Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 4 No. 1 2024, 9 - 15
DOI: *https://doi.org/10.22487/27765660.2024.v4.i1.17055*
eISSN: 2776-5660

APPLICATION OF THE LIGHTGBM ALGORITHM IN THE CLASSIFICATION OF GREENHOUSE GAS EMISSIONS

Rini Latifah[1*], Gustriza Erda[2]
[1,2]Statistic Study Program, Riau University

***e-mail**: [1]rini.latifah1099@student.unri.ac.id, [2]gustrizaerda@lecturer.unri.ac.id*

ABSTRACT

There are many negative impacts that can result from increasing greenhouse gas emissions. The negative impact that can result from increasing greenhouse gas emissions is global warming, which also has an impact on various areas of life, such as drought and rising sea levels. Therefore, it is important to know the level of greenhouse gas emissions in the future, which can be done by making predictions, so that policy planning can be carried out to reduce the impact. In this study, the classification of greenhouse gas emission levels was carried out using the lightGBM method. The aim is to see the performance of the lightGBM method in classifying greenhouse gas emissions. The data used consists of 39 independent variables that influence climate change in the world and the dependent variable is total greenhouse gas emissions. The results obtained from this research were an accuracy of 98.72%, a sensitivity of 99.42%, a specificity of 97.49%, and a MAE of 0.0128. Based on the accuracy, sensitivity, specificity, and MAE values, it can be concluded that the lightGBM method has good performance in classifying greenhouse gas emissions.

Keywords: *Classification, LightGBM, Greenhouse Gas Emissions, Data Normalization, MICE*

Cite: Latifah, R., & Erda, G., (2023). *Application of The LightGBM Algorithm in The Classification of Greenhouse Gas Emissions. Parameter: Journal of Statistics*, *4*(1), 9-15, https://doi.org/10.22487/27765660.2024.v4.i1.17055.

[*]Received: Apr 2024; Reviewed: Apr 2024; Published: Jun 2024

INTRODUCTION

Greenhouse gas (GRK) is one of the factors causing climate change due to rising temperatures on the Earth's surface. This is because greenhouse gases have binding properties and emit infrared radiation from the sun's rays (Wahyudi, 2019).The higher the GRK concentrations in the atmosphere, the greater the amount of infrared radiation trapped in the air, which promotes global warming (Yoro & Daramola, 2020). The global warming that is happening will have a negative impact on sectors like agriculture, tourism, and so on.

As many impacts are caused by these increased greenhouse gas emissions, measuring the level of greenhouse gas emissions can give an idea of what steps can be taken to reduce the impact by making predictions. One of the predictive methods that can be used is classification. Classification helps in categorizing new data based on models formed from previous data (Goldameir et al., 2021). One of the classification methods that can be used is the lightGBM method, which involves the development of gradient boosting. Gradient boosting follows the process boosting approach, which is to combine several weak machine learning models to acquire a powerful machine learning model. Gradient boosting algorithms are used for learning processes in classification and regression tasks.

LightGBM research on diabetes patients data collected from Zewditu Memorial Hospital (ZMHDD) in Addis Ababa, Ethiopia, in 2019. The use of LightGBM as one of the methods of gradient boosting has a low computational complexity that is suitable for use in regions with limited capacity, such as Ethiopia. The results of this study show that lightGBM outperforms the KNN, SVM, Naïve Bayes, Bagging, Random Forest, and XGBoost methods in terms of accuracy, AUC, sensitivity, and specificity (Rufo et al., 2021).

Previous research was also conducted by Adnan et al (2023) on climate change data with the aim of conducting classification on the variable total greenhouse gas emissions (kt of CO2 equivalent) using logistical regression. The results obtained from the study resulted in a good accuracy rate of 87.60%. Seeing the accurate value that can still be improved, this study will use a different method, namely the LightGBM method, in classifying the total greenhouse gas emissions.

The study will use the lightGBM method to classify the level of greenhouse gas emissions. The data used in this study has a considerable amount of data, so this method was chosen because it has good speed in the analysis process. Moreover, this method can also be used on imbalanced data, where the data dependencies used in the study are imbalanced data because they have different class ratios. The aim of this study is to see the goodness of the lightGBM method in the classification of greenhouse gas emissions.

MATERIALS AND METHODS

The data in this study is secondary data taken from the site https://data.worldbank.org/ (The World Bank, 2022). Contains 39 variables and 8215 lines of data that affect climate change around the world. The dependent variable in this study is the total greenhouse gas emissions (kt of CO2 equivalent), given the symbol Y. In the classification of the variable Y based on the threshold value of greenhouse emissions in 2020, it is $431\ \mathrm{MMTCO_2e}$ obtained from California Greenhouse Gas Emission Inventory Program. As for the classification rule, if the value of the variable Y is below the greenhouse gas emission limit, then the value is 0 (low class), and the value is 1 if Y is above the greenhouse gas emission threshold (high class). The method used in this research is the lightGBM method, using the google collaboratory.

LightGBM, abbreviated for light gradient boosting, is one of the developments of gradiant boosting that uses a decision-tree-based learning algorithm. LightGBM algorithms are efficient in big data training and have good performance in terms of speed (Zeng et al., 2019). These algorithms are also used in solving problems related to classification, regression, and classification. The lightGBM model is obtained by minimizing the boosting loss function based on the gradient decrease algorithm (Zhang et al., 2019). Each new model is added, and the loss function continues to decrease to obtain a variable gradient with the highest information content. In addition to minimizing loss functions and implementing gradient decrease, lightGBM has two main features: the leaf-wise tree growth method and the application of histogram-based decision tree algorithms. These two main features can effectively handle large-scale data.

According to Guolin ke (2017) algorithm LightGBM has a data training speed that is 20 times faster than the conventional gradient-boosting decision tree. This method uses gradient-based one-sided sampling (GOSS) and exclusive feature bundling (EFB) techniques to handle large-scale data so that it can increase the speed and efficiency of training algorithms. GOSS works using data that has a larger gradient, so it can produce fairly accurate predictions. EFB is a method used to reduce the number of variables by combining variables that do not interfere with each other into a single bundle (Ke et al., 2017).

LightGBM is a machine learning library with forming algorithms based on the gradient-boosting decision tree that has been developed to have higher speeds. The LightGBM library can be installed in Python or Google Collaboratory with the pip-python package. The analysis steps with lightGBM method using google collaboratory are as follows.

1. Collect climate change-related data from the BMKG and Worlbank web.
2. Do data pre-processing aimed at data cleaning. In the datasets used there are lots of empty data, so the handling used is by doing imputations to the data. The method used is Multivariate Imputation by Chained Equations (MICE). The MICE imputation process begins with the input of data into the Google collaboratory. MICE imputation is performed using the IterativeImputer() function found in the sklearn.impute package. Subsequently, the value of the model is measured based on the resulting RMSE value. The value of RMSE is calculated using the evaluation of the dual linear regression model.
3. Verification of multicolinearity of data using Variance Inflation Factor (VIF) values A VIF value greater than 10 indicates that the variable is multicolinear. VIF values from variables that are multicolinear will be deleted or not included in subsequent analysis. The VIF value can be calculated using the following equation (Gujarati & Porter, 2009).

$$\mathrm{VIF} = \frac{1}{(1 - \mathrm{R}_\mathrm{i}^2)} \quad (1)$$

4. Descriptive statistics describe the characteristics of the empirical data used.
5. Divide the data into 80% training data and 20% testing data.
6. Data normalization using a min-max scaler. Data normalisation is done so that the data used has the same scale so the algorithm used can run more effectively. The min-max normalization equation is as follows (Li & Liu, 2011).

$$\mathrm{V}' = \frac{(\mathrm{v} - \mathrm{v}_{\mathrm{min}})}{(\mathrm{v}_{\mathrm{max}} - \mathrm{v}_{\mathrm{min}})} \quad (2)$$

7. Modeling is carried out using lightGBM with binary classification, then making predictions on testing data and calculating accuracy, sensitivity, specificity, and MAE values
8. Conclusions.

Evaluation of model performance can be done using a confusion matrix. Confusion Matrix consists of rows and columns Where rows are the actual class and column is the prediction class (Caelen, 2017).

Table 1. Confusion Matrix for Two Class

		Prediction Result Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Negative (FP)	True Negative (TN)

The confusion matrix is used to evaluate the results of classifications using the confusion matrix by calculating the values of accuracy, sensitivity, and specificity. The percentage accuracy of the overall prediction is indicated by the accurate value. Sensitivity indicates the percentage accuracy of the prediction in the positive class, whereas specificity shows the percentage accuracy of the forecast in the negative class (Istiana & Mustafiril, 2023). Here's the classification performance evaluation formula.

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP})} \tag{3}$$

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{(\mathrm{TP} + \mathrm{FN})} \tag{4}$$

$$\mathrm{Spesifisity} = \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \tag{5}$$

In addition to using accuracy, sensitivity, and specificity measurement goodness models can also be used with, Mean Absolute Error (MAE). Suryanto et al (2019) explains that the value of the MAE calculates the difference between the predicted outcome and the actual value so that the absolute error averages or mean absolute errors are obtained.

$$\mathrm{MAE} = \frac{1}{\mathrm{n}}\sum\nolimits_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{y_i} - \mathrm{x_i}| \tag{6}$$

With,

y_i : Predicted value
x_i : Actual value
n : Total of data

RESULTS AND DISCUSSION

Data Cleaning

The data used in this study shows that each variable has a missing value, so the data is imputed. The imputation used is multivariate imputation by chained equations (MICE), which obtained an RMSE value of 129.10^{-5}. Judging from the relatively small RMSE value, it shows that this imputation is good to use.

Feature Selection

Then carry out feature selection by calculating the VIF value in equation (1) for each variable, i.e., if VIF is greater than 10, then the variable is deleted. Based on VIF value checking, the data that originally consisted of 39 independent variables became only 19 independent variables used for further analysis. The 19 independent variables are as follows.

Table 2. Independent Variables Research

Symbol	Variable Description
X_1	Alternative and nuclear energy (% of total energy use)
X_2	CO2 emissions (metric tons per capita)
X_3	Energy imports, net (% of
X_4	Fossil fuel energy consumption (% of total)
X_5	GDP growth (annual %)
X_6	GDP per capita (current US$)
X_7	Ores and metals exports (% of merchandise exports)
X_8	Oil rents (% of GDP)
X_9	Natural gas rents (% of GDP)
X_{10}	Mineral rents (% of GDP)
X_{11}	Forest rents (% of GDP)
X_{12}	Coal rents (% of GDP)
X_{13}	Adjusted savings: net forest depletion (current US$)
X_{14}	CO2 emissions (kg per 2015 US$ of GDP)
X_{15}	CO2 emissions from other sectors, excluding residential

X_{16}	CO2 emissions from manufacturing industries and construction (% of total fuel co)
X_{17}	CO2 emissions from electricity and heat production, total (% of total fuel combu
X_{18}	CO2 emissions from residential buildings and commercial and public services (% o
X_{19}	CO2 intensity (kg per kg of oil equivalent energy use)

Min-Max Normalization

Normalization is carried out to generalize the range in the data. The aim of normalizing data is to speed up the learning process in the analysis process. In the normalization method, the data size will be in the range of 0 to 1. Min-max normalization has the advantage of maintaining the relationship of data values when scale changes are made. The min-max normalization is carried out using equation (2).

Descriptive Statistics

Each variable in Table 2 has an outlier value. These outlier values were not deleted because they may contain important information. The total outliers in each variable are shown in Table 3 below.

Table 3. Total Outlier in Each Variable

Symbol	Variable Description	Total Outlier
X_1	Alternative and nuclear energy (% of total energy use)	511
X_2	CO2 emissions (metric tons per capita)	366
X_3	Energy imports, net (% of	765
X_4	Fossil fuel energy consumption (% of total)	109
X_5	GDP growth (annual %)	623
X_6	GDP per capita (current US$)	692
X_7	Ores and metals exports (% of merchandise exports)	866
X_8	Oil rents (% of GDP)	1082
X_9	Natural gas rents (% of GDP)	1041
X_{10}	Mineral rents (% of GDP)	823
X_{11}	Forest rents (% of GDP)	967
X_{12}	Coal rents (% of GDP)	1208
X_{13}	Adjusted savings: net forest depletion (current US$)	790
X_{14}	CO2 emissions (kg per 2015 US$ of GDP)	514
X_{15}	CO2 emissions from other sectors, excluding residential	478
X_{16}	CO2 emissions from manufacturing industries and construction (% of total fuel co	721
X_{17}	CO2 emissions from electricity and heat production, total (% of total fuel combu	403
X_{18}	CO2 emissions from residential buildings and commercial and public services (% o	423
X_{19}	CO2 intensity (kg per kg of oil equivalent energy use)	288

Table 3 shows that there is an outlier value for each variable. The outlier value range of the variable used is 109–1208 data points. The highest number of outliers is on the coal rents variable (% of GDP), which is as much as 1208; the lowest number is on the fossil fuel energy consumption variable (% of total), which is as many as 109.

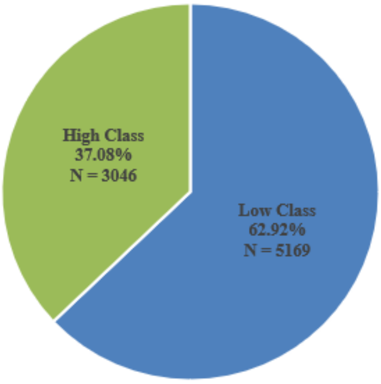

Figure 1. Percentage of dependent variable class

The dependent variable in this study consists of two classes: the lower class with the label “0” and the high class with the label "1." Based on Figure 1, it is evident that the lower and higher classes have different data ratios or data imbalances. The lower class "0" is more dominant compared to the upper class "1." The low class, after being categorized, obtains as much as 5,169 with a percentage of 62.92%, and the top class is 3046 with a percent of 37,08%. In this study, there is no data balancing because the lightGBM method is able to handle the data in the process.

Results of Prediction, Accuracy, Sensitivity, Specificity, and MAE

When making predictions using testing data, the results are shown in the following confusion matrix table.

Table 4. Confusion Matrix

		Prediction class	
		0	1
Actual class	0	1039	6
	1	15	583

Table 4 shows that there are classification errors in the lower class (“0”). There are 6 data that are supposed to be in the “0” class but are incorrectly classified in class “1.” Prediction error also occurred in class “1”, that is, 15 data were classified in class “0.” Evaluate model performance using equation (3), equation (4), equation (5), and equation (6) respectively to calculate the value of accuracy, sensitivity, specificity, and MAE.

Table 5. Accuracy, Sensitivity, Spesifisity, dan MAE

Accuracy (%)	**Sensitifity (%)**	**Spesifisity (%)**	**MAE**
98,72	99,42	97,49	0,0128

Table 5 The sensitivity value obtained is 99.42%. The sensibility value of this lightGBM model predicted the data in the low class (“0”). This indicates that the percentage level of prediction accuracy in the positive class has a good classification system. The specification value received was 97.49%. This value indicates the accuracy of the lightGBM model in correctly predicting the data for the high class (“1”). The accuration value gained by applying the lightGBM model was an accurate value of 98.72%, meaning that this model is well applied in the classification of greenhouse gas emission data. Apart from that, the MAE value was also used to assess the performance of lightGBM, the result obtained were 0.0128.

CONCLUSION

Based on the results of the analysis that has been carried out, the application of the lightGBM method to the classification of greenhouse gas emission data obtained the accuracy, sensitivity, specificity, and MAE values resulting from the process of optimization of parameters on the sequential classification of greenhouse gas emissions of 98.72%, 99.42%, 97.49%, and 0.0128. Based on the results of the four performance evaluation values of the lightGBM method, it can be concluded that this method is good for use in classifying greenhouse gas emissions.

REFERENCES

Adnan, A., Yolanda, A. M., Erda, G., Goldameir, N. E., & Indra, Z. (2023). The comparison of accuracy on classification climate change data with logistic regression. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *8*(1), 56–61.

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3), 429–450.

Goldameir, N. E., Yolanda, A. M., & Adnan, A. (2021). Classification of the human development index in indonesia using the bootstrap aggregating method. *Sinkron: Jurnal*, *6*(1), 100–106. https://jurnal.polgan.ac.id/index.php/sinkron/article/view/11173

Gujarati, D., & Porter, D. (2009). *Basic Econometrics* (Vol. 5).

Istiana, N., & Mustafiril, A. (2023). Perbandingan metode klasifikasi pada data dengan imbalance class dan missing value. *Jurnal Informatika*, *10*(2), 101–108. https://doi.org/10.31294/inf.v10i2.15540

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149–3157.

Li, W., & Liu, Z. (2011). A method of SVM with normalization in intrusion detection. *Procedia Environmental Sciences*, *11*(PART A), 256–262. https://doi.org/10.1016/j.proenv.2011.12.040

Rufo, D. D., Debelee, T. G., Ibenthal, A., & Negera, W. G. (2021). Diagnosis of diabetes mellitus using gradient boosting machine (Lightgbm). *Diagnostics*, *11*(9). https://doi.org/10.3390/diagnostics11091714

Suryanto, A. A., Muqtadir, A., & Artikel, S. (2019). Penerapan metode Mean Absolute Error (MEA) dalam algoritma regresi linear untuk prediksi padi. *SAINTEKBU: Jurnal Sains Dan Teknologi*, *11*(1).

The World Bank. (2022). *Climate change*. https://data.worldbank.org/topic/19

Wahyudi, J. (2019). *Emisi gas rumah kaca (grk) dari pembakaran terbuka sampah rumah tangga menggunakan model ipcc greenhouse gases emissions from municipal solid waste burning using ipcc model*. *XV*(1), 65–76.

Yoro, K. O., & Daramola, M. O. (2020). CO2 emission sources, greenhouse gases, and the global warming effect. In *Advances in Carbon Capture: Methods, Technologies and Applications* (pp. 3–28). Elsevier. https://doi.org/10.1016/B978-0-12-819657-1.00001-3

Zeng, H., Yang, C., Zhang, H., Wu, Z., Zhang, J., Dai, G., Babiloni, F., Kong, W., & Chuang, L. (2019). A lightGBM-based EEG analysis method for driver mental states classification. *Computational Intelligence and Neuroscience*, *2019*.

Zhang, Z., Zhao, Y., Canes, A., Steinberg, D., & Lyashevska, O. (2019). Predictive analytics with gradient boosting in clinical medicine. *Annals of Translational Medicine*, *7*(7), 152–152. https://doi.org/10.21037/atm.2019.03.29