many as 110 instances come from the normal label (class 0) and as many as 30 instances come from the abnormal label (class 1). Meanwhile, the distribution of the testing set is 53 and 17 instances respectively come from class 0 and

class 1. The training set is used for developing a classification model, and the testing set is used to evaluate the model performance.

The stages of the research are as the following:

a. Divide the dataset into training and testing parts.
b. Use the training part to build models of ridge logistic regression which applied the combination of the various learning algorithm including 'liblinear', 'newton-cg', 'lbfgs', 'sag', and 'saga' and the various hyper-parameter values including 0.000001, 0.00001, 0.0001, 0.001, 0.01, and 0.1.
c. Use the testing part to choose the best model of ridge logistic regression having the highest accuracy performance.
d. Calculate the confusion matrix and some performance measures of the ridge logistic regression best model.
e. Use the training part to build a decision tree model where both hyper-parameters (tree depth and minimum instances number at a tree leaf) are tuned by trial and error.
f. Calculate a confusion matrix and performance measures of the optimal decision tree model.
g. Display the distribution class in both scatter plot (done by PCA) and histogram (done by FLD) of the testing part.
h. Explore the important features yielded in the process of decision tree building.

5. RESULTS AND DISCUSION

In this section, we will discuss the selection of ridge logistic regression models from various learning algorithms and hyper-parameter values. Using the selected model, various performance measures are calculated on the testing set. In addition, a decision tree model was also built and various performance measures of the model were calculated on the testing set. Furthermore, the visualization of the two classes is carried out which shows the fact that they cannot be separated by a linear separator. This section concludes by showing the important features that affect the response variable.

5.1 Ridge Logistic Regression Model

The ridge logistic regression model parameters are obtained through the training model on the training set in the various learning algorithms, namely 'liblinear', 'newton-cg', 'lbfgs', 'sag', and 'saga', and also in the various hyper-parameter values, namely 0.000001, 0.00001, 0.0001, 0.001, 0.01, and 0.1. While the learning rate value was set at the value of 0.05. The combination between the learning algorithm and the hyper-parameter values as many as 30 possible ridge logistic regression models where the performance of the model's accuracy on the testing set is given in Figure 1.

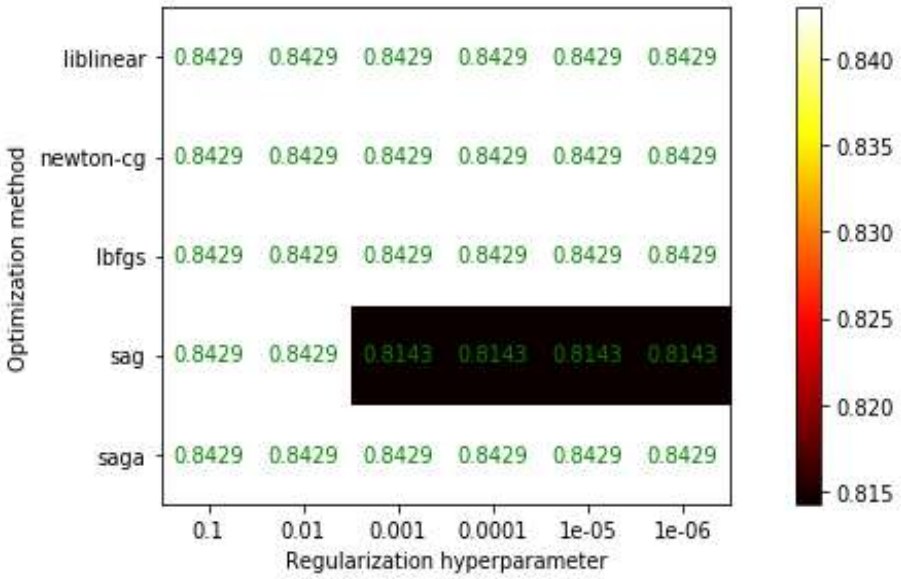

Figure 1: Performance Of The Ridge Logistic Regression Models Accuracy On The Testing Set

In Figure 1, it can be seen that only the learning algorithm 'sag' with hyper-parameters 0.001, 0.0001, 0.00001, and 0.000001 obtained a lower model performance accuracy of 81.43% whereas the other models have a performance accuracy of 84.29%. It seems that the data patterns of the dataset are quite simple, but the two classes cannot be perfectly separated by a linear classification model such as ridge logistic regression. Because all of the predictor features are binary categorical and the dataset only has as many as 9 predictor features, the reasons why we suppose this dataset has simple patterns. It is quite possible that multiple instances in the dataset which come from different classes overlap each other. The characteristics of the dataset had made both treatments of the addition of L2 regularization and the varying learning algorithm which do not have a significant effect on the logistic regression model where the result is a contradiction to the results in Özkale and Arıca [36] and in Liu and Zhang [37].

Furthermore, it is picked up that one of the models has a performance accuracy of 84.29%. We choose the ridge logistic regression model with the combination between the learning algorithm of 'liblinear' and the hyper-parameter value of 0.01. The model has the confusion matrix in the testing set presented in Table 2.

Table 2: Confusion Matrix Of The Ridge Regression Model With 'Liblinear' Algorithm And Regularization 0.01

Actual Class	Predicted Class	
	Class 0	Class 1
Class 0	48	5
Class 1	6	11

In Table 2, it is shown that there are 5 instances of class 0 which are predicted to be wrong, and as many as 6 instances of class 1 are predicted to be wrong. The main diagonal of the confusion matrix shows instances of both classes which are predicted true.

Table 3: The Ridge Logistic Regression Performance In some measures.

Perform. Metric	Precise	Recall	F1-score	Supp.
Class 0	0.89	0.91	0.90	53
Class 1	0.69	0.65	0.67	17
Accuracy		0.84		70
Macro Avg.	0.79	0.78	0.78	70
Weighted Avg	0.84	0.84	0.84	70

The performance of the ridge logistic regression model on the testing set is 84% for all of the performance measures used including accuracy, recall, precision, and F1 score. In class 1, the model performance is around 67% which means the ridge logistic regression will classify 1 of 3 instances that come from class 1 as wrongly classified. In another word, the ridge logistic regression model will classify the abnormal baby as a normal baby with a probability of 33%. While the normal baby will be classified as an abnormal baby with a probability of 10%. The detailed numerical of the model performance is presented in Table 3.

5.2 Classification of Decision Tree Model

The basic principle in decision tree modeling is divide and conquer. The best separation feature is the feature that has the highest information acquisition ratio used to divide a set of instances into 2 subsets. The maximum tree depth and the minimum number of instances in leaf nodes are both hyper-parameters that have an important role in building an optimal decision tree model. The relationship between the path and the previous node is connected by the conjunction operator. In this study as shown by the decision tree model in Figure 2, the path that connects the root node to the leaf node represents an implication where the antecedent part is the path from the root node to the child nodes until the previous last node connected by the conjunction operators. The consequent part is a leaf node where this node also contains the class label information of an instance. Thus to predict the class label of an instance, it can be done by transversal tree from root to leaf node. The class label of an instance is represented by the class label of the leaf node in the decision tree traversed.

Journal of Theoretical and Applied Information Technology
15th August 2022. Vol.100. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

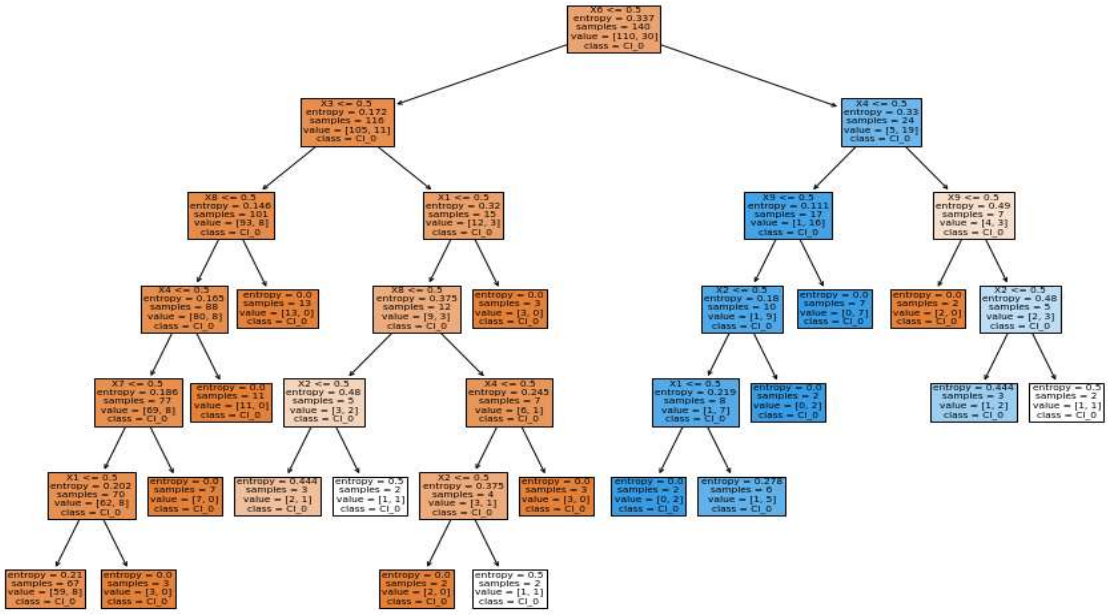

Figure 2: The Decision Tree Model With The Tree Depth 5.

The confusion matrix of the decision tree model in the testing set is presented on the Table 4 following:

Table 4: Confusion matrix of the decision tree model

Actual Class	Predicted Class	
	Class 0	Class 1
Class 0	47	6
Class 1	7	10

In Table 4, it is shown that the predicted instances of the testing set by using the decision tree model yielded 6 instances of class 0 predicted wrong, and as many as 7 instances of class 1 are predicted to be wrong. The main diagonal of the confusion matrix represented the instances predicted true by the decision tree model consisting of 47 and 10 instances respectively coming from class 0 and class 1.

Table 5: The Decision Tree Performance In Some Measures.

Perform. Metric	Precise	Recall	F1-score	Supp.
Class 0	0.87	0.89	0.88	53
Class 1	0.62	0.59	0.61	17
Accuracy		0.81		70
Macro Avg.	0.74	0.74	0.75	70
Weighted Avg	0.81	0.81	0.81	70

The performance of the decision tree model on the testing set is 81% for all of the performance measures used including accuracy, recall, precision, and F1 score. In class 1, the model performance is around 61% which means the decision tree model will classify correctly 3 of 5 instances that come from class 1. In class 0, the model performance is around 88% which means the decision tree model will classify correctly 9 of 10 instances that come from class 0. In another word, the decision tree model will classify the abnormal baby as a normal baby with a probability of around 39%. While the normal baby will be classified as an abnormal baby with a probability of around 12%. The detailed numerical of the model performance is presented in Table 5.

According to the results by Tonkin et al. [39] and by Blanquero et al. [40] stated that the performance of the decision tree model has outperformed the performance of the logistic regression model. The research has contradictory results where the decision tree model performance is lower than the logistic regression performance. The characteristic of predictor features which consist of only the binary categorical scale caused the domain values of splitting feature only limited to 2 choices. The condition leads to the selected splitting features did not have a satisfactory capability to divide the set of instances into 2 class labels. Furthermore, the decision tree model yielded did not have satisfactory performance.

Journal of Theoretical and Applied Information Technology
15th August 2022. Vol.100. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.3 Two Classes Distribution graph and Features Importance

The performance of the two classification models that have not been optimal has prompted researchers to explore further the datasets used in the implementation. Visualization of class distribution using scatter plots and histograms is expected to explain the problems above. On the other hand, the learning process of the ridge logistic regression model which is too easily leads to the hypothesis that there are features in the data set that have a very dominant effect on the target features. Exploration of the level of influence of each feature on the target feature is expected which can confirm the hypothesis above.

A dimensional reduction method such as principal component analysis (PCA) is a method used to obtain the principal components which can represent the original data in a few principal components. For supporting visualization purposes, the transformation data by using PCA to the dataset before they are divided into the training and the testing part yield the 2 principal components explained variance of around 89.6%. The Scatter plot can give a good description of class distribution in the dataset.

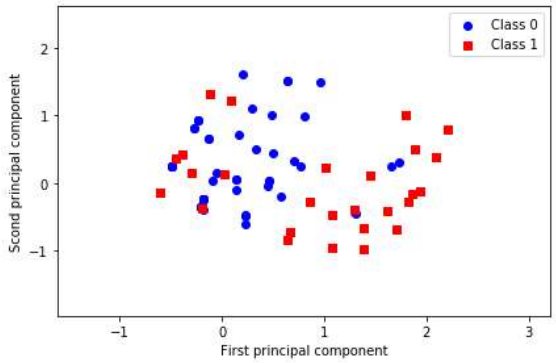

Figure 3. Scatter plot of two classes distribution in two principal components

Because the features plotted in two-dimensional coordinates are able to explain the high variability in the dataset, In Figure 2, the scatter plot visualization of the distribution of the two classes in the dataset shows a random distribution pattern which makes it impossible that the two classes in the dataset can be separated perfectly by a linear classifier model such as the ridge logistic regression.

Furthermore, Fisher's linear discriminant method was used to transform the dataset into 1 dimension. The transformed data can be visualized into a histogram of the two classes. If the two classes can be separated perfectly by the linear classification model, the graphs of the two histograms seem clearly overlap from one another.

Figure 3 shows that the two histograms overlap each other. Class 0 has a transformation value in the range between -0.9 and 0.7, while the transformation values of class 1 range from -0.9 to 1.8. The distribution of class 1 transformation values covers the entire transformation value domain.

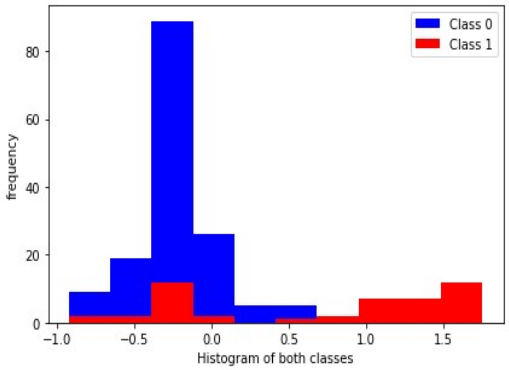

Figure 4. Histogram Of Two Classes In The Linear Discriminant Fisher Transformation

It causes the occurrence of overlapping of several instances of different classes which do not allow to be separated by a linear classification model.

In general, a decision tree can be viewed as a nonlinear classification model. Unfortunately, the decision tree performance to classify the dataset is worse than the ridge logistic regression performance. The entire features of the dataset are categorical features having 2 categories. The characteristic suppose to affect the decision tree model yielded which did not have satisfactory performance. Because the splitting feature has an important role in the process of building a decision tree, the uniform categorical features lead to the best splitting choice did not have a hard competition.

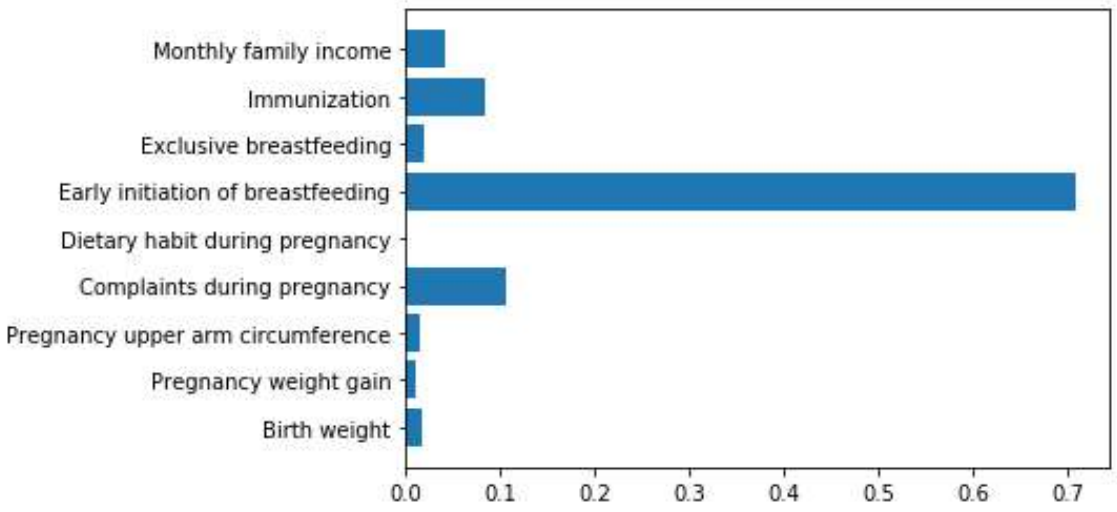

Figure 5: Bar Chart Of The Features Importance Used In The Decision Tree Building.

The decision tree model is formed based on the split consisting of feature and threshold where it is used to divide the instances set into a class. Figure 5 shows the proportion of features' importance that have a role as splitting features in building of the decision tree. The complaints during pregnancy feature has the largest proportion of importance, which it is around 70% be the splitting feature. Furthermore, the features of Early initiation of breastfeeding, Pregnancy weight gain, and Birth weight have the proportion as splitting feature around of 10%, 8%, and 4% respectively. The other features have lower than 4% as the splitting feature, even the Dietary habit during pregnancy feature has the role as splitting feature of 0%. The feature of Complaints during pregnancy dominates in the decision tree building, so the feature plays a very important role in determining the class of an instance. This condition confirms the contradictory result why the decision tree constructed has the performance (81%) which is lower than the ridge logistic regression performance (84%). In this study, the decision tree performance also is reflected by the sum of the importance level proportion of the Complaints during pregnancy and the Early initiation of breastfeeding features.

In addition, the exploration and explaining of the feature importance also can answer the occurrence of the learning process in the ridge logistic regression model where both of the regularization value and the learning algorithm types did not have a significant influence in the majority of cases. Because substantially there are only 2 important features in the dataset (very simple dataset), the addition of the L2 norm regularization penalty in the logistic regression model does not have a significant effect.

6. CONCLUSION

Implementation of ridge logistic regression modeling for the classification of Infant weight under 1 year old produces a model with an accuracy performance of around 84%. The addition of the L2 norm penalty regularization has less effect on the logistic regression model. Similarly, various learning algorithms used for the training model produce ridge logistic regression models which have an almost uniform accuracy performance of 84%. While the implementation of decision tree modeling resulted in a tree model that has a depth of 5 with an accuracy performance of 81%. Visualization of the distribution of the two classes using PCA and FLD transforms gave the pretty clear picture that some instances in the dataset overlap one to each other. In addition, the bar chart plot of the FLD projection shows that the two classes in this dataset cannot be perfectly separated by using a linear classification model such as logistic regression. However, a nonlinear classification model such as the decision tree produced counterproductive accuracy performance, which was lower than the performance of logistic regression. The feature importance exploration provides indirect confirmation of the contradictory accuracy performance in the decision tree and also the learning algorithm which did not have a significant impact to the accuracy performance of the logistic regression model. To get a more comprehensive perspective on the accuracy performance of the ridge logistic regression and decision tree models, future research should use the dataset having a mixture scale of predictor variables

(dataset having some numerical and categorical predictor features) in order for the effect of both L2 norm regularization and various learning algorithms on the performance of the logistic regression model can be explored clearly. In another hand, the type of recommended dataset above also ensures that there are many candidates for the splitting features that lead to the selected splitting feature having the capability to divide a set of instances into 2 class labels with high accuracy.

REFERENCES:

[1] C. Manfredi, A. Bandini, D. Melino, R. Viellevoye, M. Kalenga and S. Orlandi, "Automated detection and classification of basic shapes of newborn cry melody", *Biomedical Signal Processing and Control*, Vol. *45,* 2018, pp. 174-181.

[2] C.W. Lebrão, F.I. Suano-Souza, and R.O. Sarni "Is the Intrauterine Intergrowth-21 Growth Curve Better Than Fenton's for the Classification at Birth and Prediction of Postnatal Growth in Preterm Infants?", *Maternal and Child Health Journal*, Vol. *24*, No. 12, 2020, pp.1446-1453.

[3] B. Zeegers, P. Offerhaus, L. Peters, L. Budé, C. Verhoeven, and M. Nieuwenhuijze, "Impact of maternal height on birth-weight classification in singleton births at term: a cohort study in the Netherlands", *The Journal of Maternal-Fetal & Neonatal Medicine*, 2020, pp. 1-8.

[4] H. Wang, M.C. Mejia, S.J. Gonzalez, R.J. Zoorob, W. Chai, and X.L. Du, "Cancer incidence and survival trends among infants in the United States from 1975 to 2014", *Pediatric Blood & Cancer*, Vol. *68*, No. 4, 2021, pp. e28917.

[5] D. Zhang, L. Zhang, and Z. Wang, "The relationship between maternal weight gain in pregnancy and newborn weight", *Women and Birth*, Vol. *32*, No. 3, 2019, pp. 270-275.

[6] W. Chen, X. Yan, Z. Zhao, H. Hong, D.T. Bui, and B. Pradhan, "Spatial prediction of landslide susceptibility using data mining-based kernel logistic regression, naive Bayes and RBFNetwork models for the Long County area (China)", *Bulletin of Engineering geology and the Environment*, Vol. 78, No. 1, 2019, pp. 247-266.

[7] S. Nusinovici, Y.C. Tham, M.Y.C. Yan, D.S.W. Ting, J. Li, C. Sabanayagam, C.Y. Cheng, "Logistic regression was as good as machine learning for predicting major chronic diseases". *Journal of clinical epidemiology*, Vol. 122, 2020, pp. 56-69.

[8] C. Gao, and H. Elzarka, "The use of decision tree based predictive models for improving the culvert inspection process", *Advanced Engineering Informatics,* Vol. 47, 2021, pp. 101203.

[9] X. Chen, C.C. Zhu, and J. Yin, "Ensemble of decision tree reveals potential miRNA-disease associations", *PLoS computational biology*, Vol.15, No. 7, 2019, pp. e1007209.

[10] Y. Wu, Y. Ke, Z. Chen, S. Liang, H. Zhao, and H. Hong, "Application of alternating decision tree with AdaBoost and bagging ensembles for landslide susceptibility mapping", *Catena,* Vol. 187, 2020, pp. 104396.

[11] A. De Caigny, K. Coussement, and K.W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees", *European Journal of Operational Research*, Vol. 269 No. 2, 2018, pp. 760-772.

[12] J.P. Li, A.U. Haq, S.U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare", *IEEE Access*, Vol. *8*, 2020, pp. 107562-107582.

[13] H. Alqahtani, I.H. Sarker, A. Kalim, S.M.M. Hossain, S. Ikhlaq, and S. Hossain, "Cyber intrusion detection using machine learning classification techniques", *International Conference on Computing Science, Communication and Security*, 2020, pp. 121-131.

[14] S. Thapliyal, and A. Mishra, "Machine learning classification-based approach for mechanical properties of friction stir welding of copper", *Manufacturing Letters*, Vol. *29* , 2021, pp. 52-55.

[15] Marji, S. Handoyo, I.N. Purwanto, and M.Y. Anizar, "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 , 2018, pp. 3717-3728.

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

[17] I.N. Purwanto, A. Widodo, and S. Handoyo, "System For Selection Starting Lineup Of A

Football Players by Using Analytical Hierarchy Process", *Journal of Theoretical & Applied Information Technology*, Vol. 97, No. 1, 2018, pp. 19-31.

[18] H. Kusdarwati, and S. Handoyo, "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, 2018, pp. 2327-2337.

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

[20] S. Handoyo, Y.P. Chen, T.M. Shelvi, H. Kusdarwati, "Modeling Vector Autoregressive and Autoregressive Distributed Lag of the Beef and Chicken Meat Prices during the Covid-19 Pandemic in Indonesia", *journal of Hunan University Natural Science*, 2022, Vol. 49, No. 3, pp. 106-117

[21] H.N. Utami, Candra, and S. Handoyo, "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, 2020, pp. 3571-3575.

[22] S. Handoyo, 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, 2018, pp. 1015-1026.

[23] S. Handoyo, Marji, I.N. Purwanto, and F. Jie, "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, 2018, pp. 277-292.

[24] S. Handoyo, and Y.P. Chen, "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, 2020, pp. 118-122.

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

[26] A. Widodo, and S. Handoyo, The Classification Performance Using Logistic Regression And Support Vector Machine(Svm)", *Journal of Theoretical & Applied Information Technology,* Vol. 9, No. 19, 2017, pp. 5184-5193.

[27] W.H. Nugroho, S. Handoyo, and Y.J. Akri, "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, 2018, pp. 333-343.

[28] G. Wałęga, and A. Wałęga, "Over-indebted households in Poland: Classification tree analysis", *Social Indicators Research,* Vol. 15, No. 2, 2021, pp. 561-584.

[29] E. Mena, and G. Bolte, "Classification tree analysis for an intersectionality-informed identification of population groups with non-daily vegetable intake", *BMC public health,* Vol. 21, No. 1, 2021, pp. 1-14.

[30] A. Rojarath, and W. Songpan, "Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems", *Applied Intelligence,* Vol. 5, No. 7, 2021, pp.4908-4932.

[31] S.K. Behera, A.K. Rath, and P.K. Sethy, "Maturity status classification of papaya fruits based on machine learning and transfer learning approach", *Information Processing in Agriculture*, Vol. 8, No. 2, 2021, pp. 244-250.

[32] P. Lalwani, M.K. Mishra, J.S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach", *Computing*, Vol. 104, No. 2, 2022, pp. 271-294.

[33] M. Karimuzzama, N. Islam, S. Afroz, and M. Hossain, "Predicting stock market price of Bangladesh: a comparative study of linear classification models", *Annals of Data Science*, Vo. 8, No. 1, 2021, pp. 21-38.

[34] N. Yuvaraj, V. Chang, B. Gobinathan, A. Pinagapani, S. Kannan, G. Dhiman, A.R. Rajan, "Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. *Computers & Electrical Engineering*, Vol. *92*, 2021, pp. 107186.

[35] S. Handoyo, N. Pradianti, W. H. Nugroho, Y. J. Akri, "A Heuristic Feature Selection in Logistic Regression Modeling with Newton Raphson and Gradient Descent Algorithm", *International Journal of Advanced Computer Science and Applications,* 2022, Vol. 13, No. 3, pp. 118-126.

[36] M.R. Özkale, and E. Arıcan, "A first-order approximated jackknifed ridge estimator in binary logistic regression", *Computational Statistics*, Vol. 34, No. 2, 2019, pp. 683-712.

[37] T. Liu, and L. Zhang, "Application of logistic regression in web vulnerability scanning", In *2018 International Conference on Sensor Networks and Signal Processing (SNSP),* 2018, pp. 486-490.

[38] A. Gupta, R. Parmar, P. Suri, and R. Kumar, "Determining Accuracy Rate of Artificial Intelligence Models using Python and R-Studio", In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 2021, pp. 889-894.

[39] M. Tonkin, J. Woodhams, R. Bull, J.W. Bond, P. Santila, "A comparison of logistic regression and classification tree analysis for behavioural case linkage", *Journal of Investigative Psychology and Offender Profiling,* Vol. 9, No. 3, pp. 235-258.

[40] R. Blanquero, E. Carrizosa, C. Molero-Rio, D.R. Morales, "Optimal randomized classification trees", *Computers & Operations Research,* Vol. 132, 2021, pp. 105281.

[41] R. Pangrah, S. Borah, A.K. Bhoi, Ijamamf, M. Pramank, Y. Kumar, A. Haverirh, "A consolidated decision tree-based intrusion detection system for binary and multiclass imbalanced datasets", *Mathematics,* Vol. 9, No. 7, 2021, pp. 751-762.

[42] L. Fraiwan, O. Hassann, "Computer-aided identification of degenerative neuromuscular diseases based on gait dynamics and ensemble decision tree classifiers", *Plos one,* Vol. 16, No. 6, 2021, pp. e0252380.

[43] L. C. Lee, and A.A. Jemain, "On overview of PCA application strategy in processing high dimensionality forensic data", *Microchemical Journal*, Vol. *169*, 2021, pp. 106608.

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

[45] L. Sun, R. Liu, J. Xu, and S. Zhang, "An adaptive density peaks clustering method with Fisher linear discriminant", *IEEE Access*, Vol. *7*, 2019, pp. 72936-72955.

[46] Z. Zhou, G. Hooker, "Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. *15*, No. 2, 2021, pp. 1-21.

[47] M. Saarela, and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models", *SN Applied Sciences*,Vol. 3, No. 2, 2021, pp. 1-12.