Contents lists available online at TALENTA Publisher

DATA SCIENCE: JOURNAL OF COMPUTING AND APPLIED INFORMATICS (JoCAI)

Journal homepage: https://talenta.usu.ac.id/JoCAI

Preventing recession through GDP growth prediction: A classical and machine learning classification approach

Prilyandari Dina Saputri [1], *Arin Berliana Angrenani*[2], *Ika Nur Laily Fitriana*[2]

[1] *Department of Actuarial Science, Institut Teknologi Sepuluh Nopember*
[2] *Department of Statistics, Institut Teknologi Sepuluh Nopember*

email: [1] *prilyandaridina@its.ac.id,* [2]*aberlianaa@gmail.com ,* [2]*ikanurlaily97@gmail.com*

ARTICLE INFO

Article history:
Received 19 December 2022
Revised 18 June 2023
Accepted 13 May 2023

Keywords:
Accuracy
COVID-19
Data Classification
Machine Learning
Regional GDP

Corresponding Author:
prilyandaridina@its.ac.id

ABSTRACT

Classification methods are a popular method applied in many various fields of science. This research proposed a comparison of classification methods using regional GDP data for 2019-2020, before and during the COVID-19 pandemic, by predictor variables; percentage of workers, foreign direct investment (PMA), regional revenue (PAD), general allocation fund (DAU), revenue sharing fund (DBH), and the dummy of COVID-19. Economic growth, most commonly using a gross regional domestic product, is experiencing a recession or acceleration, especially before and during the COVID-19 pandemic. To represent the effect of predictor factors on categorical response variables, different machine learning classification algorithms are used, namely logistic regression, neural network (NN), random forest, support vector machine (SVM), and bayesian model averaging (BMA). Every classifier has its unique characteristic, performing well in certain datasets but not in others. Hence, it is always a quest to find the best classifier to use for a certain dataset. The results are that all selected machine learning models can classify the regional GDP growth perfectly for the training data, but, NN model outperforms the other methods with an accuracy of 100% in training and testing data. COVID-19 and the PMA are the most significant variables predicting regional GDP growth for all models. Further research relating to interpretable machine learning, such as feature interaction, global surrogate, and Shapley values, is also necessary to predict regional GDP growth using machine learning methods.

IEEE style in citing this article: [citation Heading]
P D. Saputri, A.B. Anggrenani and I.N.L. Fitriana, " Preventing recession through GDP growth prediction: A classical and machine learning 7classification approach," *DATA SCIENCE: JOURNAL OF COMPUTING AND APPLIED INFORMATICS (JoCAI)*, vol. 7, no. 2, pp. 51-68, 2023.

1. Introduction

Classification is a data mining technicality that specifies classes to a data set to help with predictions and analysis. The classification is also a function to extract data in a group to base classes or groups. A classification task starts with a data group whose category tasks do know. The classification aims to truly predict the targeted status in the data and discover how that set of attributes reaches its conclusion [1]. Classification methods categorize data to use at their highest level of effectiveness and efficiency [2].

Classification methods are a popular method applied in many various fields of science. Training data is used in classification models to develop a classification model that predicts the class label for a new sample. Classification model outputs might be discrete, as in a decision tree classifier, or continuous, as in a Naive Bayes classifier [3].

Logistic regression, neural network (NN), random forest, support vector machine (SVM), and bayesian model averaging (BMA) are some of the machine learning classification approaches used to model the effect of predictor variables on categorical response variables. Every classifier has its unique characteristic, performing well in certain datasets but not in others. Hence, it is always a quest to find the best classifier to use for a certain dataset [4]–[7]. For example, the random forest can reduce over-fitting, and its classifier is more accurate than decision trees in most cases. Also, it can give a probability over the prediction, whereas SVM cannot provide. Still, it is slow real-time prediction, challenging to implement, and has a complex algorithm. Another example is logistic regression and SVM. SVM handles outliers better as it derives maximum margin solution, and hinge loss in SVM outperforms log loss in LR [2].

Many applications in various fields of science apply classification methods. One of them is the economic field. Over time, the rate of economic growth is the most important factor in determining a region's economic performance [8]. To determine national economic growth, the Gross Domestic Product (GDP) value at constant prices is used. GDP is the total worth of goods and services generated in a certain economic sector during a given period. There's also Gross Regional Domestic Product, one of the essential metrics for determining a country's economic situation throughout time, both at current prices and at constant prices [9]. If the amount of products and services produced increased, the economy would grow. Every country and region wants a high and stable rate of economic growth. It will also benefit society's overall welfare and prosperity [10]. Regional GDP is also used as a macroeconomic indicator to evaluate economic performance and formulate various policies.

Political factors and government policies significantly impact regional GDP over time. But, in early 2020, there was a pandemic caused by a virus that affected many sectors in Indonesia. It is called Coronavirus disease (COVID-19) that caused by SARS-COV2. The outbreak started in China, late 2019, then spread worldwide, including Indonesia [11]. From 30 December to 5 July, over 184 million COVID-19 cases have been reported globally and over 2.2 million COVID-19 cases in Indonesia [12]. This virus spreads person-to-person so rapidly that many countries, including Indonesia, have implemented lockdown policies. As a result of the policies made by the spread of COVID-19, in Quarter II-2020, the Indonesian economy's GDP was only IDR 3,687.7 trillion and IDR 2,589.6 trillion, respectively, based on current and constant prices. Economic growth in the second quarter of 2020 contracted by -5.32 percent compared to the first quarter of 2020 and by -4.19 percent compared to the first quarter of 2020. This economic growth contraction has significant monetary worth. Meanwhile, cumulatively in Semester I 2019, growth contracted by -1.26% [13]. Based on this data, it can be seen that Indonesia's economic growth experienced a negative change in 2020.

Regional GDP data from the Central Bureau of Statistics (BPS) repository has been used by several researchers in studies related to machine learning classification or modeling [14]–[20]. Nasution and Matondang analyze the problem in North Sumatra Province's agricultural sector, which has the largest land area but is not the largest employer of workers. The purpose of this study is to examine the impact of leading sector workers on North Sumatra Province's regional GDP [14]. Panel data regression analysis of pooled data is used in this qualitative research. The study explains that the leading sector workers affect the regional GDP of North Sumatra Province. However, the current study only considers one variable. Malau and Loren used three variables to predict regional GDP: workers, investment, and exports [19]. These factors suspected can increase the regional GDP. The study concluded that investment and exports have no partial effect on regional GDP, but workers partially affect regional

GDP. Furthermore, according to the findings, future studies should collect more data with more variables for analysis. The regional GDP values were used by Muchisha et al. to predict real-time regional GDP growth, whether the region would be experiencing a recession or accelerate [15]. It used 18 variables, including quarterly macroeconomic and financial market indicators such as foreign direct investments. The researchers compared the performance of six popular machine learning algorithms; random forest, LASSO, ridge, elastic net, neural networks, and support vector machines, in forecasting GDP growth in real-time from 2013:Q3 to 2019:Q4. The results showed that all these models' performance outperformed the time series model. The individual model that showed the best performance is Random Forest. The effect of workers, domestic investment, foreign direct investment, and government expenditure on regional GDP in the Eastern Indonesia Region are analyzed by Istiqomah et al. [16]. This study used regression on panel data from 12 provinces from 2011 to 2016 and discovered that workers, domestic investment, foreign direct investment, and government expenditure all have a positive and significant impact on regional GDP, implying that all independent variables contributed to the region's economic growth.

Several effects of regional revenue and revenue sharing funds were proposed by Sri and Suyana to predict economic conditions [17]. The research was performed at the University of Udayana, Bali, where those variables were used to predict capital expenditure, welfare communities, and economic performance using regional GDP separately. They found that regional revenue positively affects capital expenditure, but revenue sharing funds have no significant effect during 2010-2017. But, both of them positively affect economic performance and the welfare communities. They concluded that various improvements might be made in future studies, such as the data sample and variables used.

The data analysis technique used the Panel Vector Error Correction Model (PVECM) and the Panel Granger Causality Test to determine the relationship between economic growth, using regional GDP, and local revenue components, such as regional tax revenue, regional retribution revenue, regional wealth revenue, and other legitimate revenue was proposed by Susanto and Sugiyanto [20]. They used panel data from 2005 to 2015 of 35 regencies and cities in Central Java. As a result, tax revenue, retribution revenue, and regional revenue from the previous year have a positive and significant impact on economic growth. In contrast, regional wealth revenue has a negative and significant impact on economic growth. However, the data used in this study had a substantial variance in the regional revenue structure's composition.

A study about economic growth before and during the COVID-19 pandemic was proposed by Anisah [18], where the data was cross-section with individual districts or cities in West Java Province and the three investment variables. The three investment variables are foreign direct investment, domestic direct investment, and government capital expenditures before and during the COVID-19 pandemic. The results explain that the pandemic has reduced foreign direct investment and government capital expenditure. On the other hand, it was increased domestic direct investment related to the new businesses that related to the handling of the COVID-19 pandemic. Before the pandemic, the real impact of foreign direct investment came from the first and second years after investment. In contrast, the real impact came from three years after investing in foreign direct investment during the COVID-19 pandemic.

However, existing studies rarely use classification methods to determine regional GDP is experiencing a recession or acceleration. Also, there are still very few studies related to regional GDP before and during the COVID-19 pandemic. So that an analysis of the classification method will be carried out using regional GDP data for 2019-2020, before and during the COVID-19 pandemic. This paper is organized as follows. The next section described the data and methods we used, followed by the result and discussion. In the final section, we provide the conclusion and recommendation for future research.

2. Methods

2.1 Data Source

This study used secondary data from the Central Bureau of Statistics (BPS) Indonesia. We use data in 2019 and 2020. Data for 2019 depicts conditions before COVID-19, and data for 2020 depicts conditions during COVID-19. The sample unit is a province in Indonesia in two years, consisting of 34 provinces. Thus, the number of observations is 68. We split the data into training data and testing data with a ratio of 70:30 randomly.

2.2 Data Description

This study's response variable (Y) is regional GDP, where category 1 denotes a negative regional GDP and 0 denotes a positive regional GDP. The variables employed in this study were response and explanatory variables. We use the following explanatory variables since these variables are significantly related to the economic growth in the previous research [17], [20]. The following are the explanatory variables used in this study: the percentage of workers, residents who work with the main job status as trying to be assisted by permanent workers/paid workers and workers/employees/employees; foreign direct investment (PMA), investment activities to conduct business in the territory of the Republic of Indonesia carried out by foreign investors, whether using foreign capital fully, or in joint ventures with domestic investors; regional revenue (PAD), consisting of regional taxes, regional levies, share of BUMD profits, revenues from agencies, and other revenues; general allocation fund (DAU), funds originating from APBN revenues allocated to regions with the goal of equitable distribution of financial capacity among regions to fund regional needs in the context of implementing decentralization; revenue-sharing fund (DBH), funds originating from APBN revenues allocated to regions based on a certain percentage to fund regional needs in the context of implementing decentralization [17], [20]; and dummy cases of COVID-19, with a value of 0 for no cases of COVID-19 and 1 for existing cases of COVID-19.

2.3 Binary Logistic Regression

Binary logistic regression is a statistical modeling technique that uses explanatory variables to predict binary response variables. These binary forms by categorical data, commonly "success" or "failure," and interpreted by odds ratio. In logistic regression, we used a sigmoid function instead of a linear function. The sigmoid function is able to map the predicted values to probabilities. Hence, the output is range from 0 to 1. We can classify the observation as a success if the probability is more than 0.5. The formula for the logistic regression probability is shown in equation 1 [21].

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{1}$$

where, is the estimating parameter for the explanatory or independent variable, and is the probability of “success” category on the dependent variable.

The null hypothesis of the overall model states that all regression coefficients are zero. Rejection of this null hypothesis implies that at least one regression coefficient is non-zero meaning the logistic regression equation in (1) able to predict the probability of the regional GDP status.

2.4 Neural Network

Neural network is one of the flexible nonlinear methods that apply the biological system concept. This study used a feedforward neural network consisting of the input layer, hidden layer, and output layer. The neuron in the hidden layer receives the information from the input layer, and then the obtained value is sent to the output layer. The neural network model that consists of p input and m neuron in one hidden layer can be expressed as the following equation.

$$f(\mathbf{x}_t, \mathbf{v}, \mathbf{w}) = g_2 \left\{ \sum_{j=1}^{m} v_j g_1 \left[\sum_{i=1}^{p} w_{ji} x_{it} \right] \right\} \tag{1}$$

where g_1, g_2 are the activation functions, $\mathbf{w}$ are the estimated parameters that connect the input to the hidden layer, $\mathbf{v}$ are the estimated parameters that connect the hidden to the output [22].

2.5 Random Forest

Random forest is a popular classification method developed from the decision tree, which Breiman first introduced in 2001. Many classification trees were produced in the random forest to obtain accurate predictions through majority voting. Each tree is computed independently using several predictor variables. We can arrange the number of trees and the number of subset variables used in the sample selection [23]. The algorithm of the random forest method is:

a. Select the sample dataset using the bootstrap sample.
b. In each dataset, compute the classification tree using several subset predictor variables.
c. Calculate the prediction result using the classification tree.
d. Repeat until the number of the tree is reached and produce a forest.
e. Compute the final prediction using majority voting from all classification trees.

2.6. Support Vector Machine (SVM)

SVM is a machine learning technique that employs a technique for determining a classifier function capable of classifying data into two distinct classes [24]. SVMs used kernel functions to transform non-linear classifiers in the input space to a higher dimension (feature space). The kernel function is a function that maps data to a higher-dimensional space to give the data a more structured appearance and make it easier to separate [25]. Some kernel functions that can be used are shown in Table 1 [26].

Table 1 Kernel Function SVM

Kernel	Function
Linier	$K\left(\mathrm{x},\mathrm{x}^T\right) = \mathrm{x}^T\mathrm{x}$
Polynomial	$K\left(\mathrm{x},\mathrm{x}^T\right) = \left(\mathrm{x}^T\mathrm{x}+1\right)^p$
Radial Basis Function	$K\left(\mathrm{x}_t,\mathrm{x}_u\right) = \exp\left(-\frac{1}{2\sigma^2}\left\Vert \mathrm{x}_t,\mathrm{x}_u \right\Vert^2\right)$
Sigmoid	$K\left(\mathrm{x},\mathrm{x}^T\right) = \tanh\left(\gamma \cdot \mathrm{x}^T\mathrm{x}+r\right)$

2.7 Bayesian Model Averaging (BMA)

BMA is a technique for predicting the best model by averaging the posterior distributions of all possible models. By combining several models, BMA accounts for model uncertainty [27]. Let Δ be the value to be predicted and $M_1,\ldots,M_q$ represents the entire model to be formed, the posterior distribution of the Y data can be written as equation (3).

$$\Pr\left(\Delta \middle| Y\right) = \sum_{k=1}^{q} \Pr\left(\Delta \middle| M_k, Y\right)\Pr\left(M_k \middle| Y\right) \tag{2}$$

The posterior probability for the M_k model is

$$\Pr\left(M_k|Y\right)=\frac{\Pr\left(Y|M_k\right)\Pr\left(M_k\right)}{\sum_{l=1}^{q}\Pr\left(Y|M_l\right)\Pr\left(M_l\right)} \tag{3}$$

where, the marginal likelihood of M_k model is

$$\Pr\left(Y|M_k\right)=\int\Pr\left(Y|\theta_k,M_k\right)\Pr\left(\theta_k|M_k\right)d\theta_k \tag{4}$$

with θ_k vector of parameter model, $\Pr\left(\theta_k|M_k\right)$ is prior density of θ_k in M_k model, $\Pr\left(Y|\theta_k,M_k\right)$ is likelihood, and $\Pr\left(M_k\right)$ is prior probability if M_k is the right model [28]. The posterior mean Δ is given as follows in equation (6).

$$E\left(\Delta|Y\right)=\sum_{k=0}^{q}\Pr\left(M_k|Y\right)E\left(\Delta|M_k,Y\right) \tag{5}$$

$E\left(\Delta|Y\right)$ shows the weighted expectation value Δ for each possible combination model (weight is determined by priors and models) [28].

2.8 Evaluation Criteria

The most frequently used evaluation criteria in the literature on data classification is accuracy. Accuracy is the ratio of true (positive and negative) predictions to the overall data. In GDP data, the data classes are balanced so that the measurement of model evaluation using accuracy is sufficient to measure the model's goodness [3]. The formula to calculate accuracy shown in equation (7).

$$Accuracy=\frac{TP+TN}{Total} \tag{6}$$

where, TN is true negative (refers to correctly classified negative classes), and TP is true positive (refers to correctly classified positive classes) [29]. Positive classes correspond to the negative value of GDP, while negative classes correspond to the positive value of GDP.

3. Results

In this section, you present your findings. Typically, the Results section contains only the findings, not any explanation of or commentary on the findings (see below).

3.1 Data Exploration

The predictor variables relating to the economic condition are workers, PMA, PAD, DAU, and DBH. We used a boxplot to compare the value of these variables in the condition of economic decline. The economic decline is denoted by *Y*=1, and *Y*=0 denotes the economic improvement. We have 32 provinces that experienced negative RGDP growth. As the total number of observations is 68 provinces, the ratio of event and the non-event is equal to 1:1.125. Thus, the two classes of the data are balanced. The characteristics of the economic predictor variables can be visualized in Figure 1.

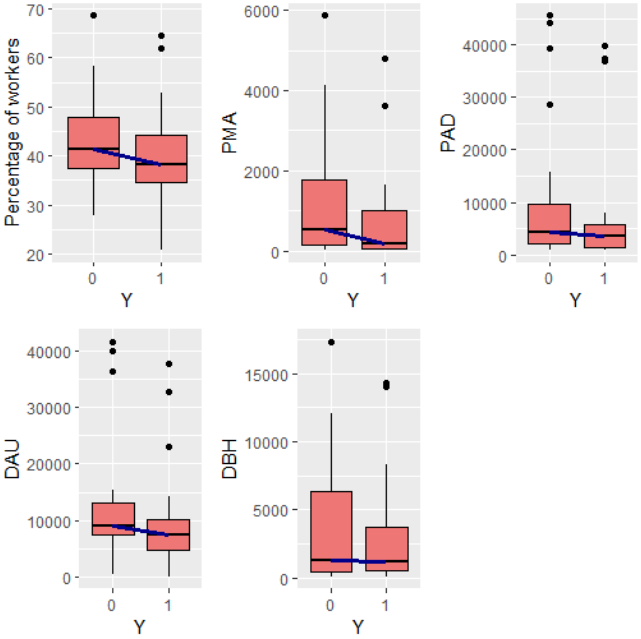

Figure 1 The characteristic of Economic Predictor Variables

The workers and PMA describe the meaningful difference in the two conditions. The lower number of workers is leading to the economic decline. The amount of PMA also indicated a similar conclusion, the province with the lower PMA are tend to experienced economic decline. This characteristic is in line with the fact that the number of workers and foreign investment are able to improve the economic condition. Compared with the other variables, the DBH variable is the most powerless predictor variable in differentiating the two categories of GDP growth.

3.2 Binary Logistic Regression

Before modeling regional GDP, multicollinearity assumption must be checked on the training data between each explanatory variable $x_1, x_2, \ldots, x_5$. The multicollinearity will occur if the VIF value is above 10 stated by Schober et al. [30]. The VIF shows that all variables have less than 10 VIF value. Therefore, binary logistic regression using all explanatory variables can be continued. The modeling output of the logistic regression shown in equation (8).

$$\pi(x) = \frac{\exp(-4.551 - 0.008x_1 - 0.001x_2 + 7.517x_{6(1)})}{1 + \exp(-4.551 - 0.008x_1 - 0.001x_2 + 7.517x_{6(1)})} \tag{7}$$

The model was transform into a natural log of the odds ratio shown in equation (9).

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = -4.551 - 0.008x_1 - 0.001x_2 + 7.517x_{6(1)} \tag{8}$$

Based on the simultaneous Likelihood Ratio Test, the value of G with a significance level (sig) 0.000 means that H_0 is rejected. It means that at least one component affects the response variable. The Wald Chi-Square statistics were used as the test for the unique contribution of each explanatory variable. The test shows that only two of six explanatory variables significantly affect the response variable. PMA and existing cases of COVID-19 have a significance value of 0.089 and 0.002 or less than 0.1 as standards for statistical significance. The logistic regression output is shown in Table 2.

Table 2 Logistic Regression Output

Variables	B	Sig.	Exp(B)
Workers	-0.008	0.921	0.992
PMA	-0.001	0.089	0.999
PAD	0.000	0.910	1.000
DAU	0.000	0.458	1.000
DBH	0.000	0.479	1.000
COVID-19(1)	7.517	0.002	0.002
Constant	-4.551	0.295	0.011

The accuracy value of the classification training data logistic regression model is 95.83 percent. Respectively, the accuracy of testing data and miss classification values were 95 percent and 5 percent. The 0.999 odds ratio for PMA indicates that for each one-point increase, the odds of the recession of the regional GDP increase by a multiplicative factor of 0.999. In summary, binary logistic regression results show that existing cases of COVID-19 are related to regional GDP status. The likelihood of a regional GDP is experiencing a recession positively related to existing cases of COVID-19.

3.3 Neural Network

In neural network modeling, we used the logistic sigmoid as activation function, standardized preprocessing for the input variables, one hidden layer consisting of several neurons. We replicated the architecture using different initial values ten times. The result of each replication on the different number of neurons is shown in Table 3.

Table 3 The Accuracy for Neural Network Models

Data	Rep	Neuron						
		1	2	3	4	5	10	15
Train	1	96%	100%	98%	100%	100%	100%	100%
	2	98%	98%	100%	98%	100%	100%	100%
	3	98%	96%	98%	100%	100%	100%	100%
	4	98%	98%	100%	100%	100%	100%	100%
	5	96%	100%	100%	100%	100%	100%	100%
	6	98%	98%	100%	100%	**100%**	100%	100%
	7	94%	98%	100%	98%	100%	100%	100%
	8	98%	96%	98%	100%	100%	100%	100%
	9	96%	98%	96%	100%	100%	100%	100%

	10	96%	100%	100%	100%	100%	100%	100%
Test	1	95%	95%	75%	90%	90%	100%	85%
	2	95%	95%	85%	85%	90%	90%	95%
	3	90%	95%	90%	90%	90%	90%	90%
	4	90%	85%	90%	90%	80%	90%	85%
	5	95%	90%	90%	85%	95%	85%	85%
	6	90%	85%	85%	95%	**100%**	95%	85%
	7	95%	95%	90%	95%	90%	85%	85%
	8	90%	95%	85%	95%	90%	90%	85%
	9	95%	90%	65%	90%	90%	90%	95%
	10	95%	95%	95%	80%	90%	85%	90%

Table 3 exhibits that the more neurons used in the model will result in the higher accuracy for training data. Architecture with more than five neurons results in the perfect accuracy for all replication in training data. This common issue can lead to over-fitting problems. Thus we select the best architecture by considering the best accuracy in testing data with the most parsimonious architecture. The highest accuracy for testing data, with the smallest number of neurons, is obtained from the five neurons in the 6th replication. Then, we analyze the feature importance of the best architecture using permutation feature importance, as in Figure 2.

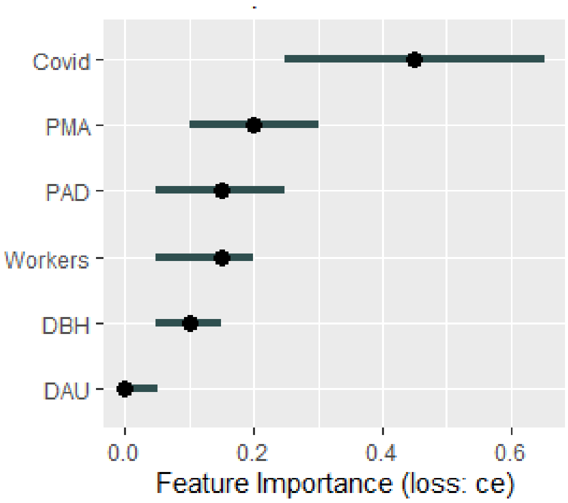

Figure 2 Feature Important Neural Network

The most important feature in predicting the regional GDP growth is the incident of COVID-19. We have Foreign Direct Investment (PMA) as the most important feature in macroeconomic variables. The importance of DAU is close to zero, which means that permuting the value of this variable was not affected the classification accuracy of the model.

3.4 Random Forest

The parameter configurations for the random forest model are obtained by combining the number of trees and the number of subset variables used in each tree. We tune the number of trees for 500, 1000, and 2000 as Statnikov suggested [31]. The accuracy of each tree using each number of variables is described by Table 4.

Table 4 The Accuracy for Random Forest Model

Data	Number of Variable	Tree		
		500	1000	2000
Train	1	100%	100%	100%
	2	100%	100%	100%
	3	100%	100%	100%
Test	1	95%	95%	95%
	2	95%	95%	95%
	3	95%	95%	95%

Since all the combinations of the number of trees and the number of variables result in identical accuracy, we select the 500 trees. The number of variables we used is two since Statnikov is suggested the number of subset variables is equal to the square root of the total number of predictor variables [31]. Using this configuration, we can extract the variable importance through the mean decrease accuracy for all variables.

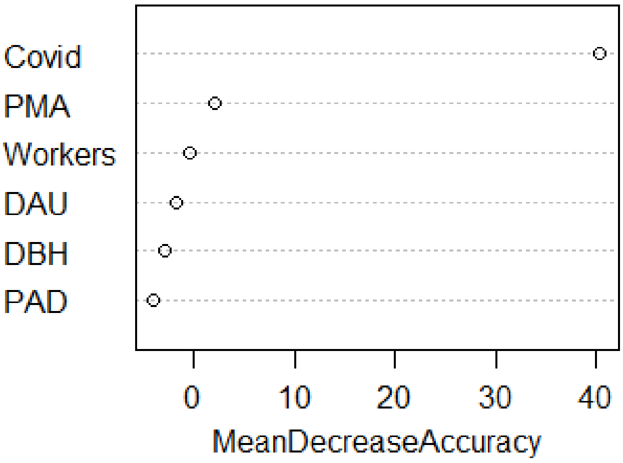

Figure 3 Mean Decrease Accuracy of Random Forest Model COVID

A huge decrease in accuracy exists for the COVID-19 variable. In addition, the most important economic predictor variable is PMA, which can decrease the accuracy by 0.5% when the value of this predictor variable is permuted.

3.5 Support Vector Machine

We use linear, polynomial, radial basis function, and sigmoid kernels in SVM modeling. In each kernel, parameter optimization is performed using a grid search. Table 5 shows the accuracy of each kernel with the most optimal parameters.

Table 5 The Accuracy of SVM With Different Kernel

Kernel	**Parameter**	**Accuracy**	
		Train	**Test**
Linear	Cost : 0.1	93.75%	95%
Polynomial	Cost :100, Degree : 3	**100%**	**95%**
RBF	Cost : 5	97.92%	95%
Sigmoid	Cost : 1	93.75%	95%

Table 5 revealed that the best kernel in SVM to separate the regional GDP class into two classes is a polynomial kernel, with accuracy in train data is perfect and in the test data 95%. Them, we conduct the feature importance of the best kernel result using permutation feature importance, as in Figure 4. Using SVM with kernel Polynomial, the essential feature in predicting the GDP growth is the COVID-19 pandemic. The other features have a negligible impact showing that the importance is close to zero, which means that permuting the value of this variable did not affect the model's classification accuracy.

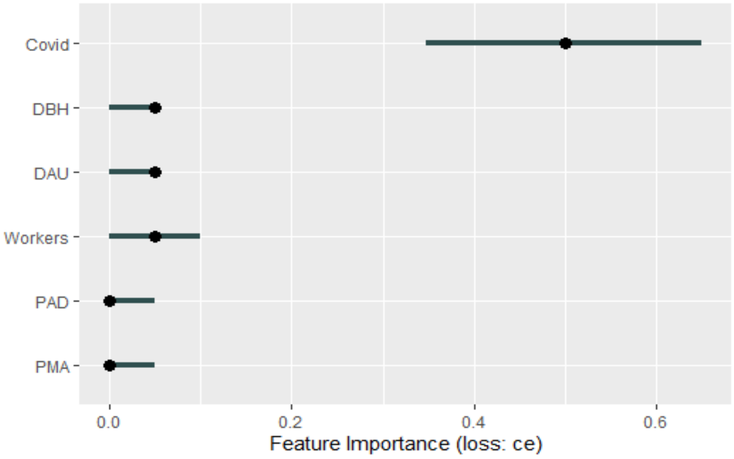

Figure 4 Feature Important SVM Using Kernel Polynomial

3.6 Bayesian Model Averaging

The BMA approach considers all possible combinations of sixvariables. Thus, regardless of the interaction between variables, there are $2^6 = 64$ models to predict the classes of regional GDP. The posterior probabilities of $Pr(b_i \neq 0 | D)$, $E(b | D)$ and $SD(b | D)$ of the remaining predictor variables in the last iteration are shown in Table 6.

The Table 6 column $Pr(b_i \neq 0 | D)$ shows the posterior probability that the coefficient is not equal to zero. Whereas $E(b | D)$ shows the posterior mean of the coefficient or the value we expect in the BMA model, the posterior standard deviation shows the posterior SD, or standard deviation, giving a measure of the coefficient of variability. Based on Table 6, the highest posterior probability are COVID-19 and PMA. Variables with a high posterior probability of $Pr(b_i \neq 0 | D)$ are important variables for predicting regional GDP. Variable COVID-19 is a variable with the highest posterior probability value so that the COVID-19 can be said to be the most crucial variable.

Table 6 Posterior Probability Selected Variables

Predictor	Pr(b_i ≠ 0 | D) (%)	*E*(b|D)	SD (b|D)
Intercept	100	-3.20	1.436
Workers	6.9	$-1.776 \cdot 10^{-3}$	$1.709 \cdot 10^{-2}$
PMA	26.6	$-2.126 \cdot 10^{-4}$	$4.823 \cdot 10^{-5}$
PAD	13.1	$7.453 \cdot 10^{-6}$	$3.232 \cdot 10^{-5}$
DAU	13.9	$9.306 \cdot 10^{-6}$	$3.366 \cdot 10^{-5}$
DBH	10.2	$9.716 \cdot 10^{-6}$	$5.663 \cdot 10^{-5}$
COVID-19$_{(1)}$	100	5.781	1.437

The posterior probability of the variable is obtained by summing up the posterior probability of the model for each variable included in the model. Table 7 again confirms that the COVID-19 is a predictor in almost all models. It shows that the COVID-19 results in a highly variable posterior probability because it is included in many models. The results of selecting the best model are presented in Table 7.

Table 7 Selected Model Using BMA

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Intercept	-3.135	-2.774	-4.084	-3.447	-2.074
Workers	-	-	-	-	-0.0257
PMA	-	$-5.114 \cdot 10^{-4}$	-	-	-
PAD	-	-	-	-	-
DAU	-	-	$5.571 \cdot 10^{-5}$	-	-
DBH	-	-	-	$6.76 \cdot 10^{-5}$	-
COVID-19$_{(1)}$	5.533	5.809	6.032	5.676	5.543
n	1	2	2	2	2
BIC	-156	-153.4	-152.9	-152.3	-152.3

PMP	0.437	0.122	0.092	0.071	0.069

Table 7 shows the model with the highest posterior model probability (PMP) of only 43.7% out of the total posterior probability, indicating that the model's uncertainty is quite high. Model 1 with PMP 0.437 indicates that Model 1 contributes 43.7% of the total posterior probability. Likewise, model 2 contributes 12.2% of the total posterior probability.

In term of contribution of each predictor variable, the COVID-19 contributes to five selected models so that it has an enormous influence on the response variable. Therefore, the COVID-19 has a significant posterior probability. PMA has the second-largest contribution compared to other variables even though it only appears in model 2. While other variables only contribute to one model with low PMP, meaning that the influence of these variables are quite small. Based on the value of Bayesian Information Criterion (BIC), the first model and the second model are the models that have the smallest BIC, indicating that the first and the second models fit better than other models. BIC shows the goodness of fit of a model. The smaller the BIC value, the better the model formed.

The visualization of variables and selected models using the BMA is shown in Figure 5. In Figure 5, the selected variable by BMA is shown on the vertical axis, and the selected BMA model is displayed on the horizontal axis. Models are sorted in order based on the largest to smallest posterior model (PMP) probability from left to right. The heatmap shows that the COVID-19 as a predictor variable in all models. The blue color indicates that the coefficient is negative, and the red color indicates the coefficient in the model is positive. The COVID-19 variable has a red color, meaning that the COVID-19 variable makes predictions towards class 1, which in this case means a slowdown in regional GDP.

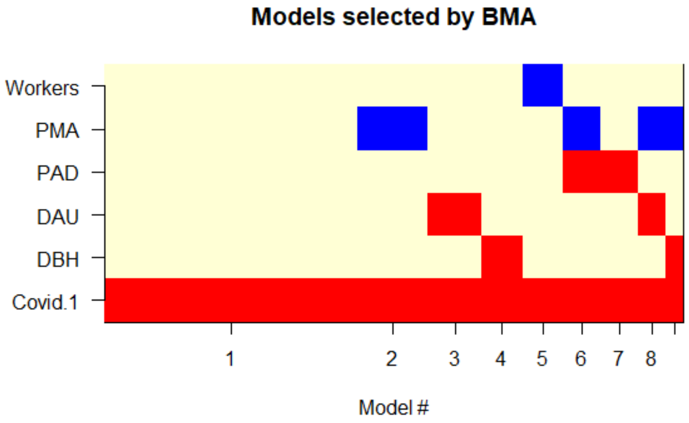

Figure 5 Selected Variables and Models with BMA

3.7 Comparison

Model selection in predicting the growth of regional GDP is based on the best accuracy in testing data. In case the accuracy results in the same performance, we also consider the accuracy of the training data.

Table 8 The Accuracy for All Methods

Method*	Accuracy		Variable Importance
	Train	Test	
LR	95.83%	95.00%	COVID, PMA
BMA	93.75%	95.00%	COVID, PMA
SVM	**100.00%**	95.00%	COVID
NN	**100.00%**	**100.00%**	COVID, PMA
RF	**100.00%**	95.00%	COVID, PMA

*LR=Logistic Regression, BMA= Bayesian Model Averaging, SVM= Support Vector Machine, NN=Neural Network, RF=Random Forest

As in Table 8, the prediction accuracy of all five classifiers is quite impressive across the board, with all five of them classifying the regional GDP correctly over 90%. The most significant variables are the occurrence of COVID-19 and PMA as an economic variable. This result is in line with the previous research conducted by Muchisha et al. [15], Istiqomah et al. [16], and Anisah [18]. It was also suitable with the data exploration, where the PMA variable is able to differentiate two classes in GDP growth. The higher PMA of a province, the lower probability the province would experience an economic decline. The neural network has a slightly higher training and testing data accuracy than the other four contenders, which can perfectly classify regional GDP growth for training and testing data. The important variables in the neural network model are also consistent with other methods, that is, COVID-19 and PMA. Therefore, the neural network models have satisfying performance and appropriately explain the important variable. The machine learning methods are able to be interpreted through permutation feature importance. Supported by the excellent performance, the machine learning method is promising to be applied in various fields.

4. Conclussion

This section involves describing the results obtained from the research and drawing similarities and differences between the research and previous others from methods, data, and results.

This study employed several classical and machine learning approaches to predict regional GDP growth. The regional GDP growth is classified into two classes. The first class is the condition of positive regional GDP growth, which implies the improvement in regional GDP. The second class is the condition of negative regional GDP growth, which means the decline in the economy. The methods used to classify regional GDP growth are binary logistic regression, bayesian model averaging, support vector machine, neural network, and random forest. We found that all the selected machine learning models are able to classify the regional GDP growth perfectly for the training data. The neural network model outperforms the other methods with an accuracy of 100% in training and testing data. We can get the important variable directly from the parameter estimation process in the classical methods. In addition, the permutation process provides the important variables in machine learning methods. The important variable for all methods provides a similar result. COVID-19 and the PMA variable are the most important variables in predicting regional GDP growth.

In conclusion, machine learning methods are able to be interpreted through the permutation process with the appropriate results. Supported by the excellent performance, the machine learning method is promising to be applied in various fields. However, we are not able to test the statistical significance of the input variable to the output variable. Further research relating to interpretable machine learning, such as feature interaction, global surrogate, and Shapley values, is also necessary to predict regional GDP growth using machine learning methods [32]. The use of fuzzy algorithm can also be applied to the machine learning methods [33], [34]. In addition, the application of big data, including time series dataset along with the appropriate methods can be promising [35]–[37].

However, describe whether the problems have been researched successfully according to the objectives using the proposed methods. This should involve the description of the analysis conducted, cause and benchmark of success/failure, and the unfinished part of the research followed with the steps to be taken as follow up process.

References

[1] J. Jelina, R. Sasikala, and M. Phil, “Multi Class Classification Methods on Data Analysis Using Data Mining Techniques,” *IOSR J. Eng.*, no. Section 2, pp. 1–4, 2018, [Online]. Available: www.iosrjen.org.

[2] R.-M. Ştefan, “A Comparison of Data Classification Methods,” *Procedia Econ. Financ.*, vol. 3, no. 12, pp. 420–425, 2012, doi: 10.1016/s2212-5671(12)00174-8.

[3] A. Tharwat, “Classification assessment methods,” *Appl. Comput. Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.

[4] K. Kumar and K. Chaturvedi, “An audio classification approach using feature extraction neural network classification approach,” *Int. Conf. Data, Eng. Appl.*, no. 2, pp. 1–6, 2020, doi: 10.1109/IDEA49133.2020.9170702.

[5] M. Yang, S. Cui, Y. Zhang, J. Zhang, and X. Li, “Data and Image Classification of Haematococcus pluvialis Based on SVM Algorithm,” *China Autom. Congr.*, pp. 522–525, 2021, doi: 10.1109/CAC53003.2021.9727433.

[6] J. V. Rissati, P. C. Molina, and C. S. Anjos, “Hyperspectral Image Classification Using Random Forest and Deep Learning Algorithms,” *Lat. Am. GRSS ISPRS Remote Sens. Conf.*, pp. 132–132, 2020, doi: 10.1109/lagirs48042.2020.9165588.

[7] P. Suksomboon and A. Ritthipakdee, “Performance Comparison Classification using k-Nearest Neighbors and Random Forest Classification Techniques,” *Int. Conf. Big Data Anal. Pract.*, pp. 43–46, 2022, doi: 10.1109/IBDAP55587.2022.9907218.

[8] B. Sasongko, S. Bawono, and B. H. Prabowo, “The Economic Performance of China in Trade War: The Case Study of Three Global Economic Crises in 1997–2020,” *Environ. Soc. Gov. Perspect. Econ. Dev. Asia*, vol. 29B, pp. 1–11, 2021.

[9] A. B. Abel, B. Bernanke, and D. Croushore, *Macroeconomics*. Addison Wesley Longman. Inc, 2005.

[10] N. Oktaviana and N. Amalia, “Gross Regional Domestic Product Forecasts Using Trend Analysis: Case Study of Bangka Belitung Province,” *JESP*, vol. 19, no. 2, pp. 142–151, 2018.

[11] R. Nuraini, “Kasus Covid-19 Pertama, Masyarakat Jangan Panik,” *Portal Informasi Indonesia*, p. https://indonesia.go.id/narasi/indonesia-dalam-ang, 2020.

[12] R. D. A. Saptoyo, “Update Corona Dunia 5 Juli: 184 Juta Kasus Covid-19 | Angka Kematian akibat Tak Vaksinasi,” 2021. https://www.kompas.com/tren/read/2021/07/05/112600765/update-corona-dunia-5-juli--184-juta-kasus-covid-19-angka-kematian-akibat?page=all (accessed Jul. 10, 2021).

[13] D. Wuryandani, “Dampak Pandemi COVID-19 Terhadap Pertumbuhan Ekonomi Indonesia 2020 dan Solusinya,” 2020.

[14] H. F. Nasution, Z. Matondang, K. Nasution, and D. W. Nasution, “The Role Of Leading Sector Labor On The GRDP Of North Sumatera Province,” *Al-Masharif J. Ilmu Ekon. dan Keislam.*, vol. 9, no. 1, pp. 76–92, 2021.

[15] N. Tamara, N. D. Muchisha, Andriansyah, and A. M. Soleh, “Nowcasting Indonesia’s GDP Growth Using Machine Learning Algorithms,” *IJSA*, vol. 5, no. 2, pp. 355–368, 2021.

[16] Istiqomah, A. A. Wibowo, E. Yunianti, and D. S. Gunawan, “Determinants of Gross Regional Domestic Product in Eastern Indonesia Region,” *Trikonomika*, vol. 18, no. 1, pp. 18–24, 2019.

[17] L. Sri and U. M. Suyana, “The Effect of Local Government Own Revenue and Revenue Sharing Funds on Econoomic Performance and Community Welfare Throught Capital Expenditure of Regency/City in Bali Province, Indonesia,” *RJOAS*, vol. 7, no. 91, pp. 67–87, 2019.

[18] N. Anisah, “Investment Development Before and During the Covid-19 Pandemic and Impact on Regional Economy in West Java,” *J. Ekon. Pembang.*, vol. 19, no. 2, pp. 81–96, 2021.

[19] Catherine, L. L. Lilyana, S. H. Selvia, and Y. N. Malau, “Pengaruh Investasi, Tenaga Kerja, dan Ekspor Terhadap PDB di Provinsi Sumatera Utara Periode 2017-2019,” *J. Ilm. MEA (Manajemen, Ekon. Akuntansi)*, vol. 4, no. 3, pp. 1711–1721, 2020.

[20] W. Susanto and C. Sugianto, “The Structure of Regional Original Revenue and Its Effect on Economic Growth: Facts from Regencies and Cities in Central Jawa,” *Indones. J. Dev. Plan.*, vol. 3, no. 1, pp. 68–102, 2019.

[21] A. Agresti, *An Introduction to Categorical Data Analysis*, 3rd ed. USA: New Jersey: Wiley, 2019.

[22] Suhartono, P. D. Saputri, F. F. Amalia, D. D. Prastyo, and B. S. S. Ulama, “Model Selection in Feedforward Neural Networks for Forecasting Inflow and Outflow in Indonesia.pdf,” *Commun. Comput. Inf. Sci.*, vol. 788, pp. 978-981-10-7242–0, 2017.

[23] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.

[24] V. N. Vapnik, *The Nature of Statistical Learning Theory 2nd Edition*. New York: Springer, 2000.

[25] T. de O. Nogueira *et al.*, “Imbalance classification in a scaled-down wind turbine using radial basis function kernel and support vector machines,” *Energy*, vol. 238, no. C, p. 122064, 2022.

[26] S. Abe, *Support Vector Machines for Pattern Classification*. London: Springer, 2010.

[27] M. Hinne, Q. F. Gronau, D. van den Bergh, and E. J. Wagenmakers, “A Conceptual Introduction to Bayesian Model Averaging,” *Adv. Methods Pract. Psychol. Sci.*, vol. 3, no. 2, pp. 200–215, 2020, doi: 10.1177/2515245919898657.

[28] T. M. Fragoso, W. Bertoli, and F. Louzada, “Bayesian Model Averaging: A Systematic Review and Conceptual Classification,” *Int. Stat. Rev.*, vol. 86, no. 1, pp. 1–28, 2018, doi: 10.1111/insr.12243.

[29] H. N. K. AL-Behadili and K. R. Ku-Mahamud, “Hybrid K-Nearest Neighbour and Particle Swarm Optimization Technique for Divorce Classification,” *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 11, no. 4, pp. 1447–1454, 2021.

[30] P. Schober and L. A. Schwarte, “Correlation coefficients: Appropriate use and interpretation,” *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, 2018, doi: 10.1213/ANE.0000000000002864.

[31] A. Statnikov, L. Wang, and C. F. Aliferis, “A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification,” *BMC Bioinformatics*, vol. 9, pp. 1–10, 2008, doi: 10.1186/1471-2105-9-319.

[32] C. Molnar, *Interpretable machine learning” in A Guide for Making Black Box Models Explainable*. 2022.

[33] R. Kanagaraj, E. Elakiya, N. Rajkumar, K. Srinivasan, and S. Sriram, “Fuzzy Neural Network Classification Model for Multi Labeled Electricity Consumption Data Set,” *Int. Conf. Smart Syst. Inven. Technol.*, pp. 1037–1041, 2022, doi: 10.1109/icssit53264.2022.9716415.

[34] G. Liu, L. Wang, L. Fei, D. Liu, and J. Yang, “Hyperspectral Image Classification Based on Fuzzy Nonparallel Support Vector Machine,” *Glob. Conf. Robot. Artif. Intell. Inf. Technol.*, pp. 242–246, 2022, doi: 10.1109/GCRAIT55928.2022.00058.

[35] K. Djouzi, K. Beghdad-Bey, and A. Amamra, “A new adaptive sampling algorithm for big data classification,” *J. Comput. Sci.*, vol. 61, no. March, p. 101653, 2022, doi: 10.1016/j.jocs.2022.101653.

[36] T. Altay and M. G. Baydoğan, “‘A new feature-based time series classification method by using scale-space extrema,’” *Eng. Sci. Technol. an Int. J.*, vol. 24, no. 6, pp. 1490–1497, 2021, doi: 10.1016/j.jestch.2021.03.017.

[37] N. Li and B. Zhang, “Research on big data classification method based on improved KNN algorithm,” *Int. Conf. Comput. Inf. Big Data Appl.*, pp. 827–830, 2022.