The Comparison of Accuracy on Classification Climate Change Data with Logistic Regression

Arisman Adnan[1)*], Anne Mudya Yolanda[2)], Gustriza Erda[3)], Noor Ell Goldameir[3)], Zul Indra[5)]
[1)2)3)4)]Department of Statistics Universitas Riau, Indonesia , [5]Department of Information System Universitas Riau, Indonesia
[1)]arisman.adnan@lecturer.unri.ac.id, [2)]annemudyayolanda@lecturer.unri.ac.id, [3)]gustrizaerda@lecturer.unri.ac.id, [4)] noorellgoldameir@lecturer.unri.ac.id, [5)]zulindra@lecturer.unri.ac.id

Submitted : Nov 9, 2022 | **Accepted** : Nov 29, 2022 | **Published** : Jan 1, 2023

Abstract: Machine learning methods can be used to generate climate change models. The goal of this study is to use logistic regression machine learning algorithms to classify data on greenhouse gas emissions. The data used is climate change data of several countries obtained from The World Bank, with total greenhouse gas emissions as the response variable and 61 other attributes as explanatory variables. This data is preprocessed using min-max normalization to handle unbalanced ranges, and then the data is split into 70% training data and 30% testing data. Based on the logistic regression modeling, it was discovered that the data from the min-max transformation resulted in better modeling than the data modeling without the transformation process. The accuracy, precision, sensitivity, and specificity of the transformation are 87.60%, 87.76%, 87.04%, and 88.14%, respectively

Keywords: Classification, climate change; logistic regression; machine learning; transformation data

INTRODUCTION

The presence of the industrial revolution has changed many aspects of human life. One of the negative consequences of industrialization that is currently being highlighted is global warming as a result of climate change. This is a critical issue because climate change is expected to worsen in the coming decades. These effects can be seen throughout the region and in many important societal sectors such as health, agriculture and food security, water supply, transportation, energy, ecosystems, and others.

Climate change, according to the National Oceanic and Atmospheric Administration (NOAA), is a long-term shift in weather statistics (including the average) (Top et al., n.d.). Climate change can be identified only after a long period of time. Many studies on climate change have been conducted to date, the majority of which indicate that global temperatures will rise, though the magnitude is uncertain. NOAA states that there two causes of climate change: natural variability and human-caused changes (Top et al., n.d.). Variability due to nature climate change is a natural part of the Earth's variability, caused by interactions between the atmosphere, sea, and land, as well as changes in the amount of solar radiation reaching the Earth. Significant evidence for large-scale climate change on Earth's past can be found in the geological record. A plot of temperature data from the Antarctic ice core for the last 420,000 years depicts an example of this variability.

In dealing with climate change, machine learning has made significant contributions, such as enabling automated monitoring via remote sensing (by pinpointing deforestation, collecting building data, and assessing damage after disasters) and accelerating the process of scientific discovery (suggesting new materials for batteries, construction, and carbon capture). Machine learning can also optimize systems to increase efficiency, such as combining delivery, designing carbon markets, and reducing food waste, as well as speed up simulation processes in computing through hybrid modeling, such as climate models (Rolnick et al., 2019).

The Intergovernmental Panel on Climate Change (IPCC) predicts that if global greenhouse gas emissions are not curtailed and drastically reduced, the world will face catastrophic consequences within the next 30 years. In its 2021 report, the IPCC employs a variety of data analyses, one of which is regression analysis, which is used to examine the relationship between cumulative CO_2 emissions and global surface temperature increases. The analysis of historical data shows that the observed increase in global surface temperature in °C since 1850-1900 is a function of historical cumulative CO_2 emissions in $GtCO_2$ from 1850 to 2019. According to the model, global surface temperatures will rise by 3° C or more by 2050 as global warming continues (The

*name of corresponding author

Intergovernmental Panel on Climate Change Working Group I Sixth Assessment Report/AR6 Group 1 IPCC, 2021).

Modeling with statistical machine learning is one integral part of a strategy that can generate statistically significant evidence to formulate variables affecting climate change from multiple areas. Machine learning can also lead to interdisciplinary innovations in agriculture, forestry, geospatial, and other disciplines. Machine learning algorithms can be used to formulate climate system variables, improve forecasts across time scales, and generate policy recommendations about climatic changes (Huntingford et al., 2019). According to the aforementioned description, this study aims to classify greenhouse gas emissions data using machine learning algorithms called logistic regression. The findings of this study are expected to provide a pattern of climate change as then effective programs to deal with greenhouse gas emissions can be developed.

LITERATURE REVIEW

Machine learning is a term used to describe fields of science and engineering that can perform a variety of valuable tasks without being explicitly programmed to do so (Burkov, 2019). Machine learning can also be defined as a subfield related to the development of algorithms that rely on a collection of examples of phenomena that can come from nature, be created by humans, or be generated by other algorithms (Burkov, 2019). Simply stated, machine learning teaches machines, or in this case software, how to complete a task by providing some case examples as training data. The goal is to gain insight into cases from the data set by learning more about the underlying patterns and relationships. The resulting set of rules (also known as a model) can then be applied to a new data set (Richert & Coelho, 2013). Because of the assertion that this algorithm provides promising accuracy results, the use of machine learning algorithms in classification has increased significantly.

In 2021, researchers use a machine learning algorithm to classify Human Development Index (HDI) data in Indonesia based on regencies/cities in 2020. Based on the findings, the best machine learning algorithm is Support Vector Machine (SVM), which has an average accuracy of 95.9% and good predictive quality with an accuracy of 96.04 (Yolanda et al., 2021). (Yolanda et al., 2021)Another study found that the classification of HDI in Indonesia in 2020 based on regencies/cities provided a balanced accuracy of 91.67% (Goldameir et al., 2021). The findings of these two studies showed that machine learning algorithms are very effective at predicting HDI in the coming year and could be used by the government to determine development programs that must be implemented based on regional priorities in the area based on HDI category.

Another research used machine learning algorithms with ensemble methods to classify rainfall data (rainy and non-rainy) at the Sultan Syarif Kasim II Pekanbaru Meteorological Station from January 1, 2018 to July 31, 2021 (Adnan et al., 2021). Machine learning analytics based on historical data could provide a new approach to delivering weather forecasts. Furthermore, this analysis can be applied to more complex cases, such as climate change, to determine model in climate change.

There have also been numerous applications of machine learning algorithms on various climate change indicators. The results of climate change data analysis can be used to gain insight from the data, and the results can be used as a reference and recommendation in decision making or for policymakers. Research that uses machine learning algorithms includes the application of machine learning to analyze the physical causes of climate change: Case studies of extreme rainfall in the Midwest, United States (Davenport & Diffenbaugh, 2021), application of machine learning to mitigate climate change that is geographically different in urban areas (Milojevic-Dupont & Creutzig, 2021), Uncertainty analysis the impact of climate change on the frequency of floods using a hybrid machine learning method (Anaraki et al., 2021), and the application of machine learning for climate change risk assessment (Zennaro et al., 2021).

This study applied logistic regression model for predicting greenhouse gas emissions. Logistic regression is an analysis that uses independent variables that can be nominal, ordinal, interval, or ratio level to test a predictor or set of predictors of a binary dependent variable (Connely, 2020). The logistic regression equation shows how the likelihood of an event (one category of the response variable) is related to each of the explanatory variables (Nayebi, 2020). Models are created using logistic regression based on the estimate that adequately describes the associations. In addition, one of the novel aspects of this study is the comparison of performance with logistic regression analysis on historical data with and without data transformation. Given that there are significant differences in historical global climate data, this transformation is carried out as a standardization in order to get a best model and prediction.

METHOD

The data used in this study is weather data from several countries obtained from The World Bank (The World Bank, 2022), with total greenhouse gas emissions as the dependent attribute and 61 other attributes as independent attributes. The data processing process begins with data cleaning, transformation, and splitting, and data modeling using logistic regression in R Studio software.

*name of corresponding author

a. The data cleaning stage involves deleting missing values and then categorizing the total data on greenhouse gas emissions into categorical, with a value of 1 if the total data on greenhouse gas emissions is greater than the median value and 0 if the total data on greenhouse gas emissions is less.
b. Checking the correlation between independent variables (multicollinearity test). If the attribute has a VIF value greater than 10, then the attribute is not included in the next analysis.
c. Data transformation steps are performed to eliminate data trends (trends), minimize data variances, and perform smoothing. The transformation is carried out using min-max normalization with the equation below.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

d. The data is split by dividing the training and testing data by a percentage: 70% training and 30% testing. At this point, two data sets will be created: Data set 1 and Data set 2.
e. For each data set that is created, the modeling is carried out using logistic regression with two data set, with and without transformation. Because we have two classes in this study, 0 and 1, we can define our case study as a binary classification problem. As a result, we can formulate the results as a probability from [0,1], and if it is greater than a certain threshold, say 0.5 (AI Publishing, 2020), then greenhouse gas emissions are greater than the median, and if it is less than the threshold, then it is less than the median. Each modeling data set will be compared to the accuracy, sensitivity, and specificity values to determine which modeling and dataset generates the best results.

RESULT

Data Cleaning

The first step in data analysis is to ensure that there is no missing data in the data set. If the dataset has missing data, then the data in the same row will be deleted. Based on the results of data cleaning, there were 1,683 data with average of total greenhouse gas emissions was 321,455. The smallest and largest value of total greenhouse gas emissions was 2,260 and 8,786,120 respectively, as follows:

Table 1 Statistic Descriptive

Mean	Q1	Q2	Q3	Min	Max
321455	23205	66710	230560	2260	8786120

Data on total greenhouse gas emissions were categorized based on the median value. If total greenhouse gas emissions data was greater than 66,71 then it would be categorized as 1 (high). Meanwhile, if the data was less than 66,71 it would be categorized as 0 (low). Hence, there was 842 or 50.02% total greenhouse gas emissions data that joined to category 1 (high) while 841 other data (49.98%) categorized into 0 (low). It can be seen that the categorization of data based on the median value made the total greenhouse gas emissions data divided into balanced categories.

Multicollinearity Test

The multicollinearity test was carried out with the aim of knowing whether in a regression model there was a correlation between the predictors. To find the presence or absence of multicollinearity in the regression model, it can be seen from the value of the variance inflation factor (VIF) which measures the variability of the selected predictors that not explained by the others. The cut off value used is the VIF value above the number 10. If the VIF value is > 10, it is stated that multicollinearity occurs so that the attributes is excluded from the modeling.

Table 1 Predictors that have VIF larger then 10

Atribut	VIF
Capture.fisheries.production	132352.0088
Land.area	131873.9038
Total.greenhouse.gas.emissions	80014.29006
CO2.emissions.kt	54247.88458
Methane.emissions	1960.501305
Nitrous.oxide.emissions	353.2839138
Agricultural.nitrous.oxide.emissions.1	104.5445805
Methane.emissions.in.energy.sector	29.79308817
Agricultural.methane.emissions.1	25.90075751

*name of corresponding author

Atribut	VIF
Nitrous.oxide.emissions.in.energy.sector.1	20.45374194
Total.fisheries.production	19.75700527

Based on multicollinearity checking as shown in Table 2, it was found that there were 10 attributes that had a VIF value of more than 10. Then, these attributes were removed from the modeling so that only 52 attributes were used in modeling the total greenhouse emissions.

Data Transformation

Data transformation is carried out because of the unbalanced range of values for each attribute so that it can affect the quality of the data results. Data transformation aims to normalize the data so that the attributes used have data with the same range. The transformation process was carried out using Min-Max normalization. Min-Max normalization transformation is a normalization method by performing a linear transformation of the original data so as to produce a balance of comparison values between the data before and after the process. All data on each attribute was reduced by the minimum value of the attribute then divided by the range of values in that attribute. This transformation produced new normalized data with the smallest value 0 and the largest value 1. This data transformation was expected to increase the accuracy of the results of dataset classification.

Data Splitting

From the 1,683 available data, the data set was then divided with the following conditions:

1) Data set 1: data without transformation with 70% training data and 30% testing data.

2) Data set 2: transformed data with 70% training data and 30% testing data;

Retrieval of data training and testing data was selected randomly by specifying set.seed (932) in R studio. The determination of the seed was done so that every data retrieval is carried out at a later time, it still provides the same data. The series of data sharing is as follows:

Table 2 Number of training and testing data

Category	Training	Testing
Data set 1	0=588 data (50,30%) 1=581 data (49,70%)	0=254 data (49,42%) 1=260 data (50.58%)
Data set 2	0=588 data (50,30%) 1=581 data (49,70%)	0=254 data (49,42%) 1=260 data (50.58%)

Based on table 2, it can be seen that the two datasets have almost the same proportions, both in the proportion of training data and in testing data. The training data used in the two datasets were 1169 data with 588 being total data on greenhouse gas emissions with category 0 and the other 581 being data for total greenhouse gas emissions with category 1. In data testing, the proportion of data on total greenhouse gas emissions with category 1 was slightly higher than the data on total greenhouse gas emissions in category 0. The comparison was 50.58% for category 1 and 49.42% for category 0.

Model Evaluation

Based on logistic regression testing, there were two models formed, namely the dataset model 1 (data without transformation) and the model for dataset 2 (data with transformation). Both models indicate that the models formed are the good models. These indicate that one or more attributes forming the models are attributes that affect the model. These are also reinforced by the significance test, it is found that all attributes have a significant effect with a significance level of 95%. Then, in evaluating the performance of the Machine Learning (ML) algorithm, the Confusion Matrix reference values (Accuracy, Precision, Sensitivity, and Specificity) were used to represent the predictions and actual (actual) conditions of the data generated by the ML algorithm. The results of the comparison of the four models formed are presented in the following table:

Table 3 Comparison of Confusion Matrix

Data set	Description	Accuracy	Precision	Sensitivity	Specificity
Dataset 1	without transformation	0.8619	0.8617	0.8583	0.8654
Dataset 2	with transformation	0.8760	0.8776	0.8704	0.8814

Based on Table 4, it is found that all values in the Confusion Matrix indicator for dataset 2 (data with transformations) provided a higher value than the values generated by dataset 1 (data without transformation).

The accuracy value of the transformed data was 0.8760, that was 0,01% higher than the data without transformation. This indicates that the model generated by the transformed data was better because it could predict accurately 87.60% of the data, while dataset 1 can only predict correctly 86.195 of the data.

Furthermore, the precision value in the transformation data was 0.8776, which is 0,0159 higher than the precision value in the data without transformation. This explains that the transformed data produced a higher ratio of positive correct predictions than the data without transformation. This was in line with the sensitivity value generated by the transformed data which was higher than the untransformed data. The transformation data had value of 0.8704, which means that 87.04% of the transformation data modeling predicted true positives compared to all the original data which were positive.

In contrast to sensitivity, specificity is the proportion or ratio of the correctness of negative predictions compared to the overall negative data. The specificity value for the transformation data was 0.8814 while the specificity value for the transformation data is 0.8654. It means that the transformed data could describe the ratio of negative predictions to the overall negative data which is higher than the data without transformation. In the transformation data, 88.14% of the negative data were correctly predicted, while the data without transformation only predicted 86.54% of the negative.

DISCUSSIONS

Preparing data is a very important preprocessing stage in data mining, the main reason is because the quality of the input data greatly affects the quality of the resulting analysis output. The dataset used in this study has a different range of values for each attribute. This can affect the modeling because it can cause the attribute to malfunction which has a much smaller value compared to other attributes (Nasution et al., 2019). To overcome this, preprocessing the data through data transformation by means of min-max normalization so that the range of values in the attribute becomes the same by performing a linear transformation of the original data so as to produce a balance of comparison values between the data before and after the process. This transformation produces new normalized data with the smallest value 0 and the largest value 1.

Based on the results of logistic regression modeling, it was found that the data from the min-max transformation resulted in better modeling than the data modeling without going through the transformation process. This is evident from the results of the accuracy, precision, and specificity values possessed by the transformation data which are higher than the values of accuracy, precision, and specificity possessed by the transformation data. The values of accuracy, precision, sensitivity, and specificity in the transformation data are 87.60%, 87.76%, 87.04% and 88.14%, respectively. This percentage is a fairly high percentage and is good in logistic regression modeling.

CONCLUSION

The transformation process carried out before data modeling affects the accuracy of the data classification results. Modeling using min-max normalization data shows better results than modeling without transformation. The values of accuracy, precision, sensitivity, and specificity of the transformation data modeling also showed high results, namely 87.60%, 87.76%, 87.04% and 88.14%, respectively.

ACKNOWLEDGMENT

This work was financially supported by LPPM Universitas Riau 2022 with the grant number 1500/UN19.5.1.3/PT.01.03/2022 from DIPA Universitas Riau 2022.

REFERENCES

Adnan, A., Yolanda, A. M., & Natasya, F. (2021). A comparison of bagging and boosting on classification data: Case study on rainfall data in Sultan Syarif Kasim II Meteorological Station in Pekanbaru. *Journal of Physics: Conference Series*, *2049*(1), 012053. https://doi.org/10.1088/1742-6596/2049/1/012053

AI Publishing. (2020). *Regression Models with Python for Beginners* (First). AI Publishing LLC.

Anaraki, M. V., Farzin, S., Mousavi, S.-F., & Karami, H. (2021). Uncertainty Analysis of Climate Change Impacts on Flood Frequency by Using Hybrid Machine Learning Methods. *Water Resources Management*, *35*, 199–223.

Burkov, A. (2019). The Hundred Pages Machine Learning Book. In *Expert Systems*. Andriy Burkov. https://doi.org/10.1111/j.1468-0394.1988.tb00341.x

Connely, L. (2020). Logistic Regression. *Medsurg Nursing*, *29*(5).

Davenport, F., & Diffenbaugh, N. (2021). Using Machine Learning to Analyze Physical Causes of Climate Change: A Case Study of U.S. Midwest Extreme Precipitation. *Geophysical Research Letters*, *48*(15). https://doi.org/http://dx.doi.org/10.1029/2021GL093787

Goldameir, N. E., Yolanda, A. M., Adnan, A., & Febrianti, L. (2021). Classification of the Human Development

Index in Indonesia Using the Bootstrap Aggregating Method. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *6*(1), 100–106. https://doi.org/10.33395/sinkron.v6i1.11173 e-ISSN

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environ. Res. Lett.*, *14*(124007). https://doi.org/doi.org/10.1088/1748-9326/ab4e55

Milojevic-Dupont, N., & Creutzig, F. (2021). Machine learning for geographically differentiated climate change mitigation in urban areas. *Sustainable Cities and Society*, *64*(102526). https://doi.org/doi.org/10.1016/j.scs.2020.102526

Nayebi, H. (2020). *Advanced statistics for testing assumed casual relationships*. Springer. http://www.springer.com/series/14538%0Ahttp://dx.doi.org/10.1007/978-3-030-54754-7

Richert, W., & Coelho, L. P. (2013). *Building Machine Learning Systems with Python*. PACKT Publishing.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2019). *Tackling Climate Change with Machine Learning*. https://doi.org/10.48550/arxiv.1906.05433

The Intergovernmental Panel on Climate Change Working Group I Sixth Assessment Report/AR6 Group 1 IPCC. (2021). *Climate Change 2021 The Physical Science Basis: Summary for Policymakers*. The Intergovernmental Panel on Climate Change.

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

Top, G., Anomaly, C., & Anomaly, F. (n.d.). *Global Climate Report-Annual 2010*.

Yolanda, A. M., Adnan, A., Goldameir, N. E., & Rizalde, F. A. (2021). The Comparison of Accuracy on Classification Data with Machine Learning Algorithms (Case Study: Human Development Index by Regency/City in Indonesia 2020). *International Conference on Advanced Technology and Multidiscipline (ICATAM) 2021*.

Zennaro, F., Furlan, E., Simeoni, C., Torresan, S., Aslan, S., Critto, A., & Marcomini, A. (2021). Exploring machine learning potential for climate change risk assessment. *Earth-Science Reviews*, *20*(103752). https://doi.org/doi.org/10.1016/j.earscirev.2021.103752