Modeling Multioutput Response Uses Ridge Regression and MLP Neural Network with Tuning Hyperparameter through Cross Validation

Waego Hadi Nugroho[1], Samingun Handoyo[2], Hsing-Chuan Hsieh[3], Yusnita Julyarni Akri[4], Zuraidah[5], Donna DwinitaAdelia[6]
Department of Statistics, Brawijaya University, Malang 65145, Indonesia[1, 2]
Department of Electrical Eng. And Computer Sci.-IGP, National Yang Ming Chiao Tung University[2]
Hsinchu 30010, Taiwan[2]
Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu 30010, Taiwan[3]
Department of Midwifery, Tribuana Tunggadewi University, Malang 65144, Indonesia[4]
Study Program of Islamic Banking, State Islamic Institute of Kediri, Kediri 64127, Indonesia[5]
Department of Midwifery, Wira Husada Nusantara Health Polytechnic, Malang 65144, Indonesia[6]

Abstract*—The multiple regression model is very popular among researchers in both field of social and science because it is easy to interpret and have a well-established theoretical framework. However, the multioutput multiple regression model is actually widely applied in the engineering field because in the industrial world there are many systems with multiple outputs. The ridge regression model and the Multi-Layer Perceptron (MLP) neural network model are representations of the predictive linear regression model and predictive non-linear regression model that are widely applied in the world of practice. This study aims to build multi-output models of a ridge regression model and an MLP neural network whose hyperparameters are determined by a grid search algorithm through the cross-validation method. The hyperparameter that produces the smallest RMSE value in the validation data is chosen as the hyperparameter to train both models on the training data. The hyperparameter in question is a combination of learning algorithms and alpha values (ridge regression), a combination of the number of hidden nodes and gamma values (MLP neural network). In the ridge regression model for alpha in the range between 0.1 and 0.7, the smallest RMSE is obtained for all learning algorithms used. While the MLP neural network model specifically obtained a combination of the number of nodes = 18 and gamma = 0.1 which produces the smallest RMSE. The ridge regression model with selected hyperparameters has better performance (in the RMSE and R2 value) than the MLP neural network model with selected hyperparameters, both on training and testing data.

Keywords*—Filter approach; hyperparameter tuning; multi-response; neural network; ridge regression

I. Introduction

The health of the mother during pregnancy and the condition of the baby at birth greatly affect the health and intelligence of the younger generation which is the continuation of the sustainability of a nation. Several factors, including the condition of pregnant women, food intake of pregnant women, and health conditions of the family environment affect the condition of the baby at birth, such as stunting events [1]. The model for predicting the occurrence of a class category (stunting or not stunting) is called a classification model. Comparison of the performance of binary classification models, among others, was carried out by Widodo and Handoyo [2] comparing logistic regression and support vector machine, Handoyo et al [3] comparing logistic regression and Linear Discriminant, while Nugroho et al [4] comparing logistic regression and decision tree on the multiclass label response.

If the response variable has a numerical scale such as the length of the baby at birth [5], the predictive model is called a regression model. Santosa et al [6] used a partial least square approach to explain the effect of factors on maternal and child conditions on stunting where this factor is a latent variable. On the other hand, Sajjad et al [7] used multi-output modeling of response variables. Heating and cooling loads with predictor variables were factors related to the layout of a building. Multi-output response variables derived from the condition of the baby at birth (latent variables) consisting of several numerical indicators are very possible and also a challenge when building a model based on a multi-output system.

Regression modeling using machine learning methods has been applied in various fields, including industrial product design by Turetsky et al [8], wind speed prediction by Barhmi et al [9], prediction of imported soybean prices in Indonesia by Handoyo and Chen [10], and also prediction of beef and chicken prices by Handoyo et al [11]. In general, a model that is free from overfitting problems will have satisfactory performance. The ridge regression model is a multiple regression model which is given a penalty of l2 norm [12]. The regularization technique on the neural network is done by adding a penalty l2 norm to the loss function as an attempt to overcome the overfitting problem [13]. However, tuning hyperparameters on ridge regression and neural networks with regularization is generally done by trial and error. In order for these models to have an optimal combination of hyperparameters (producing the best performance), Tso et al

[14], and also Belete and Huchaiah [15] used the k-folds cross-validation method for hyperparameter tuning.

This study aims to build a multioutput model of ridge regression and MLP neural network on the survey dataset with the predictor and response variables derived from the latent variables. The selection of predictor variables that are free from multicollinearity elements is carried out using the filter approach method. In the ridge regression model, the learning algorithm and alpha values are tuned, while the MLP neural network model is carried out to tune the nodes number in the hidden layer and the gamma value using the grid search method. Evaluation of model performance with RMSE and R2 is carried out on both training and testing data.

II. Related Works

Broadly speaking, there are 2 types of modeling in machine learning, namely supervised learning (predictive modeling) and unsupervised learning (descriptive modeling). An unsupervised learning model is characterized by the dataset used in the model building that does not contain a response variable [16]. The response variable measurement unit scale has a critical role, namely if the response variable on a numerical scale will lead to regression modeling, whereas if the response variable is on a categorical scale it will lead to a classification model. Modeling the dengue fever status of a village [17] and modeling the baby's weight status at birth [18] are examples of classification modeling. Regression modeling generally aims to determine the magnitude of the influence of the predictor variable on the numerical response variable and also predicts an unknown value of the response variable based on the values of the predictor variables of a certain instance [19-21]. The research above only involves a single response variable and there has also been no effort to produce a model that is free from overfitting problems.

Often researchers do not pay attention to the unit of measure for each variable contained in the dataset where the action will lead to the incorrect model construction. A commensurate nature of all variables involved in modeling must be maintained so that the arithmetic operations on all formulas used can be guaranteed validity [22]. In addition, correlations between predictor variables should be avoided in order to produce a model that has a low bias value. The selection of predictor variables that are independent of each other can be done before the process of building a model, namely the filter approach method [23]. The filter approach method will greatly reduce the computational cost of complex models involving many parameters [24]. The advantage of the filter approach method is that it reduces the number of predictor variables and still maintains the predictor variables in their original form.

Decision-making in the real world must take into account many factors related to the system being studied. A multi-output model can be a classification or a regression model which if it is given an input, can predict unknown multi-output simultaneously [25]. Assessment of product quality in the food, beverage, and fragrance industries uses a lot of semantic odor perception descriptors. Li et al [26] designed an odor perception descriptor selection mechanism based on a multi-output machine learning model including multiple regression and neural network to find the main odor perception descriptors. Shams et al [27] compared the performance of Multiple Linear Regression and MLP neural networks to predict SO2 concentration in the air of Tehran. The predictor variables used include meteorological parameters, urban traffic data, urban green open space information, and selected time parameters, while the response variable is the daily concentration of SO2. The MLP model has a better performance than the regression model. Siavash et al [28] predict turbine performance using multiple linear regression and a neural network considering as many as 4 channel opening angles as response variables. The performance of the neural network model is more satisfactory than the multiple regression model. The performance comparison between the regression model and the MLP in the above study did not involve tuning the hyperparameters of both models.

The ridge regression model is widely used in practice because of its ease of interpretation, use, and strong theoretical guarantees. In many cases, the model hyperparameter is tuned by using cross-validation, but when the spectrum of the covariate matrix is almost flat and the observations in the observed model are not too high then cross-validation will be detrimental [29]. Meanwhile, van de Wiel et al [30] proposed fast hyperparameter tuning, and Meanti et al [31] proposed Efficient Hyperparameter tuning in the kernel of ridge regression based on cross-validation of data. Tuning hyperparameters on a neural network model using cross-validation data, among others, was carried out by Blume et al [32], and also by Lainder et al [33]. Although there is controversy over the advantages and disadvantages of applying the cross-validation method to set up model hyperparameters, this method is systematic and fair.

III. Proposed Method

A model is called a simple model if the model only involves a few predictor variables and the relationship between predictor variables is linear. The most sought-after models are those that are simple and have high performance. Variable selection is needed to avoid multicollinearity between predictor variables and also to reduce the number of predictor variables.

A. Variable Selection and Data Formatting

In the dataset, each variable is related to its respective units of measure, giving rise to very diverse units of measure. The difference in the unit of measure for each of these variables must be handled in order to meet the rules in arithmetic operations. All variables before being analyzed must have an equivalent unit of measure (commensurate measure). The min-max transformation given to eq.(1) is a simple way to satisfy the commensurate measures of each variable [34].

$$N_{i,} = \frac{P_i - P_{min}}{P_{max} - P_{min}} \tag{1}$$

Where $N_{i,}$ is the normalized value of the i-th instance, P_i is the observed value of the i-th instance, and P_{min}, P_{max} are respectively the minimum and maximum value of the predictor variable P. The eq. (1) will be used to transform all values of the predictor variable P into the range of [0,1] and without a unit of measures.

Variable selection with the filter approach is computationally inexpensive because the selection process does not involve the prospective model to be built. The selection of variables is only based on the level of dependence between two variables. The measuring scale of two variables evaluated for dependence lead to a kind of statistical test, namely the dependence between two categorical variables is evaluated by a chi-square test through a contingency table, and the dependence between numerical and categorical variables is evaluated by a one-way ANOVA test, and the dependence between 2 numerical variables is evaluated by correlation test [35]. The Pearson correlation formula given i.e. Eq. (2) measures a degree of dependence between two numerical variables.

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}} \qquad (2)$$

Where r represents a level of dependency between 2 numerical variables namely the x and y variables. The coefficient correlation r has a value in the range between -1 and 1. The value of r = 0 means that there is perfectly no dependency between 2 variables, while the value of r = abs(1) i.e. absolute 1 means that there is a perfect dependence between two variables. To make a simple task in evaluating dependency between two numerical variables, the value of threshold = 0.7 is set. If the value of r is less than absolute 0.7 then the two variables are declared to have no dependency, for the opposite condition means that the two variables have a dependency and as a result one of these variables must be dropped from the dataset [36].

Modeling in machine learning always provides the out-sample data, which is a subset of data obtained by splitting the dataset that separates from the data to build the model. Out-sample data is used to test the model's performance or often referred to as the testing data.

Fig. 1 presents the splitting of the dataset into training and testing parts, also into sub-training, and validation data [37]. In Fig. 1, Initially, the dataset was randomly divided into the training subset (80%) and the testing subset (20%). Furthermore, the training subset is divided randomly into k-fold which are used to form the sub-training and validation data. In this process, k pairs of sub-training and validation data were obtained. For example, if the fold 1 is as the validation data, the other k-1 folds are as the sub-training data, if the fold 2 is as the validation data, the other k-1 folds are as the sub-training data and so on. Model candidates are trained on all sub-training data with each candidate hyperparameter and the model's performance is evaluated on the corresponding validation data. The grid search method is a way to find the model's hyperparameters that give the best average performance on the validation data.

Training					Testing
fold_1	fold_2	fold_3	...	fold_k	
val_1	sub training				
...	...	...	...	...	
sub training				val_k	

Fig. 1. The Formatting of the Training Data into k-fold Ceoss Validation.

B. Multioutput Multiple Regression and Ridg Regression

In multiple linear regression, if there is more than 1 response variable, it will lead to multi-response modeling which in machine learning is better known as multi-output regression modeling. A simple multi-output regression modeling diagram is given in Fig. 2 as the following.

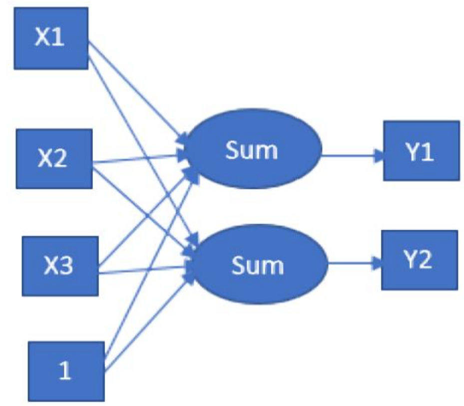

Fig. 2. The Multioutput Multiple Regression Diagram.

In Fig. 2, it is illustrated that there are three predictor variables, namely X1, X2, and X3 as inputs for a system that performs summation operation. This system produces two outputs, namely, Y1 and Y2. In addition, the input system also has a bias of 1. The diagram when expressed in the form of a mathematical formula is as follows:

$$\begin{aligned} Y_1 &= b_1 + w_{11}X_1 + w_{12}X_2 + w_{13}X_3 \\ Y_2 &= b_2 + w_{21}X_1 + w_{22}X_2 + w_{23}X_3 \end{aligned} \qquad (3)$$

$$Y = w^T X \qquad (4)$$

Basically, regression model training is a process to obtain weight and bias values that minimize the loss function, which usually takes MSE (Mean Square Error) as the loss function in machine learning modeling given in the following formula:

$$MSE = \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \qquad (5)$$

The MSE value was optimized using ordinary least squares and can be obtained as an analytical (close form) solution. However, this analytical formula will be problematic if there is strong multi-collinearity between the predictor variables [38-39]. The MSE of a multi-output system is similar to the MSE in Eq. (5) where each Yi and the associated prediction have at least 2 values.

If there are large predictor variables in the multiple regression model, a penalty will be given to the MSE loss function so that the new model is called ridge regression having the loss function formula as the following.

$$MSE_{ridge} = \frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \alpha\|w\|_2^2 \qquad (6)$$

Eq. (6) is a loss function that must be minimized and is a non-linear function in the w parameter and also contains the alpha hyperparameter [40-41]. In this research, various learning algorithms and alpha hyperparameter values were tested. The combination of the learning algorithm and the alpha value that produces the minimum loss function of the ridge regression is selected as the model's hyperparameters.

C. Multi-layer Perceptron Neural Network

A neural network is known as a reliable non-linear model for modeling a complex system. The main difference with a multiple regression model is that the weights on the neural network cannot be interpreted as in the multiple regression model, but the magnitude of the weights only indicates the strength or weakness of the relationship between two adjacent nodes. In addition, in the neural network model, each node uses a certain formula called the activation function which is generally a non-linear function [42]. The diagram of an MLP neural network model is illustrated in Fig. 3 as follows:

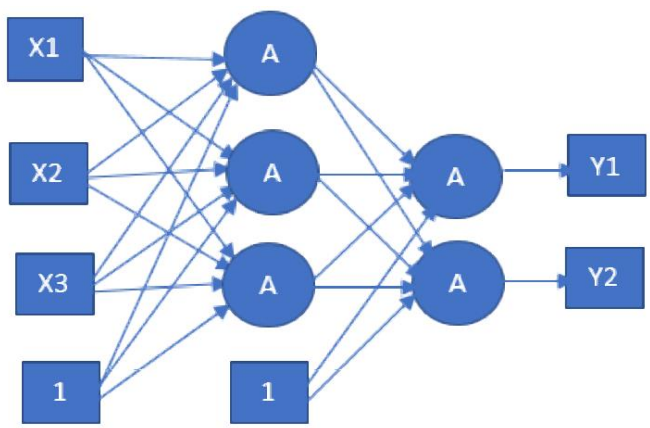

Fig. 3. The Multioutput MLP Neural Network Diagram.

The MLP neural network is characterized by the presence of a hidden layer located between the input layer and the output layer. When this hidden layer is dropped, it will be formed a diagram that is very similar to Fig. 2 except for the existing activation function in the output layer. The neural network model without a hidden layer (very similar to the multiple regression model) is known as the perceptron model. The most popular activation function is known as ReLu which stands for Rectified Linear Units. The ReLu formula is presented in Eq. 7 as follows:

$$ReLu\ (z) = \begin{cases} z & for\ z > 0 \\ 0 & for\ z \leq 0 \end{cases} \quad (7)$$

Where z is a linear combination between the input and the associated weight. For example, the output of the first node on the hidden layer uses the combination linear of $z = b_1 + w_{11}X_1 + w_{12}X_2 + w_{13}X_3$, and the first output of the MLP is obtained by using Eq. (8) as the following:

$$Y_1 = ReLu(output\_A1) \quad (8)$$

Where $output\_A1 = b_h + w_{11} * A_1 + w_{12} * A_2 + w_{13} * A_3$ and $A_1, A_2, and\ A_3$ are respectively $ReLu\ (z1)$, $ReLu\ (z2)$, and $ReLu\ (z3)$. In the neural network term, the process calculates the value of the neural network output such as in Eq. (8) is called the forward step.

The loss function of the MLP neural network is the same as the loss function of the multiple regression in Eq. (5) and Eq. (6) (with and without penalty term). Because the MLP neural network model involves a non-linear activation function such as Relu, the optimal weights for the neural network model cannot be obtained using an analytical solution (close-form solution). The backpropagation algorithm which consists of a forward step and a backward step is used to train this model in obtaining the optimal weights. The forward step aims to calculate the predicted value (network output), while the backward step is the process of updating the weights by applying the gradient descent method and the chain rule to obtain a gradient descent at nodes that are further back [43].

IV. Describing Data and Research Stages

This research uses datasets collected by the Center for Child Development Studies at the Midwifery Academy of Wira Husada Nusantara Malang in 2022. In the dataset, there are 696 sample points which are explained by 20 predictor variables and 3 response variables. The predictor variables were derived from several factors including the condition of pregnant women (variables X1 to X11), food intake of pregnant women (variables X12 to X16), and the health condition of the family environment (variables X17 to X20) which the factors were supposed to affect the condition of the baby at the time of birth (variables Y1 to Y3). The predictor variables consist of 12 Likert scales and 8 ratio scales, while the three response variables are all ratio scales. Table I presents the variables in the dataset along with their minimum and maximum values.

TABLE I. The Variables and Range Values

Symbol	Variable	Min.	Max.
X1	Weight at first check	37	80
X2 (Likert)	Frequency of checks during pregnancy	2	4
X3	Weight gain during pregnancy	1	14
X4	Height at first check	139	170
X5	Circumference of the upper arm	18	36
X6	Body Mass Index	16.4	33.3
X7	Normal blood pressure	80	190
X8	Hemoglobin level	5.8	15.4
X9 (Likert)	The protein level in urine	1	4
X10 (Likert)	The number of complaints during pregnancy	1	5
X11	Gestational age when the baby is born	23	42
X12 (Likert)	Consumption of iron element intake	1	4
X13 (Likert)	Consumption of vegetable protein	1	4
X14 (Likert)	Consumption of animal protein	1	4
X15 (Likert)	Consumption of protein from milk intake	1	4
X16 (Likert)	Consumption of vitamin intake	1	3
X17 (Likert)	Family Income per month	1	5
X18 (Likert)	Quantity and quality of drinking water	1	5
X19 (Likert)	Condition of sanitary facilities	1	5
X20 (Likert)	Cleanliness of the house and environment	1	5
Y1	Baby weight at birth	1	4
Y2	Baby body length at birth	30	55
Y3	Baby health score visually	2	10

Differences in the unit of measurement for variables in this dataset must be addressed before modeling is carried out. All variables involved in building the model must have the same unit of measurement (commensurate measures). Therefore, it is necessary to preprocess all ratio-scaled variables using the minimax transformation. By using the formula in equation (1), the value of the transformation results in the range of 0 to 1, so that finally all variables in the dataset have the same unit of measurement.

Furthermore, broadly speaking, the stages of the process to produce the best model in this study are as follows:

1) Selecting variables using the filter approach method.

2) Splitting the dataset into training and testing subsets.

3) Dividing the training subset into k folds and formatting k fold cross validation data.

4) Tuning the hyperparameter model using the grid search method.

5) Multioutput regression modeling using training subset.

6) Ridge regression multi-output modeling with the best hyperparameters using the training subset.

7) MLP neural network multioutput modeling with the best hyperparameter using the training subset.

8) Evaluating the model performance in both training and testing subsets.

V. Result and Discussion

The quality of the input data used to build a model significantly determines the model's performance. In this section, we will discuss variable selection using the filter approach. The multioutput regression model with the least square of the parameter estimation was built based on the training data, the multioutput ridge regression model with the best hyperparameters obtained through cross-validation was built using training data, and the MLP neural networks model with the best hyperparameters obtained through the hyperparameter tuning process was built using the training data. Furthermore, the performance of the models is evaluated on both the training and testing data.

A. Evaluate Independency among PredictorVariables

In developing a regression model, one of the conditions that must be met is that there is causality between the response and predictor variables. Causality has a meaning that the response variable is influenced by predictor variables. In multiple regression, the predictor variables must also meet the condition that they must be independent of one another. The measuring scale of each variable will determine the appropriate evaluation method to check the independence between the two variables. The independence between the two numerical variables can be evaluated by their correlation value [35]. Table II presents the Spearmen correlation value between two predictor variables presented in the form of a matrix. The matrix main diagonal has a value of 1, which indicates the correlation value in the same variable. Because of the limited space, Table II only presents half part of its column elements. The correlation value of two different variables is expressed in the cells outside the main diagonal.

TABLE II. The Spearman's Correlation between 2 Predictor Variables

Variable	X2	X9	X10	X12	X13	X14	X15	X16	X17	X18
X2	1	0.2	-0.13	-0.23	-0.13	-0.26	-0.01	0.03	-0.26	-0.23
X9	0.2	1	0.23	-0.77	-0.68	-0.69	-0.5	-0.63	-0.66	-0.66
X10	-0.13	0.23	1	-0.22	-0.18	-0.19	-0.24	-0.19	-0.22	-0.21
X12	-0.23	-0.77	-0.22	1	0.79	0.75	0.63	0.74	0.8	0.79
X13	-0.13	-0.68	-0.18	0.79	1	0.68	0.58	0.64	0.65	0.66
X14	-0.26	-0.69	-0.19	0.75	0.68	1	0.47	0.55	0.68	0.65
X15	-0.01	-0.5	-0.24	0.63	0.58	0.47	1	0.43	0.56	0.58
X16	0.03	-0.63	-0.19	0.74	0.64	0.55	0.43	1	0.56	0.53
X17	-0.26	-0.66	-0.22	0.8	0.65	0.68	0.56	0.56	1	0.83
X18	-0.23	-0.66	-0.21	0.79	0.66	0.65	0.58	0.53	0.83	1
X19	-0.3	-0.68	-0.21	0.8	0.68	0.72	0.61	0.5	0.81	0.83
X20	-0.26	-0.66	-0.23	0.79	0.68	0.71	0.6	0.5	0.8	0.82
X1	-0.23	-0.34	-0.1	0.43	0.33	0.4	0.26	0.26	0.41	0.41
X3	-0.2	-0.62	-0.23	0.72	0.64	0.68	0.52	0.46	0.75	0.77
X4	0.11	0.23	0.04	-0.28	-0.2	-0.19	-0.19	-0.19	-0.23	-0.22
X5	-0.2	-0.4	-0.15	0.53	0.39	0.47	0.33	0.36	0.49	0.49
X6	-0.25	-0.4	-0.11	0.5	0.38	0.43	0.31	0.32	0.47	0.46
X7	0.12	0.9	0.22	-0.67	-0.6	-0.58	-0.43	-0.58	-0.53	-0.56
X8	-0.2	-0.55	-0.15	0.62	0.51	0.57	0.38	0.45	0.59	0.56
X11	-0.26	-0.49	-0.18	0.56	0.44	0.53	0.42	0.39	0.57	0.57

In this study, two numerical variables are considered independent if the Spearmen correlation value is less than 0.7. An evaluation of the correlation values in Table II can be done either by row or column as the basis for selection. Look at the correlation value in column 1 (the variable X2 as the basis), it appears that all correlation values are less than 0.7. It can be interpreted that the variable X2 is independent of all other variables, so X2 is selected as a predictor that has no impact on other variables. Next, the correlation value in column 2 (the variable X9 as the basis) is found 2 correlation values that are greater than 0.7, they are the correlation value between variables X9 and X12, as well as the correlation value between variables X9 and X7. This shows that the three variables, namely X9, X12, and X7 are not independent of each other. The three variables can be represented by one of them as the selected predictor variable. In this research, so that the variable selection process is more structured, the variable that acts as the basis for selection (the variable X9) is determined as the predictor variable. Meanwhile, the row and column associated with variables X12 and X7 are removed or dropped from the correlation matrix member (not considered again in the next predictor variable selection process). The selection process was continued by considering the next column (the variable X10) as the selection bases where there were not find correlation values in the X10's column which are greater than 0.7. The variable X10 is selected as the member of the predictor variables without dropping other rows and columns.

As a summary of the predictor variable selection process in the forwarding next columns given a result that the selection process on the basis of variables X9, X14, X17, and X1 caused as many as 8 variables to be excluded from the set of predictor variables, namely variables X12, X7, X19, X20, X18, X3, X5, and X6. Thus the dataset used to build and evaluate the model in this study consists of 12 predictor variables and 3 response variables. The selected variable rows (12 variables) are the variables that have a role as the basis of the selection, and furthermore they as the predictor variable selected as independent variables or input variables of the model to be built.

B. *Multioutput Regression Model*

Initially, the resulting dataset obtained from the selection variables was divided into a training subset (80%) and a testing subset (20%). The training subset data is used to build the model, while the testing subset data is used to evaluate the model's performance. The splitting of the training subset data into five folds aims to form five pairs of sub-training and validation data. The five data pairs will be used for hyperparameter tuning. The emphasis of this research is actually getting a multi-output ridge regression model having a combination of hyperparameters (solver method and alpha value) which produces the smallest MSE value in the validation data. However, the author also considers it necessary to obtain a multi-output multiple regression model with the ordinary least square estimate as the benchmark model. The multi-output multiple regression model was built based on the training subset data obtained coefficients which are presented in Table III.

On the response variable Y1 (Baby weight at birth), the predictor variable X11 (Gestational age when the baby is born) has a very significant effect (13.893). It is followed by variables X16 (Consumption of vitamin intake), X10 (The number of complaints during pregnancy), and X4 (Weight at first check) which have an effect on the response variable of Baby weight at birth respectively 2.478, 1.907, and 1.662. On the response variable Y2 (Baby body length at birth), the 4 predictor variables with a moderate effect are X9(The protein level in urine), X10 (The number of complaints during pregnancy), X16 (Consumption of vitamin intake), and X15 (Consumption of protein from milk intake) where they have an effect on the response variable of Baby body length at birth respectively 0.16, 0.16, 0.137, and -0.132 respectively. In addition, the 4 predictor variables with a large effect on the response variable Y3 (Baby health score visually) are X14 (Consumption of animal protein), X17 (Family Income per month), X9(The protein level in urine), and X13(Consumption of vegetable protein) with the effect magnitude of -0.993, 0.555, -0.553, and 0.536 respectively. It is clear that the influence of the predictor variables on the response variable of Baby weight at birth is very large, while their influence on the response variable of Baby health score visually is greater than their influence on the response variable of Baby body length at birth.

Before building the multi-output ridge regression model using the training subset data, in this study, hyperparameter tuning (solver method and alpha value) was carried out using 5 folds cross-validation data that had been formed based on the training subset data. For each pair of fold cross-validation data, the model's performance is calculated on the validation data. For example, for solver of 'svd' and alpha of 0.1, parameter estimation is carried out with the 1st sub-training fold data, and then the MSE value is calculated on the 1st fold validation data. The parameter estimation is carried out with the 2nd sub-training fold data and the MSE value is calculated on the 2nd fold validation data. The above computation process is carried out up to the 5th sub-training fold and the 5th fold validation data. So each pair of both solver and alpha was performed five times parameter estimation and five times calculation of MSE value using different sub-training and validation data. Fig. 4 presents the average MSE of each combination of solver and alpha in the validation data that it is presented in the form of a heap map.

TABLE III. THE COEFFICIENTS OF MULTIOUTPUT REGRESSION MODEL

Resp.	X1	X2	X4	X8	X9	X10	X11	X13	X14	X15	X16	X17
Y1	0.231	0.127	1.662	-0.017	0.955	1.907	13.893	0.067	0.893	-0.063	2.478	0.223
Y2	0.004	0.029	0.022	-0.028	0.16	0.16	-0.063	-0.202	-0.09	-0.132	0.137	-0.053
Y3	-0.046	-0.01	0.11	0.219	-0.553	0.004	0.094	0.536	-0.993	-0.025	0.264	0.555

The hyperparameter tuning with grid search method and k-folds cross-validation requires a lot of computation tasks in estimating model parameters on sub-training data and calculating the MSE model performance on validation data. The MSE value in Fig. 4 was obtained from the average of 5 MSE values from five validation data and from 5 models generated from five sub-training data. So in this case, parameter estimation and MSE calculations were carried out 150 times. The smallest MSE average value is 1.561 which occurs at alpha values of 0.1 and 0.3 in all solver methods except the 'sag' solver method which has an MSE value of 1.562. If the MSE value used only considers 2 decimal digits, then all combinations of solver and alpha result in the MSE = 1.56 in all solver methods with an alpha value of less than 0.8.

The multi-output ridge regression model is built by choosing one combination of hyperparameters (solver = 'sag' and alpha = 0.5) having the smallest MSE using the training subset data. The resulted coefficients of the model are presented in Table IV.

The predictor variable having the largest effect on the response variable Y1(Baby weight at birth) is the variable X11(Gestational age when the baby is born). It has a very significant effect of13.677 which is followed by variables X16(Consumption of vitamin intake), X10(The number of complaints during pregnancy), and X4(Weight at first check). They have an effect on the response variable of Baby weight at birth respectively 2.441, 1.891, and 1.636. For the response variable Y2(Baby body length at birth), the four predictor variables have a moderate effect namely X10(The number of complaints during pregnancy), X9(The protein level in urine), X16(Consumption of vitamin intake), and X15(Consumption of protein from milk intake). They have an effect on the response variable of Baby body length at birth respectively 0.156, 0.155, 0.136, and -0.133. In addition, the four predictor variables with a large effect on the response variable Y3 (Baby health score visually) are X14(Consumption of animal protein), X17(Family Income per month), X9(The protein level in urine), and X13(Consumption of vegetable protein) with the effect magnitude of -0.985, 0.556, -0.554, and 0.544 respectively. It is clear that the influence of the predictor variables on the response variable of Baby weight at birth is very large, while their influence on the response variable of Baby health score visually is greater than their influence on the response variable of Baby body length at birth.

C. MLP Neural Network Model

Neural network modeling is a type of non-linear modeling that is complex because it involves setting two groups of hyperparameters, namely it related to network architecture and it related to network training processes. The hyperparameters in the network architecture include the number of inputs, the number of outputs, the number of hidden layers, the number of modes in each hidden layer, the activation function employed, the minimized cost function, and others. The hyperparameters in network training include learning algorithms, learning rate values, number of iterations, tolerance values, number of mini-batches, gamma regularization values, and others.

Because the dataset in this study consists of 12 predictor variables and 3 response variables, this leads to neural networks having the architecture of the number of inputs = 12 and the number of outputs = 3. Several hyperparameters were determined by the researcher through a trial and error process, namely the activation function = ReLu, the loss function = MSE, learning algorithm = SGD (stochastic gradient descent with learning rate = 0.01, and momentum value = 0.9), number of iterations = 100, and number of mini-batches = 30. There are two hyperparameters that are considered very important, namely, the number of nodes in the hidden layer and gamma values in L2 norm regularization are determined using the grid search method using the cross-validation data. The variations in the number of hidden nodes that were tested were [12, 18 ,30, 42, 60, 78], while the variations in gamma values were [0.001, 0.005 ,0.01, 0.05, 0.1, 0.5].

The process of finding the combination of the number of hidden nodes and the gamma value that produces the minimum average MSE value in the validation data is similar to that carried out in the process of obtaining the combination of the solver method and alpha value in multioutput ridge regression modeling. In essence, for each combination of the number of hidden nodes and lambda values, network training is carried out on five sub-training data and the MSE value is calculated for the five corresponding validation data, and finally, the average of the five MSE values obtained is calculated. After the average MSE for all combinations of the number of hidden nodes and gamma value is obtained, then in order to facilitate the process of the grid search method, the average MSE value is presented in a heap map in Fig. 5.

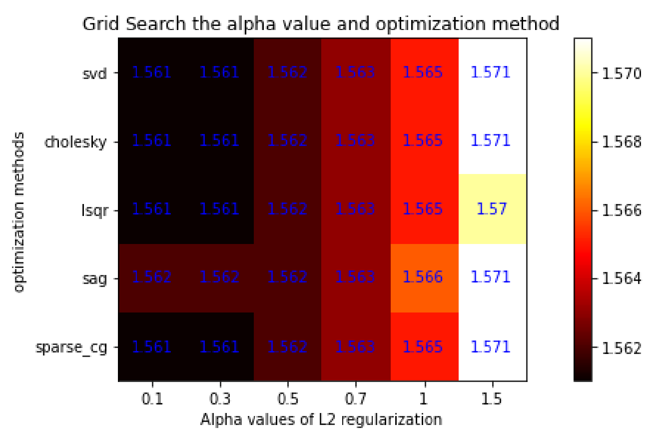

Fig. 4. The Heap Map for Grid Search of the Multi-output Ridge Regression Hyperparameters.

TABLE IV. THE COEFFICIENTS OF MULTIOUTPUT RIDGE REGRESSION MODEL

Resp.	X1	X2	X4	X8	X9	X10	X11	X13	X14	X15	X16	X17
Y1	0.235	0.124	1.541	-0.015	1.002	1.827	12.882	0.093	0.865	-0.048	2.304	0.228
Y2	0.001	0.027	0.02	-0.027	0.138	0.142	-0.071	-0.19	-0.097	-0.135	0.136	-0.05
Y3	-0.046	-0.011	0.113	0.222	-0.554	-0.004	0.12	0.574	-0.951	-0.026	0.269	0.562

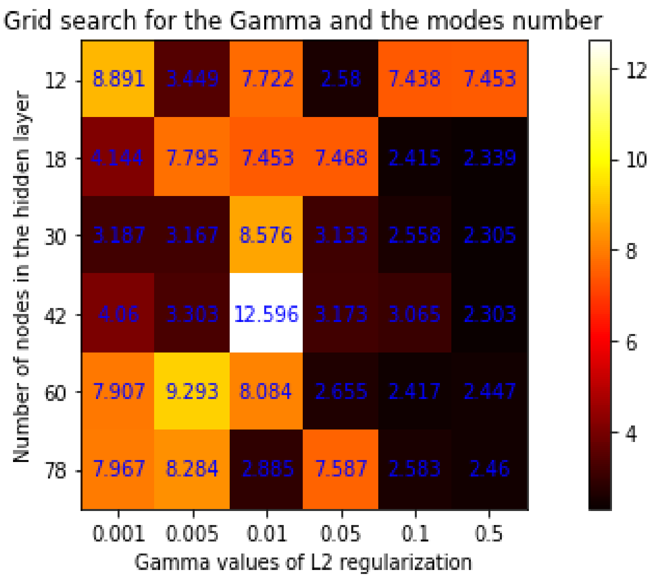

Fig. 5. The Heap Map for Grid Search of MLP Neural Network Hyperparameter.

As explained in the proposed method session, this research uses an MLP neural network architecture whose main feature is that there is only one hidden layer. Based on the average MSE value in the validation data on the heap map in Fig. 5, it is clear that the average MSE value is in the range between 2.415 and 12.596 which is expressed in the darkest (black) and lightest (white) colors. The Heap map also implies that changes in the two hyperparameters greatly affect the MSE average. The combination of the number of nodes and the gamma value that produces the minimum MSE is a combination of the number of nodes = 18 and the value of gamma = 0.1 which will be used to train the network on the training subset data, and then calculate its performance on both the training and testing subset data.

Neural network training with features that include the number of inputs = 12, the number of nodes in the hidden layer = 18, the number of outputs = 3, the number of iterations = 100, and the mini-barch size = 30, the learning algorithm = SGD (learning tare = 0.01 and the momentum value = 0.9) and the value of gamma regularization = 0.1 in the training subset data obtained by the network weight values. The distribution of the resulting network weights in the hidden layer is given in Fig. 6. While the distribution of the resulting network weights in the output layer is given in Fig. 7.

The total number of weights in the hidden layer is (12*18) + (1*18) = 234 where the magnitude of these weights only states the fire strength between each network input and each node in the hidden layer. Based on Fig. 6, it appears that the weight value with the highest frequency (more than 60 pieces) on the average class value = 0.00. The negative weights are about 60 pieces, while the rest are positive weights. Thus, the positive weight dominates in the hidden layer. The effect of the gamma regularization value is to ensure that the weights with very small values become zero, resulting in the number of zeros occupying the mode value of the histogram in Fig. 6.

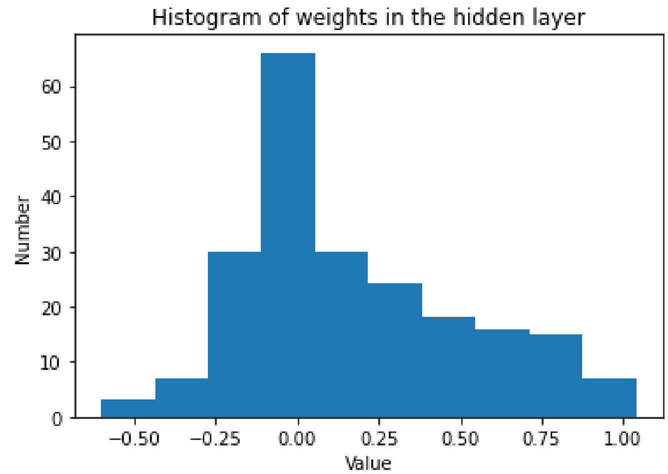

Fig. 6. The Hidden Layer Weights Distribution.

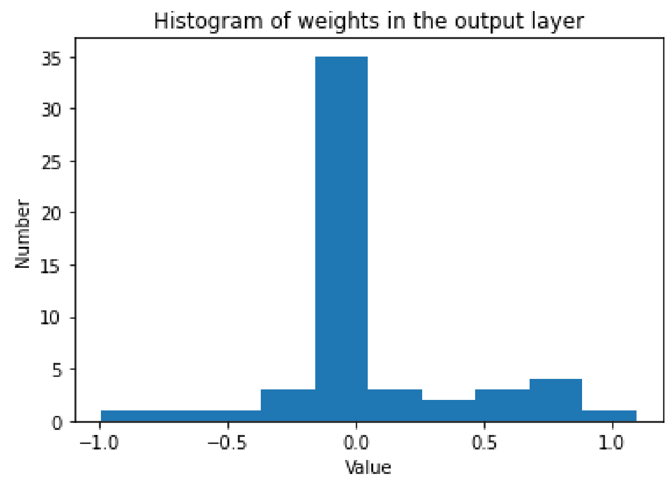

Fig. 7. The Output Layer Weights Distribution.

In the output layer of this neural network model, there are weights as many as (18*3) + (1*3) = 57 pieces. The weights relate the 18 hidden nodes to the 3 output nodes and also relate one bias in the hidden layer to the 3 output nodes. The weights around zero occupy the mode value of the histogram in Fig. 7 which is due to the effect of the l2-norm regularization. The distribution of weights in the output layer is almost similar to the distribution of weights in the hidden layer where positive weights dominate.

D. Discussion

In this section, a discussion is given of the results obtained in the previous session, and also the performance of the model is calculated both on the training and testing subset data using the RMSE and R2 measures.

Based on Table V, in general, the coefficient of the ridge regression model has a slightly smaller effect on the response variables than the coefficient of multiple regressions. This is due to the effect of giving the l2-norm regularization value in the ridge regression model. The predictor variable X11 (Gestational age when the baby is born) has a very dominant effect (13.677) on the response variable Y1 (Baby weight at birth), the predictor variable X9 (The protein level in urine), and X16 (Consumption of vitamin intake) have the greatest influence (0.156) on the response variable Y2 (Baby body length at birth), and the predictor variable X14 (Consumption of animal protein) has the greatest effect (-0.985) on the response variable Y3 (Baby health score visually). The predictor variables X10, X16, and X9 have a considerable influence on two response variables at once. The performance of both regression models and also the MLP neural network model are given in Table VI.

TABLE V. THE COMPARATION OF THE PREDICTOR VARIABLES AFFECTS ON EACH RESPONSE VARIABLE

Response	Predictor	Regress	Ridge regress
Y1(Baby weight at birth)	X11(Gestational age when the baby is born)	13.893	13.677
	X16(Consumption of vitamin intake)	2.478	2.441
	X10(The number of complaints during pregnancy)	1.907	1.891
	X4(Weight at first check)	1.662	1.636
Y2(Baby body length at birth)	X9(The protein level in urine)	0.16	0.155
	X10(The number of complaints during pregnancy)	0.16	0.156
	X16(Consumption of vitamin intake)	0.137	0.136
	X15(Consumption of protein from milk intake)	0.132	0.133
Y3(Baby health score visually)	X14(Consumption of animal protein)	-0.993	-0.985
	X17(Family Income per month)	0.555	0.556
	X9(The protein level in urine	-0.553	-0.554
	X13(Consumption of vegetable protein)	0.536	0.544

TABLE VI. THE PERFORMANCE MODEL ON BOTH TRAINING AND TESTING SUBSET DATA

	Training		Testing	
Model	MSE	R2	MSE	R2
Regression	1.5073	0.6469	1.4362	0.6085
Ridge Regression	1.5074	0.6468	1.4336	0.6098
Neural network	2.2714	0.5823	2.1063	0.5574

All of the developed models have similar performance's characteristics which are the RMSE value in the testing subset is smaller than the RMSE value in the training subset. while the R2 value in the training subset is greater than the R2 value in the testing subset. The multiple regression models consistently have better performance than the ridge regression and MLP neural network model in both the training and testing subsets, although the performance difference between the multiple regression and ridge regression models is very small. This result is in contradiction with the level of complexity in the model building where the MLP neural network model involves as many as 291 weights and also hyperparameter tuning which requires expensive computations. The coefficients of the multiple regression model are obtained based on the close form solution by the ordinary least square method. In another hand, the coefficients of the ridge regression model are obtained using a numerical optimization method that involves several hyperparameters.

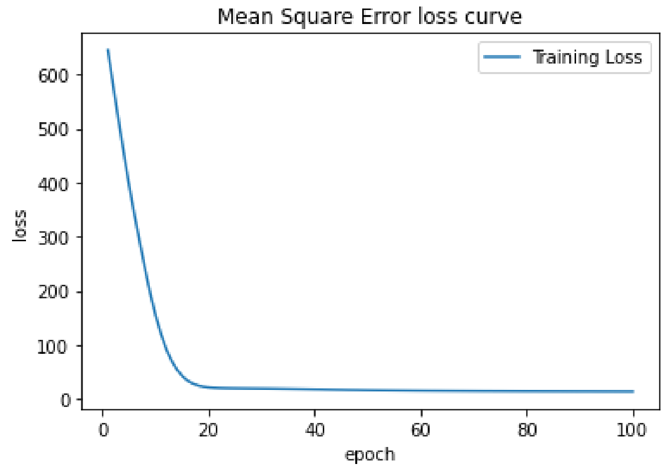

Fig. 8. The Learning Curve of the Logistic Regression Model.

A closed-form solution can be obtained because the predictor variables are independent of each other so the resulting quadratic matrix is not singular. This is one of the impacts of the selection of predictor variables. The outperformance of the multiple regression model also shows that the relationship between the predictor variables and the response variables is based on a linear system. If a linear system is modeled with a non-linear model (such as ridge regression or neural network) it will result in an unsatisfactory performance caused by over modeling. The curve loss function in Fig. 8 strengthens the above argument.

As previously mentioned, the MLP neural network model involves 291 weights that must be optimized using a training subset of 556 instances. The loss function curve in Fig. 8 shows that the value of the loss function has sloped at less than 20 iterations. This means that the model training process is very fast, which only requires updating the weights less than 20 times. This indicates that the system being modeled is a linear system so if it is modeled with a non-linear model, it causes a lot of useless resources or an inefficient modeling process which ultimately results in unsatisfactory model performance.

VI. CONCLUSION

The equivalence of measure units in the dataset must receive careful attention because arithmetic operations on all mathematical formulas can only work if all operands (variables) involved in the formula must be commensurate. The min-max transformation is often applied to satisfy the commensurate nature of the variable. Multi-collinearity between predictor variables must be overcome so that the influence of predictor variables on response variables is unbiased. The Spearman correlation value can be used as a basis for variable selection with a filter approach if the predictor variables are all numerical scale (interval or ratio).

The complexity of a model does not always result in better performance. In this study, the multiple regression model has the best performance compared with the ridge regression and the MLP neural network model. Even the MLP neural network model has the highest RSME and the lowest R2 value compared to the other two models and its performance gap is moderately large. In this dataset, both the predictor and the response variables are manifest variables that construct the variables of predictor and response latent. So it is an interesting idea if in future research this dataset is modeled with another approach such as the structural equation modeling method using the partial least squares algorithm.

REFERENCES

[1] Hadisuyitno J and Riyadi B D, “Determinant Factors of Stunting Events of Toddler in Batu City Indonesia” Systematic Reviews in Pharmacy, vol. 12, no. 1, pp. 231-234, 2021.

[2] 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.

[3] 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.

[4] Nugroho W H, Handoyo S, Akri Y J and Sulistyono A D, “Building Multiclass Classification Model of Logistic Regression and Decision Tree Using the Chi-Square Test for Variable Selection Method“ Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences, vol. 49, no. 4, pp. 172-181, 2022.

[5] TaiwoI A, Adeleye A and Uzoma I C, “A possible model for estimating birth length of babies from common parental variables using a sample of families in Lagos, Nigeria” African Health Sciences, vol. 21, mo. 1, pp. 349-356, 2021.

[6] Santosa A, Arif E N and Ghoni DA, “Effect of maternal and child factors on stunting: partial least squares structural equation modeling” Clinical and Experimental Pediatrics, vol. 65, no. 2, pp. 90-102, 2022..

[7] Sajjad M, Khan S U, Khan., Haq I U, Ullah A, Lee M Y and Baik S W, “Towards efficient building designing: Heating and cooling load prediction via multi-output model” Sensors, vol. 20, no. 22, pp. 6419,2020.

[8] Turetskyy A, Wessel J, Herrmann C and Thiede S, ”Battery production design using multi-output machine learning models” Energy Storage Materials, vol. 38, pp. 93-112, 2021.

[9] Barhmi S, Elfatni O and Belhaj I, “Forecasting of wind speed using multiple linear regression and artificial neural networks” Energy Systems, vol. 11, no. 4, pp. 935-946, 2020.

[10] 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.

[11] Handoyo S, Chen Y P, Shelvi T M and Kusdarwati H, “Modeling Vector Autoregressive and Autoregressive Distributed Lag of the Beef and Chicken Meat Prices during the Covid-19 Pandemic in Indonesia” Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences, vol. 49, no. 3, pp. 220-231, 2022.

[12] Assaf A G, Tsionas M and Tasiopoulos A, “Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression” Tourism Management, vol. 71, pp. 1-8, 2019.

[13] Salgado C M, Dam R S F, Salgado W L, Werneck R R A, Pereira C M N A and Schirru R, “The comparison of different multilayer perceptron and General Regression Neural Networks for volume fraction prediction using MCNPX code” Applied Radiation and Isotopes, vol. 162, pp. 109170, 2020.

[14] Tso W W, Burnak B and Pistikopoulos E N, “HY-POP: Hyperparameter optimization of machine learning models through parametric programming” Computers & Chemical Engineering, vol. 139, pp. 106902, 2020.

[15] Belete D M and Huchaiah M D, “Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results” International Journal of Computers and Applications, pp. 1-12, 2021.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Debbouche N, Ouannas A, Batiha I M and Grassi G, “Chaotic dynamics in a novel COVID-19 pandemic model described by commensurate and incommensurate fractional-order derivatives” Nonlinear Dynamics, pp. 1-13, 2021.

[23] Wang D, Zhang Z, Bai R and Mao Y, “A hybrid system with filter approach and multiple population genetic algorithm for feature selection in credit scoring” Journal of Computational and Applied Mathematics, vol. 329, pp. 307-321,2018.

[24] Nayak S K, Rout P K, Jagadev A K and Swarnkar T, “Elitism based multi-objective differential evolution for feature selection: A filter approach with an efficient redundancy measure” Journal of King Saud University-Computer and Information Sciences, vol. 32, no. 2, pp. 174-187, 2020.

[25] Xu D, Shi Y, Tsang I W, Ong Y S, Gong C and Shen X, “Survey on multi-output learning” IEEE transactions on neural networks and learning systems, vol. 31, no. 7, pp. 2409-2429, 2019.

[26] Li X, Luo D, Cheng Y, Wong K Y and Hung K, “Identifying the Primary Odor Perception Descriptors by Multi-Output Linear Regression Models” Applied Sciences, vol. 11, no. 8, 3320, 2021.

[27] Shams S R, Jahani A, Kalantary S, Moeinaddini M and Khorasani N, “The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models for predicting SO2 concentration” Urban Climate, vol. 37, pp. 100837, 2021.

[28] Siavash N K, Ghobadian B, Najafi G, Rohani A, Tavakoli T, Mahmoodi E and Mamat R, “Prediction of power generation and rotor angular speed of a small wind turbine equipped to a controllable duct using artificial neural network and multiple linear regression” Environmental research, vol. 196, pp. 110434, 2021.

[29] Stephenson W, Frangella Z, Udell M and Broderick T, “Can we globally optimize cross-validation loss? Quasiconvexity in ridge regression” Advances in Neural Information Processing Systems, vol. 34, pp. 24352-24364, 2021.

[30] van de Wiel M A, van Nee M M and Rauschenberger A, “Fast cross-validation for multi-penalty high-dimensional ridge regression” Journal of Computational and Graphical Statistics, vol. 30, no. 4, pp. 835-847,2021.

[31] Meanti G, Carratino L, De Vito E and Rosasco L, “Efficient Hyperparameter Tuning for Large Scale Kernel Ridge Regression” In

International Conference on Artificial Intelligence and Statistics pp. 6554-6572, 2022.

[32] Blume S, Benedens T and Schramm D, “Hyperparameter Optimization Techniques for Designing Software Sensors Based on Artificial Neural Networks” Sensors, vol. 21, no. 24, pp. 8435, 2021.

[33] Lainder A D and Wolfinger R D, “Forecasting with gradient boosted trees: augmentation, tuning, and cross-validation strategies: Winning solution to the M5 Uncertainty competition” International Journal of Forecasting., 2022.

[34] Il Choi H, “Assessment of aggregation frameworks for composite indicators in measuring flood vulnerability to climate change” Scientific Reports, vol. 9, no. 1, pp. 1-14,2019.

[35] Handoyo S, Pradianti N, Nugroho W H and Akri Y J, “A Heuristic Feature Selection in Logistic Regression Modeling with Newton Raphson and Gradient Descent Algorithm “International Journal of Advanced Computer Science and Applications, vol. 13, no. 3, pp. 119-126, 2022.

[36] Cai J, Luo J, Wang S and Yang S, “Feature selection in machine learning: A new perspective” Neurocomputing, vol. 300, pp. 70-79, 2018.

[37] Elgeldawi E, Sayed A, Galal A R and Zaki A M, “Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis” In Informatics, vol. 8, no. 4, p. 79, 2021.

[38] Arashi M, Saleh A M E and Kibria B G, “Theory of ridge regression estimation with applications” John Wiley & Sons, 2019.

[39] Liu H, Cai J and Ong Y S, “Remarks on multi-output Gaussian process regression” Knowledge-Based Systems, vol. 144, pp. 102-121, 2018.

[40] 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.

[41] Marji and Handoyo S,” Performance of Ridge Logistic Regression and Decision Tree in the Binary Classification” Journal of Theoretical & Applied Information Technology, vol. 100, no. 13, pp. 4698-4709, 2022.

[42] Hwang J, Lee J and Lee K S, “A deep learning-based method for grip strength prediction: Comparison of multilayer perceptron and polynomial regression approaches” Plos one, vol. 16, no. 2, e0246870, 2021.

[43] Plonis D, Katkevičius A, Gurskas A, Urbanavičius V, Maskeliūnas R and Damaševičius R, “Prediction of meander delay system parameters for internet-of-things devices using pareto-optimal artificial neural network and multiple linear regression” IEEE Access, vol. 8, pp. 39525-39535, 2020.