JOURNAL OF DATA SCIENCE | Vol.2024:05
eISSN:2805-5160

Comparison of Recursive Feature Elimination and Boruta as Feature Selection in Greenhouse Gas Emission Data Classification

Riko Febrian[1], Anne Mudya Yolanda[2]

[1,2] Department of Statistics, Faculty of Mathematics and Natural Sciences, Riau University
Bina Widya Campus Km 12.5 Simpang Baru Pekanbaru 28293 - Indonesia

[*]**Email**: [1]riko.febrian1098@student.unri.ac.id, [2]annemudyayolanda@lecturer.unri.ac.id

Abstract

Classification analysis is a supervised learning method that can be utilized to categorize levels of greenhouse gas emissions. Regular monitoring of greenhouse gas emissions is essential for relevant agencies to devise prevention and mitigation programs that address climate change. In classification analysis, enhancing model performance is correlated with the number of features or variables utilized, thus necessitating feature selection in its application. This study compares feature selection methods for classifying greenhouse gas emission levels, specifically wrapper feature selection, recursive feature elimination, and boruta. The Support Vector Machine (SVM) algorithm is employed to evaluate classification performance, focusing on binary classification into "high" and "low" categories in this study. The results indicate that classification performance improves with feature selection and recursive feature elimination compared to scenarios without feature selection or with Boruta feature selection. By employing three out of the thirty-nine features, accuracy, sensitivity, and specificity of 98.95%, 99%, and 97% were achieved, respectively.

Keywords

Classification, greenhouse gas emissions, feature selection, recursive feature elimination, boruta.

Introduction

Global warming has become a term that is familiar and even heard in the ears of ordinary people. It has long been realized that global warming is something that must be addressed immediately because there are so many negative impacts that come along with it. Global warming is the cause of various natural disasters such as floods, landslides, increases in average daily temperatures, and other natural disasters (Istianah, 2023)

Global warming, as one of the issues of climate change, can indeed be linked to natural phenomena. Events that have occurred throughout Earth's history, such as volcanic eruptions, variations in solar radiation, tectonic movements, and even small changes in the sun's orbit, have shown observable impacts on warming and cooling trends on the planet (Turrentine, 2022). However, the increase in average world temperature, or global warming, cannot be separated from human actions. Human behavior that does not protect the natural environment contributes greatly

Submission: 3 May 2024; **Acceptance:** 28 May 2024

to causing this to happen. Behaviors such as not using public transportation, burning excessive waste, burning factories, and many other examples of behavior that cause the greenhouse effect (greenhouse effect). All of this can increase gas emissions such as CH_4 (methane), CO_2 (carbon dioxide), N_2O (dinitrous oxide), CFCs (chlorofluorocarbons). Some of the sunlight that reaches the Earth's atmosphere will be reflected into space, and some will be absorbed by the land and oceans to warm the Earth. The heat from the earth should return to space, but because the atmosphere is filled with gas emissions above, ultimately the heat is trapped back in the earth's atmosphere and causes global warming (Anggraeni, 2015).

To deal with this global warming attack, a monitoring process is needed that is carried out periodically to build support for policymaking. Variables such as the amount of greenhouse gas emissions are one indicator that can be used as a basis for determining policy. By utilizing this data, it is hoped that relevant institutions can better control the development of global warming.

In the development of technological science, there is a branch of science that studies the use of data, namely data mining. Data mining is the process of automatically finding useful information in large repositories. Data mining techniques are used to explore a database to find useful patterns that may not have been previously known. Apart from that, data mining can also can predict events in the future (Steinbach et al., 2006). The main tasks that can be carried out in data mining include estimation, description, clustering, classification, association, and prediction (Larose & Larose, 2005).

The main function that is often used in data mining is classification. Classification is a job that pays attention to data patterns and categorizes them into predetermined classes. The classification works by building a model from training data, which is then used to classify new data into the available classes (Utomo & Mesran, 2020). When running a classification algorithm, you can, of course, use various data as predictor variables. However, the number of variables or the number of features used does not always have a positive correlation with the forecasting results, often there are noise features. So before building a model, it is necessary to clean up the noise features so that the modeling process becomes more efficient and the accuracy results are more optimal (Narayanan et al., 2013). The process of selecting this feature is called a method of feature selection.

Feature selection is a useful method for selecting relevant features or removing redundant or irrelevant features and can reduce computational costs or computing time. This is done so that the algorithm that will be run can work more efficiently and effectively, which in this case is expected to increase accuracy in the process (Arifin, 2015). When used, the wrapper method often produces better performance than the filter and embedded method.

A related study comparing feature selection methods has been carried out previously by Silalahi, Murfi, and Satria (2017). This research discusses classification performance through different feature selection stages, among others *variance threshold*, *univariate chi-square*, *Recursive Feature Elimination (RFE),* and *Extra Trees Classifier* (ETC). The results of this research are methods variance *threshold* and *univariate chi – square* makes the resulting accuracy decrease, while the method *Recursive Feature Elimination* (RFE) and *Extra Trees Classifier* (ETC) can increase accuracy. The highest accuracy is provided by the method *Recursive Feature Elimination* (RFE).

The feature selection method that is also often used is *boruta* and *chi square,* as research conducted by Bhalaji, Sundhara Kumar, and Selvaraj (2018).In this research, two datasets were used, including census data in the United States and data*iris.* The selected features are then applied to various classification methods. According to the results obtained on the Iris dataset, there is no

difference in the classification results of these two feature selection methods because all features are categorized as important. However, in the United States census dataset, boruta feature selection produces better classification accuracy in almost all classification methods tried, namely *Naive Bayes, Decision tree, random forest, dan gradient boosting*.

In general, this research will examine how the classification model performs when using feature selection methods such as *Recursive Feature Elimination* and *Boruta.* This method will be applied to the dataset of World *Bank Climate Change Data* that was obtained from the *World Bank Open Data page.* So, the main focus of this research is to compare the performance of each of the methods mentioned above and find out what independent variables influence the dependent variable.

Methodology

Data

The data that will be used in this research is secondary data sourced from the website https://www.worldbank.org/. The data used is data from 1990 to 2020 which contains climate change with 8215 observations and 40 variables. The dependent variable in this research is *Total greenhouse gas emissions (kt of CO_2 equivalent).* Meanwhile, 39 other variables are independent.

Recursive Feature Elimination

Recursive Feature Elimination is one of the feature selection methods which is often used in selecting features. This method focuses on reducing features that are irrelevant or have weak relationships and is carried out repeatedly recursively until it produces the best features that will be used to build the model (Kuhn & Johnson, 2019)

Feature selection is carried out by ranking based on the attribute coefficient values. The ranking value will be directly proportional to the attribute coefficient value. The feature with the lowest attribute coefficient will be removed at each iteration, then the attribute coefficient is recalculated with the remaining available features. This is intended to see which features are superior and play the most role (Pratama et al., 2022)

Boruta

Boruta is among the most commonly utilized feature selection techniques and constitutes a component of the wrapper feature selection method. This approach operates by assessing the degree of feature relevance aided by shadow features, which comprise randomly permuted values and serve as duplicates of the original features (Anand et al., 2021). Boruta draws upon the foundational concept of the Random Forest Classifier, characterized by the deliberate introduction of randomness within its decision framework. This stochastic system serves the purpose of discerning the pivotal features inherent in a dataset. By embracing randomness, Boruta endeavors to distinguish between salient features and extraneous noise, thus facilitating the precise selection of features essential for predictive modeling endeavors (Kursa & Rudnicki, 2010).

The steps in Boruta feature selection process are as follows:

1. Creation of shadow features by generating copies of the original features, which are subsequently shuffled.
2. Execution of a tree-based algorithm on the entire dataset to compute feature importance.
3. Evaluation of whether the original features exhibit greater importance compared to the shadow features.
4. Iteration of these steps for each cycle. If an original feature demonstrates superior importance over its shadow counterpart, it is designated as significant.

Classification*Support Vector Machine* (SVM)

Classification is part of data *mining* which is often used to group observations into predetermined classes (Larose & Larose, 2005). *Support Vector Machine* (SVM) is a classification algorithm that was originally discovered by Leg Vapnik, Issabelle Guyon and Bernhard Boser in 1992. SVM maps data non-linearly so that the data turns into data with higher dimensions. SVM focuses on creating hyperplane which can divide data in linear conditions by mapping it non-linearly accurately. SVM can be relied on to divide data from two classes (Widodo et al., 2013).

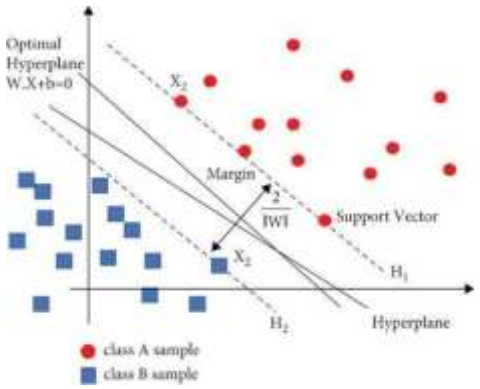

Figure 1. Illustration of classification with SVM

Evaluation and Performance Method

In carrying out classification, it is hoped that the resulting analysis will be classified correctly, so it is necessary to evaluate and validate the model that has been created to see how the model maps the actual data. Evaluation of model performance can be done by creating cross-tabulations between actual data and predicted results. From the tabulation obtained, it will be possible to calculate the accuracy, sensitivity, and specificity (Mumtazah, 2021). Table 1 is a form of cross-tabulation often referred to as *Confusion Matrix.*

Table 1. *Confusion Matrix*

		Prediction Class	
		Positive (P)	Negative (N)
Actual Class	Positive (P)	TP	FN
	Negative (N)	FP	TN

From the Confusion matrix then the values for accuracy, sensitivity, specificity, and the following calculations will be calculated:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \times 100\% \quad (1)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \times 100\% \quad (2)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP} \times 100\% \quad (3)$$

The analysis steps in this research are outlined as follows:

1. Data Collection: Gather the data to be examined from the World Bank website.
2. Descriptive Analysis: Conduct descriptive analysis to explore the dataset.

3. Feature Selection Analysis:
    a. Perform feature selection on the dataset using recursive feature elimination and Boruta methods, as discussed previously.
    b. Collect important features based on the results of feature selection from each method.
    c. Split the data into training (70%) and testing (30%) sets, utilizing a random state of 12, to apply the Support Vector Machine classification algorithm to the training data that has undergone feature selection.
    d. Test the classification results without feature selection by comparing them with the testing data.
4. Evaluation Performance: Evaluate the performance of each feature selection method.
5. Validation and Comparison: Validate and compare the results to determine the most effective method.
6. Interpretation: Interpret the findings derived from the analysis conducted.
7. Conclusion: Draw conclusions based on the results obtained.

Results and Discussion

Descriptive Analysis

The dependent variable remains in numerical form; however, for classification purposes, it necessitates categorization. Drawing from the California Greenhouse Gas Emission Inventory Program, emissions surpassing 431 million metric tonnes of carbon dioxide equivalent (MMTCO2e) are deemed 'high.' Consequently, variable Y is dichotomized into 'low' and 'high' categories based on this threshold. Table 2 below displays a descriptive analysis of variable Y.

Table 2. Descriptive Analysis

Code Variable	N	Mean	With	Min	Max
Y	8215	1440479	82990	10	45873848

Following categorization, the analysis reveals 5,169 observations falling within the 'low' category and 3,046 observations within the 'high' category. Table 2 provides a descriptive analysis of variable Y.

The dependent variable remains in numerical form, necessitating categorization for classification purposes. As per the California Greenhouse Gas Emission Inventory Program, emissions exceeding 431 million metric tonnes of carbon dioxide equivalent (MMTCO2e) are deemed 'high.' Consequently, variable Y is partitioned into 'low' and 'high' categories based on this threshold. Following categorization, it was found that there were 5,169 observations classified under the 'low' category and 3,046 observations under the 'high' category.

Recursive Feature Elimination

The Recursive Feature Elimination (RFE) method works by eliminating one variable with the smallest absolute coefficient. This research uses the Recursive Feature Elimination with Cross Validation (RFECV) method. In RFECV, in each iteration, the classification performance is calculated, and the number of variables with the best performance is selected.

In the first iteration, the smallest variable coefficient value is obtained, namely X_7 or GDP growth (annual%), with a value of 0.0059, therefore, this variable is not included in the next iteration. In the second iteration, the smallest variable coefficient value is obtained, namely X_{10} or Oil Rent(% of GDP), with a value of 0.0284, and this variable will not be used in the next iteration. In the third iteration, the smallest variable coefficient value is obtained, namely X_8 or GDP per capita (current US$), so this variable will not be included in the next iteration. This process continues until all variables are exhausted. At each iteration, the classification performance is calculated, which can be seen in Figure 2.

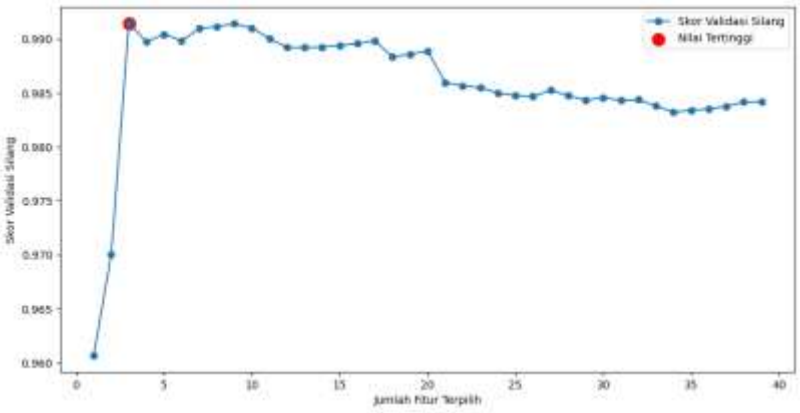

Figure 2. Cross-validation score for each feature of the iteration results

Figure 2 shows that the best classification occurs when using 3 variables. These variables include CO emissions (kt), Nitrous oxide emissions (equivalent to one thousand metric tons of CO_2), CO emissions of liquid fuel consumption (kt).

Boruta

Feature selection method with boruta runs on the classification algorithm method with the tree-based *model* as previously explained, a classification model is used to select these features *random forest* with parameters*n_jobs*=1, *n_estimators*=100, *max_depth* =5, and *random_state*=12. Feature selection boruta selects useful features by giving a rating to each feature.

Important features will get a rating of 1, while features other than 1 are considered unimportant in classification. From the feature selection that has been obtained, the method of boruta resulted in 17 features categorized as important. These features will be used for the next classification process.

Application of Classification with SVM

After obtaining important features according to each feature selection method, these features are then used to build a classification model for greenhouse gas emissions data. Table 3 is the confusion *matrix* classification results.

Table 3. *Confusion Matrix*

Treatment	Real	Prediction	
		Low	Height
Recursive Feature Elimination	Low	1548	3
	Height	23	891
Boruta	Low	1496	55
	Height	416	498

In simple terms, from Table 3 it can be seen that classification using *Recursive Feature Elimination* produced fewer prediction errors than the other two treatments. Feature selection boruta produces more classification errors than without feature selection. The following is a classification *report* from this analysis.

Table 4. *Classification Report*

Methods	Number of Feature	Accuracy	Sensitivity	Specificity
RFE	3	98,95%	99%	97%
Boruta	18	80,89%	78%	54%

Table 4 reveals that Recursive Feature Elimination emerges as the optimal method for classifying greenhouse gas emissions data. Utilizing merely 3 out of the 39 features yields superior accuracy, sensitivity, and specificity values—98.95%, 99%, and 97% respectively. In contrast, Boruta FS achieves lower rates of 80.89%, 78%, and 54%. This underscores the notion that a larger feature set does not necessarily augment model accuracy, as evidenced by the superiority of a minimal feature subset over the larger one.

Conclusions

In classification analysis, the quality of input variables profoundly impacts the predictive model's efficacy, adhering to the principle of "garbage in, garbage out" prevalent in data mining. Consequently, meticulous preprocessing of data significantly influences modeling outcomes. In this study, feature selection notably enhanced model accuracy, with both recursive feature elimination and Boruta showcasing efficacy. Particularly, employing SVM in conjunction with recursive feature elimination demonstrated superior performance, achieving accuracy, sensitivity, and specificity rates of 98.95%, 99%, and 97%, respectively, utilizing only 3 out of the 39 features.

References

Anand, N., Sehgal, R., Anand, S., & Kaushik, A. (2021). Feature selection on educational data using Boruta algorithm. *International Journal of Computational Intelligence Studies*, *10*(1), 27. https://doi.org/10.1504/ijcistudies.2021.113826

Anggraeni, D. Y. (2015). Pengungkapan Emisi Gas Rumah Kaca, Kinerja Lingkungan, Dan Nilai Perusahaan. *Jurnal Akuntansi Dan Keuangan Indonesia*, *12*(2), 188–209. https://doi.org/10.21002/jaki.2015.11

Arifin, M. (2015). Ig-Knn Untuk Prediksi Customer Churn Telekomunikasi. *Simetris : Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *6*(1), 1. https://doi.org/10.24176/simet.v6i1.230

Bhalaji, N., Sundhara Kumar, K. B., & Selvaraj, C. (2018). Empirical study of feature selection methods over classification algorithms. *International Journal of Intelligent Systems Technologies and Applications*, *17*(1–2), 98–108. https://doi.org/10.1504/IJISTA.2018.091590

Istianah, L. (2023). *Gunung Djati Conference Series*, *Volume 19 (2023) CISS 4*. *19*, 104–111.

Kuhn, M., & Johnson, K. (2019). Feature Engineering and Selection. In *Feature Engineering and Selection*. https://doi.org/10.1201/9781315108230

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Larose, D. T., & Larose, C. D. (2005). Discovering Knowledge in Data: an Introduction to Data Mining. In *Journal of the American Statistical Association* (Vol. 100, Issue 472). https://doi.org/10.1198/jasa.2005.s61

Mumtazah, n. A. (2021). *Perbandingan hasil metode support vector machine (svm) dengan ensemble smote bagging dan smote boosting pada data kelulusan mahasiswa unimus*. Universitas muhammadiyah semarang.

Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8206 LNCS*, 194–201. https://doi.org/10.1007/978-3-642-41278-3_24

Pratama, A. R. I., Latipah, S. A., & Sari, B. N. (2022). Optimasi Klasifikasi Curah Hujan Menggunakan Support Vector Machine (Svm) Dan Recursive Feature Elimination (Rfe). *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *7*(2), 314–324. https://doi.org/10.29100/jipi.v7i2.2675

Silalahi, D. K., Murfi, H., & Satria, Y. (2017). Studi Perbandingan Pemilihan Fitur untuk Support Vector Machine pada Klasifikasi Penilaian Risiko Kredit. *EduMatSains : Jurnal Pendidikan, Matematika Dan Sains*, *1*(2), 119–136. http://ejournal.uki.ac.id/index.php/edumatsains/article/view/238

Steinbach, M., Pi, V. I., Ku Mar, N., San, B., Newyork, F., Toronto, L., Tokyo, S., Madrid, S., Munich, M., Capetown, P., & Montreal, H. (2006). *Introduction to Data Mining*.

Turrentine, J. (2022). *What are the causes of climate change?* NRDC. https://www.nrdc.org/stories/what-are-causes-climate-change#natural

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *Jurnal Media Informatika Budidarma*, *4*(2), 437. https://doi.org/10.30865/mib.v4i2.2080

Widodo, P. ., Handayanto, R. T., & Herlawati, H. (2013). Penerapan Data Mining dengan Matlab. *Bandung: Rekayasa Sains*.