Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 4 No. 1 2024, 1 - 8
DOI: *https://doi.org/10.22487/27765660.2024.v4.i1.17051*
eISSN: 2776-5660

WORLD GREENHOUSE GAS EMISSION ANALYSIS: EVALUATING CLASSIFICATION ACCURACY USING SUPPORT VECTOR MACHINE (SVM)

Kurnia Ramadani[1*], Gustriza Erda[2]

[1,2]Statistics Study Program, Riau University

***e-mail**: *kurnia.ramadani0394@student.unri.ac.id*

ABSTRACT

The phenomenon of Heatwaves has struck several countries across the globe due to climate change. This climate change has led to an increase in greenhouse gas emissions surpassing the limits set by the IPCC Fourth Assessment Report GWPs. This study utilizes the Support Vector Machine (SVM) classification method to identify and categorize greenhouse gas emission data from 1990 to 2020 using four kernels function such as linear, polynomial, radial basis function (RBF), and sigmoid. The SVM method demonstrates excellent performance in constructing classification models with a polynomial kernel function. This is evidenced by high values of training accuracy, testing accuracy, and F1-score, accompanied by short training and testing analysis times. Successively, these values are 97.39%, 97.69%, 96.82%, 0.59 seconds, and 0.22 seconds.

Keywords: *Classification, SVM, greenhouse gas emission, climate change.*

Cite: Ramadani, K., & Erda, G. (2024). *World Greenhouse Gas Emission Analysis: Evaluating Classification Accuracy using Support Vector Machine (SVM). Parameter: Journal of Statistics, 4*(1), 1-8, https://doi.org/10.22487/27765660.2024.v4.i1.17051.

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

INTRODUCTION

The phenomenon of heatwaves caused by climate change has affected several countries across the globe (Setiawati, 2023). Climate change occurs due to greenhouse gas emissions comprising water vapor (H_2O), carbon dioxide (CO_2), nitrous oxide (N_2O), methane (CH_4), hydrofluorocarbons (HFC), perfluorocarbons (PFC), and sulfur hexafluoride increasing in Earth's atmosphere. The IPCC Fourth Assessment Report GWPs set the greenhouse gas emission limit at 431 $MMTCO_2eq$, where China's emission volume stands at 15.7 $GtCO_2eq$ (Crippa et al., 2023), far exceeding this limit. This condition needs attention to minimize greenhouse gas emissions in the future by identifying and categorizing greenhouse gas emissions using machine learning algorithms in supervised learning techniques. Support vector machine (SVM) has the ability to classify research data effectively, as demonstrated by many studies that have utilized or compared support vector machine classification methods with other approaches. Supervised learning is a part of machine learning and artificial intelligence that utilizes labeled data to train algorithms (ibm, 2024). The supervised learning technique used is the classification method. According to the Indonesian Dictionary (KBBI), classification is the arrangement of classes made hierarchically into a group or category based on established standards according to needs (KEMDIKBUD, 2024). The classification method used in this research is the classification method in supervised learning, namely the Support Vector Machine (SVM) method.

The use of SVM methods is widely applied in several classification-related research studies. For instance, the SVM classification method has been used in the classification of the Human Development Index (HDI) by (Yolanda et al., 2023), where the SVM classification method achieved an accuracy of 95.9% and prediction quality with an accuracy of 96.04%. Additionally, it was used by Bintang Girik Allo et al., (2023) to classify breast cancer problems, where the SVM method provided a classification accuracy of 81.816%. SVM classification research was also conducted by Pasaribu et al., (2021) to classify data from community health centers in Bandar Lampung City, where the SVM method achieved a classification accuracy of 99.78%. The SVM method was also utilized in the classification of elementary school accreditation data in Magelang Regency by Anna et al., (2014) where the classification had a classification accuracy of 93.902%.

Research using the support vector machine method in the environmental field, especially regarding climate change, is rarely conducted. Climate change-related research has been carried out by Adnan et al., (2023) using logistic regression methods with the results 87.60% accuracy, 87.76% precision, 87.04% recall, and 88.14% spesificity. This study will classify greenhouse gas emissions using the support vector machine (SVM) method by applying four different kernel functions and evaluate which kernel function performs best in classifying greenhouse gas emissions.

MATERIALS AND METHODS

Data Sources and Research Variables

The data used in this study are secondary data sourced from several institutions, namely BMKG, the World Bank, and Kaggle, consisting of climate change-related data from 1990 to 2020. The analysis in this study utilizes the Python programming language on Google Colab.

This study uses 40 variables with 1 dependent variable, namely greenhouse gas emissions, and 39 independent variables that influence the dependent variable. The dependent variable is the total greenhouse gas emissions in kiloton CO_2 equivalent categorized into two, namely high (1) if the dependent variable value is greater than 431 $MMTCO_2eq$ and low (0) if the dependent variable value is less than 431 $MMTCO_2eq$.

The independent variables in this study are descibed in Table 1.

Table 1. Independent variables

Variables	Explanation
$\mathbf{X_1}$	Alternative and nuclear energy (% of total energy)
$\mathbf{X_2}$	CO_2 emissions (metric tons per capita)
$\mathbf{X_3}$	CO_2 emissions (kilotons)
$\mathbf{X_4}$	Net energy import (% of energy use)
$\mathbf{X_5}$	Fossil fuel energy consumption (% of total)
$\mathbf{X_6}$	GDP (current US$)
$\mathbf{X_7}$	GDP growth (% annual)
$\mathbf{X_8}$	GDP per capita (current US$)
$\mathbf{X_9}$	Ore and metal exports (% of merchandise exports)

Variables	Explanation
X_{10}	Oil rent (% of GDP)
X_{11}	Natural gas rent (% of GDP)
X_{12}	Mineral rent (% of GDP)
X_{13}	Forest rent (% of GDP)
X_{14}	Coal rent (% of GDP)
X_{15}	Energy depletion (current US$)
X_{16}	Mineral depletion (current US$)
X_{17}	Gross fixed capital formation (current US$)
X_{18}	Net forest depletion (current US$)
X_{19}	CO_2 damage (current US$)
X_{20}	Fisheries production (metric tons)
X_{21}	Aquaculture production (metric tons)
X_{22}	Nitrous oxide emissions (thousand metric tons of CO_2 equivalent)
X_{23}	Nitrate oxide emissions in energy sector (thousand metric tons of CO_2 equivalent)
X_{24}	Nitrous oxide emissions in agriculture (thousand metric tons of CO_2 equivalent)
X_{25}	Methane emissions
X_{26}	Methane emissions in energy sector (thousand metric tons of CO_2 equivalent)
X_{27}	Methane emissions in agriculture (thousand metric tons of CO_2 equivalent)
X_{28}	CO_2 emissions from solid fuel consumption (kilotons)
X_{29}	CO_2 emissions from liquid fuel consumption (kilotons)
X_{30}	CO_2 emissions (kg per 2015 US$ of GDP)
X_{31}	CO_2 emissions from gas fuel consumption (kilotons)
X_{32}	Surface area (square kilometers)
X_{33}	Land area (square kilometers)
X_{34}	CO_2 emissions from other sectors
X_{35}	CO_2 emissions from manufacturing and construction
X_{36}	CO_2 emissions from electricity and heat production
X_{37}	CO_2 emissions from residential buildings and commercial and public services
X_{38}	CO_2 intensity (kg per kg of oil equivalent energy use)
X_{39}	Electricity production from renewable sources, excluding hydroelectric (% of total)
Y	Total greenhouse gas emissions (kilotons of CO_2 equivalent)

Analysis Steps

The analysis steps in this study are as follows:

1. Collecting climate change factor data from various sources.
2. Converting the dependent variable into categories 1 and 0.
3. Splitting the data into training data and testing data with a ratio of 70% for training data and 30% for testing data.
4. Creating a classification model with Support Vector Machine (SVM) on the training data using linear, polynomial, RBF, and sigmoid kernel functions.
5. Training the classification model with Support Vector Machine (SVM) using linear, polynomial, RBF, and sigmoid kernel functions.
6. Comparing the evaluation results.
7. Interpreting the comparison evaluation results.
8. Making conclusions from the analysis results.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) operates on the principle of maximizing the margin and optimizing the hyperplane to find the best hyperplane. The hyperplane (Decision boundary) is the optimal separator between two high-dimensional classes that can be determined by maximizing the

margin from the support vectors or the data points closest to the hyperplane (Vapnik & N., 1995). Support vectors, hyperplanes, and margins are depicted in Figure 1.

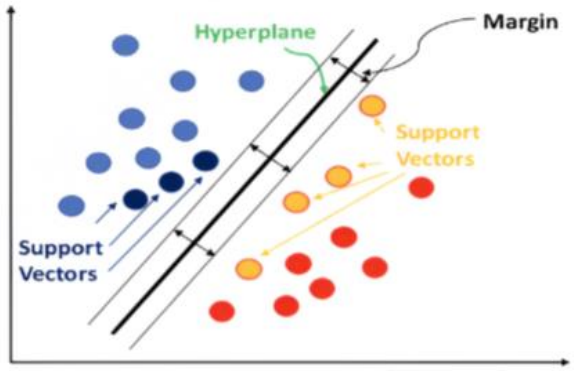

Figure 1. Support Vector Machine (SVM) (Rizwan, 2023)

The Support Vector Machine (SVM) employs a linear decision function model with the general form as follows.

$$\boldsymbol{f(x) = w\phi(x) + b} \quad (1)$$

where, w and b are two parameters that can be calculated in this estimation. $\phi(x)$ is the basis function.

In the Linear Support Vector Machine, the separator between classes is referred to as a linear function. Training data is represented as ($x_i y_i$) where i = 1, 2,..., N, x_i = {x_1, x_2…,x_q}is the set of attributes (features) for the i-th training data, and y_i, ∈ {-1,+1} indicates the class label. The linear SVM classification hyperplane can be denoted as (Prasetyo, 2012):

$$\boldsymbol{wx_i + b = 0}. \quad (2)$$

The margin of the hyperplane is given by the distance between the two hyperplanes from the two classes. The notation is summarized as:

$$\boldsymbol{\|w\|\, x\, d = 2 \text{ or } d = \frac{2}{\|w\|}}. \quad (3)$$

Minimize:

$$\boldsymbol{\frac{1}{2}\|w\|^2} \quad (4)$$

subject to:

$$\boldsymbol{y_i(wx_i + b) \geq 1, i = 1, 2, \ldots N}. \quad (5)$$

This optimization can be solved by maximizing the Lagrange multiplier:

$$\boldsymbol{Lp(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{t-1}^{N} a_i y_i (wx_i + b) - 1}. \quad (6)$$

To simplify, equation (6) must be transformed into the Lagrange function itself. The Lagrange multiplier equation is expanded to:

$$\boldsymbol{Lp = \frac{1}{2}\|w\|^2 - \sum_{t-1}^{N} a_i y_i (wx_i) - b\sum_{t-1}^{N} a_i y_i + \sum_{t-1}^{N} a_i}. \quad (7)$$

Equation (7) represents the primal Lagrange model, and it transforms into the Lagrange multiplier duality as Ld whose equation becomes to maximize:

$$\boldsymbol{Ld = \sum_{t-1}^{N} a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j x_i x_j}. \quad (8)$$

With

$$\boldsymbol{a_i \geq 0\ (i = 1, 2, \ldots, l) \sum_{i=1}^{l} a_i y_i = 0}. \quad (9)$$

Support Vector Machine (SVM) is designed to solve linear cases. However, in reality, linear cases are rarely encountered. Therefore, Support Vector Machine (SVM) is modified to solve non-linear cases using a kernel function. Support vectors in Support Vector Machine (SVM) are easier to obtain using this kernel function. The higher the value of support vectors, the higher the accuracy (Satriyo et al., 2003). The kernel function is mathematically represented as follows (Steinwart & Christmann, 2008):

$$\boldsymbol{K(x_i, x_j) = \phi(x_i), \phi(x_j)} \quad (10)$$

where ϕ denotes the mapping from x to feature space. With its function depicted as follows:

$$\boldsymbol{f(x) = \sum_{j=1}^{N} \alpha_j y_j K(x_i, x_j) + b}. \quad (11)$$

Some kernel functions in SVM (Pratiwi & Setyawan, 2021) are:

Table 2. Kernel function

Name of the function	Function
Linear Kernel	$K(x_i, x_j) = x_i^T \cdot x_j$
Polynomial Kernel	$K(x_i, x_j) = (x_i^T \cdot x_j + 1)^p$
RBF (*Radial Basis Function*) Kernel	$K(x_i, x_j) = exp\left(-\frac{1}{2\sigma^2}\left\|\left\|x_i - x_j\right\|\right\|^2\right)$
Sigmoid Kernel	$K(x_i, x_j) = \tan(x_i^T \cdot x_j + 1)$

Evaluation

In this study, the evaluation used is the confusion matrix, where the confusion matrix table is explained in Table 3.

Table 3. *Confusion matrix*

		Prediction	
		Positive (1)	*Negative (0)*
Actual	*Positive (1)*	*TP*	*FN*
	Negative (0)	*FP*	*TN*

This research will focus on examining the accuracy and F1-score values from the analysis results. Explanations regarding accuracy and F1-score are as follows:

a. Accuracy measures how well the model makes correct predictions from the total predictions made. Accuracy is described in the following formula (12).

$$\boldsymbol{accuracy = \frac{TP + TN}{FP + FN + FN + TN}} \quad (12)$$

b. F1-Score is a calculation that describes the balance between precision and sensitivity. F1-Score is described in the following formulas (13) and (14).

$$\mathbf{F1 - Score = 2\ x\ \frac{recall\ x\ precission}{recall + precission}} \quad (13)$$

$$\mathbf{F1 - Score = 2\ x\ \frac{(\frac{TP}{TP + FN})\ x\ (\frac{TP}{TP + FP})}{(\frac{TP}{TP + FN}) + (\frac{TP}{TP + FP})}} \quad (14)$$

RESULTS AND DISCUSSION

The descriptive statistics of dependent variable is depicted in Table 4.

Table 4. descriptive statistics of dependent variable

Descriptive Statistics	Value (kilotons of CO_2 equivalent)
Mean Value	1.440.479,2
Standard Deviation Value	4.089.782,3
Maximum Value	45.873.848
Minimum Value	10
Range Value	45.873.838

Based on Table 4, dependent variable has an mean value of 1.440.479,2 $KtCO_2eq$. Which means, on average, each country have as mush greenhouse gas emission as that with the standard deviation value 4.089.782,3 $KtCO_2eq$. It means, that the classification accuracy values are quite spread out from the mean value of 1.440.479,2 $KtCO_2eq$. This indicates a significant amount of variability in the accuracy of the SVM model's classification of greenhouse gas emissions.

Besides that, the dependent variable has a maximum value of 45.873.848 $KtCO_2eq$ with a minimum value of 10 $KtCO_2eq$. Range value that is shows the difference between the maximum value and the minimum value and also indicates the overall spread of accuracy values of 45.873.838 $KtCO_2eq$. The maximum value of 45.873.848 $KtCO_2eq$ have been far from the limits set by the IPCC at 431 $MMTCO_2eq$ or equivalent to 431.000 $KtCO_2eq$, this is a matter of concern and must be dealt with one of which can be done through accuracy classification analysis as in this research paper.

The categorical proportion of the dependent variable in the study is depicted in Figure 2.

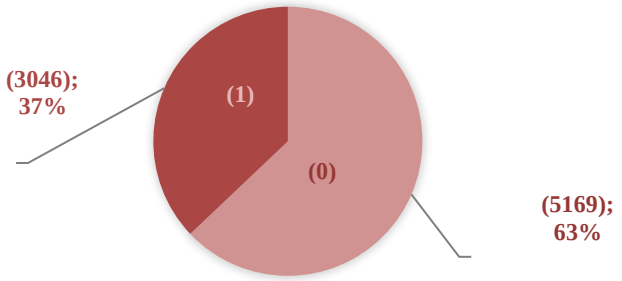

Figure 1. Categorical proportion of the dependent variable

Based on Figure 2, it is found that the proportion of the dependent variable category (1) is 37% with a data count of 3046, and the low category is 63% with a data count of 5169. Although the low category proportion (0) is larger than the high category (1), more attention should be given to the high category (1) because countries included in this category have greenhouse gas emissions exceeding the established limit. This is an important concern because its negative impact can significantly increase if not addressed from the outset.

The categorized dependent data into high (1) and low (0) categories will be used in further classification analysis to see how the data is categorized with the influencing independent variables. To continue the next analysis steps, the research data is divided into 70% training data, which is 5750 data, and 30% testing data, which is 2465 data.

The results of this classification are performance evaluations that will examine the accuracy, analysis time, and F1-score values from the analysis. The accuracy evaluation results for each SVM kernel function on the training data are detailed in Table 5.

Table 5. Support vector machine (SVM) training classification results

Name of the function	Training accuracy (%)	Training time (s)
Linear Kernel	92,00%	1,23
Polynomial Kernel	**97,39%**	**0,59**
RBF (Radial Basis Function) Kernel	91,98%	1,88
Sigmoid Kernel	70,99%	3,89

Based on Table 5, it is found that each kernel function has different accuracy and analysis time values.

The linear kernel function has a training data accuracy of 92.00% with an analysis time of 1.23 seconds. The polynomial kernel function has a training data accuracy of 97.39% with an analysis time of 0.59 seconds. The radial basis function (RBF) kernel function has a training data accuracy of 91.98% with an analysis time of 1.88 seconds. The sigmoid kernel function has a training data accuracy of 70.99% with an analysis time of 3.89 seconds. This means that the support vector machine (SVM) classification method has built a classification model using training data that can classify data with a certain accuracy and with different analysis times for each kernel function. In the analysis in Table 4, it is found that the polynomial kernel function has the best accuracy value of 97.39% with an analysis time of 0.59 seconds. This means that the analysis using the polynomial kernel function with training data can correctly predict approximately 97 data with approximately 3 errors on a scale of 100 in 0.59 seconds.

The results of this classification are performance evaluations that will examine the accuracy, analysis time, and F1-score values from the analysis. The accuracy evaluation results for each SVM kernel function on the testing data are detailed in Table 6.

Table 6. Support vector machine (SVM) testing classification results

Name of the function	Testing accuracy (%)	Testing time (s)
Linear Kernel	92,98%	0,42
Polynomial Kernel	97,69%	0,22
RBF (Radial Basis Function) Kernel	93,90%	0,39
Sigmoid Kernel	71,27%	0,40

Based on Table 5, it is found that each kernel function has different accuracy values and analysis times. The linear kernel function has a testing accuracy of 92.98% with an analysis time of 0.42 seconds. The polynomial kernel function has a training accuracy of 97.69% with an analysis time of 0.22 seconds. The radial basis function (RBF) kernel function has a training accuracy of 93.90% with an analysis time of 0.39 seconds. The sigmoid kernel function has a training accuracy of 71.27% with an analysis time of 0.40 seconds.

This implies that the classification model built using training data can correctly classify testing data to the extent of the testing accuracy within a specific analysis time. Table 5 shows that the polynomial kernel function has the best testing accuracy at 97.69% with an analysis time of 0.22 seconds. This indicates that the polynomial kernel function can correctly classify approximately 98 data points with around 2 errors in a scale of 100.

The results of accuracy evaluation for training and testing models using each SVM kernel function with training data accuracy (atr), testing data accuracy (ats), training time (ttr), testing time (tts), and F1-score (F1-Sc) are outlined in Table 7.

Table 7. Classification results of the support vector machine (SVM) method

Name of the function	atr (%)	ttr (s)	ats (%)	tts (s)	F1-*Sc*
Linear Kernel	92,00%	1,23	92,98%	0,42	89,67%%
Polynomial Kernel	**97,39%**	**0,59**	**97,69%**	**0,22**	**96,82%**
RBF (Radial Basis Function) Kernel	91,98%	1,88	93,90%	0,39	89,55%
Sigmoid Kernel	70,99%	3,89	71,27%	0,40	37,12%

Based on Table 7, it is found that the training data accuracy for all kernel functions has improved in their testing accuracy, followed by decreasing analysis times.

The SVM analysis on the Linear kernel function shows that the training and testing data accuracy are 92.00% and 92.98% respectively, with a training analysis time of 1.23 seconds and a testing analysis time of 0.42 seconds. The Linear kernel function has a balanced data accuracy with an F1-score of 96.82%.

The SVM analysis on the Polynomial kernel function is the kernel function with the highest training and testing data accuracy compared to the other three functions. The training and testing data accuracy for this function are 97.39% and 97.69% respectively. The analysis time used for this function is also the shortest, with 0.59 seconds for training data analysis and 0.22 seconds for testing data analysis. The Polynomial kernel function also provides an F1-score of 96.82%.

The SVM analysis on the Radial Basis Function (RBF) kernel function has a training accuracy of 91.98%, testing accuracy of 93.90%, training analysis time of 1.88 seconds, testing analysis time of 0.39 seconds, and an F1-score of 89.55%.

The SVM analysis on the Sigmoid kernel function has a training accuracy, testing accuracy, training analysis time, and testing analysis time of 70.99%, 71.27%, 3.89 seconds, and 0.40 seconds respectively, with an F1-score of 37.12%.

CONCLUSION

Based on the analysis results of greenhouse gas emissions classification using the support vector machine (SVM) method, it is found that the SVM method has good accuracy values and analysis times in building classification models using the polynomial kernel function. This is evidenced by the training accuracy (atr) of 97.39% with an analysis time of 0.59 seconds. The accuracy results increase in the prediction accuracy of the model using testing data with a testing accuracy of 97.69%, accompanied by a faster analysis time of 0.22 seconds. In addition to good model accuracy and predictions, the SVM method also provides a good balance in data predictions as evidenced by a relatively high F1-score of 96.82%. These results are sufficient to demonstrate that in classification analysis using climate change-related data, Support Vector Machine (SVM) is a good method to use.

These results indicate a good outcome in classification analysis, as shown by the nearly 100% accuracy values and better prediction accuracy in testing data compared to the accuracy values of the classification model in training data accuracy. The results of the classification analysis that can be said to be used in subsequent research to consider its continuation in terms of both the development of the scope of its research, as well as development of methods.

REFERENCES

Adnan, A., Yolanda, A. M., Erda, G., Ell, G. N., & & Indra, Z. (2023). The Comparison of Accuracy on Classification Climate Change Data with Logistic Regression. *SinkrOn : Jurnal Dan Penelitiak Teknik Informatika*.

Anna, P. O., Wilandari, Y., & Ispriyanti, D. (2014). Penerapan Metode SVM Pada Data Akreditasi Sekolah Dasar Di Kabupaten Magelang. *Jurnal Gaussian*, *3*(8), 811–820.

Bintang Girik Allo, C., Sandy Ade Putra, L., Roona Paranoan, N., & Vincentius, A. G. (2023). *Comparing Logistic Regression and Support Vector Machine in Breast Cancer Problem*.

Crippa, M., Guizzardi, D., Pagani, F., Banja, M., Muntean, M., Schaaf, E., Becker, W., Monforti-Ferrario, F., Quadrelli, R., Risquez Martin, A., Taghavi-Moharamli, P., Köykkä, J., Grassi, G., Rossi, S., Melo, J., Oom, D., Branco, A., San-Miguel, J., & Vignati, E. (2023). GHG emissions of all world countries. In *Publications Office of the European Union* (Issue KJ-NA-31-658-EN-N (online),KJ-NA-31-658-EN-C (print)).

IBM. (2024). *What is supervised learning?*. Retrieved from https://www.ibm.com/topics/supervised-learning#:~:text=the next step-,What is supervised learning%3F,data or predict outcomes accurately

KEMDIKBUD. (2024). *Klasifikasi*. Retrieved from Kbbi.Kemdikbud.Go.Id. https://kbbi.kemdikbud.go.id/entri/klasifikasi

Pasaribu, I., Lumbanraja, F. R., Shofiana, D. A., & Aristoteles, A. (2021). Klasifikasi Kejadian Hipertensi Dengan Metode Support Vector Machine (SVM) Menggunakan Data Puskesmas Di Kota Bandar Lampung. *Jurnal Pepadun*, *2*(2), 183–190. https://doi.org/10.23960/pepadun.v2i2.56

Prasetyo, E. (2012). *Data mining konsep dan aplikasi menggunakan MATLAB* (Nikodemus (ed.); Data minin).

Pratiwi, N., & Setyawan, Y. (2021). Analisis Akurasi Dari Perbedaan Fungsi Kernel Dan Cost Pada Support Vector Machine Studi Kasus Klasifikasi Curah Hujan Di Jakarta. *Journal of Fundamental Mathematics and Applications (JFMA)*, *4*(2), 203–212. https://doi.org/10.14710/jfma.v4i2.11691

Rizwan. (2023). *Mastering support vector machine (SVMs)*. Medium.Com. https://medium.com/@rizwan44007/mastering-support-vector-machines-svms-f45c0d9eb33c

Setiawati, S. C. I. (2023). *Sepanjang 2023 Dunia Panas Bak Neraka, Ini Penyebabnya!* https://www.cnbcindonesia.com/research/20230916115625-128-473012/sepanjang-2023-dunia-panas-bak-neraka-ini-penyebabnya

Steinwart, I., & Christmann, A. (2008). Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*. https://api.semanticscholar.org/CorpusID:661123

Vapnik, & N., V. (1995). The Nature of Statistical Learning. In *Theory* (p. 334).

Yolanda, A., Adnan, A., Goldameir, N., & Rizalde, F. (2023). *The comparison of accuracy on classification data with machine learning algorithms (Case study: Human development index by regency/city in Indonesia 2020)*. https://doi.org/10.1063/5.0118720