Available at https://jtsiskom.undip.ac.id (31 January 2022)
DOI:10.14710/jtsiskom.2022.14105

Data scaling performance on various machine learning algorithms to identify abalone sex

Willdan Aprizal Arifin[*)], Ishak Ariawan, Ayang Armelita Rosalia, Lukman, Nabila Tufailah

Marine Information System, Universitas Pendidikan Indonesia
Jl. Dr. Setiabudi No.229, Isola, Sukasari, Bandung City, West Java 40154, Indonesia

How to cite: W. A. Arifin, I. Ariawan, A. A. Rosalia, L. Lukman, and N. Tufailah, "Data scaling performance on various machine learning algorithms to identify abalone sex," *Jurnal Teknologi dan Sistem Komputer*, vol. 10, no. 1, pp. 26-31, 2022. doi: 10.14710/jtsiskom.2022.14105, [Online].

Abstract – This study aims to analyze the performance of machine learning algorithms with the data scaling process to show the method's effectiveness. It uses min-max (normalization) and zero-mean (standardization) data scaling techniques in the abalone dataset. The stages carried out in this study included data normalization on the data of abalone physical measurement features. The model evaluation was carried out using k-fold cross-validation with the number of k-fold 10. Abalone datasets were normalized in machine learning algorithms: Random Forest, Naïve Bayesian, Decision Tree, and SVM (RBF kernels and linear kernels). The eight features of the abalone dataset show that machine learning algorithms did not too influence data scaling. There is an increase in the performance of SVM, while Random Forest decreases when the abalone dataset is applied to data scaling. Random Forest has the highest average balanced accuracy (74.87%) without data scaling.

Keywords *– data scaling; machine learning algorithms; min-max normalization; zero-mean standardization*

I. Introduction

The study of machine learning is essential for addressing fundamental scientific questions [1]. Machine learning offers various methods that can be applied in marine science [2]. The global challenges in marine science are necessary to realize the collected data's potential through auto-mating more of the analysis [3]. One of them is the use of machine learning to identify marine species [4]-[6]. It supports data-driven learning, resulting in automated decision-making [2].

Abalone is one of the giant marine gastropod mollusks, and they are economically significant seafood in aqua-culture worldwide [7]. Eight features of the abalone dataset (length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, ring) were used to determine the three sexes of Abalone, which are Female (F), Male (M), and Infant (I). Machine learning has great potential to improve the quality and range science of abalone sex identification by identifying trends and distribution patterns within the abalone dataset [2].

Data scaling or normalization is an additional step between the two possibilities in the dataset to be tested: normal and non-normal distribution. The normal data distribution is ready to be processed, while the non-normal data needs to be normalized. In general, normalization techniques or data scaling have an important role in data preprocessing [8]-[10]. It involves sub-processed data from multiple sources. Data may need to be reformatted, normalized, and aggregated [11]. The normalized data can improve intrusion detection accuracy [12].

Combining data normalization techniques with machine learning algorithms improves the performance [13]. Li and Liu [14] show that the method using SVM with normalization has much better performance than the method using SVM without normalization because Min-Max normalization has better performance in speed and accuracy. Identifying marine species using several machine learning algorithms has been implemented. Moitinho-Silva [4] used machine learning to predict HMA-LMA status in marine sponges. This research comparing several machine learning algorithms shows that the high classification performance uses a random forest algorithm trained with phylum. It produces weighted mean accuracy of 96.90% and ±5.75 of weighted standard deviation.

In Ambarwati et. al. [13], the objects used were medicinal plant leaves with the SVM, KNN, ANN, and Naïve Bayes algorithms, Li and Liu [14] for intrusion detection systems with the SVM algorithm, and Moitinho-Silva [4] for marine species (involving corals, seaweeds, plankton, and fish) with the random forest algorithm. The prior studies were only limited to the use of machine learning algorithms without modification of changes to the dataset. This study aims to analyze the performance of the machine learning algorithms on data scaling techniques in identifying abalone sexes based on eight features.

Data scaling or normalization is analyzed for its impact on several machine learning algorithms to identify abalone sexes. Four classifiers are used: Random Forest, Naïve Bayesian, Decision Tree, and

[*)] Corresponding author (Willdan Aprizal Arifin)
Email: willdanarifin@upi.edu

Copyright ©2022, The authors. Published by Department of Computer Engineering, Universitas Diponegoro

Submitted: 15 February 2021; Revised: 22 July 2021; Accepted: 10 August 2021; Published: 31 January 2022

SVM. Furthermore, the determination to use data scaling techniques can be appropriately applied to the algorithms. The eight variations of the features were carried out by data scaling or data normalization to analyze their impact on the machine learning algorithms. Each feature of the abalone dataset is compared before and after applying the min-max (normalization) and zero-mean (standardization) techniques to test their performances.

II. Research Method

The dataset used in this study is the abalone dataset taken from Kaggle[1]. This data has been grouped based on sex, which are Female (F), Male (M), and Infant (I) [15]. The research stages are data collecting, preprocessing, scaling, sharing, modeling, analysis, and evaluation. The research stages can be seen in Figure 1.

A. Preprocessing data

Data balancing is carried out at this stage using the Synthetic Minority Oversampling Technique (SMOTE) method [16]. SMOTE is one of the most popular methods for dealing with data distribution imbalances [17]. The SMOTE method finds the closest neighbor to data as much as K for each data in the minority class. After that, synthetic data is made as much as the desired percentage of duplication between minor data and K as much data as randomly selected neighbors [18].

B. Data scaling

Data scaling or normalization converts numeric values in a data set to a general scale without distorting differences in the range of values. Data normalization will help accelerate the learning process in machine learning. In the Abalone dataset, data normalization was carried out using min-max and zero-mean [14].

Min-max normalization changes the data size from the original range so that all values are in the range 0 and 1. The min-max normalization can be expressed in (1) where w_i is an original value, w_{norm} is a normalized value, w_{min} is a minimum value, and w_{max} is a maximum value.

$$w_{norm} = \left(\frac{w_i - w_{min}}{w_{max} - w_{min}} \right) \quad (1)$$

The Zero-Mean normalization method is based on the mean and standard deviation. Standardizing a dataset involves changing the value distribution scale so that the observed mean (mean) is 0 and the standard deviation is 1. Standard deviation is calculated using (2) where z_{mean} is the average of the data. Normalization can be calculated using (3) where z_i is an original feature vector, z_{mean} is the mean of the feature vector, z_{std} is a standard deviation, and $z_i^{'}$ is a result value of standardization.

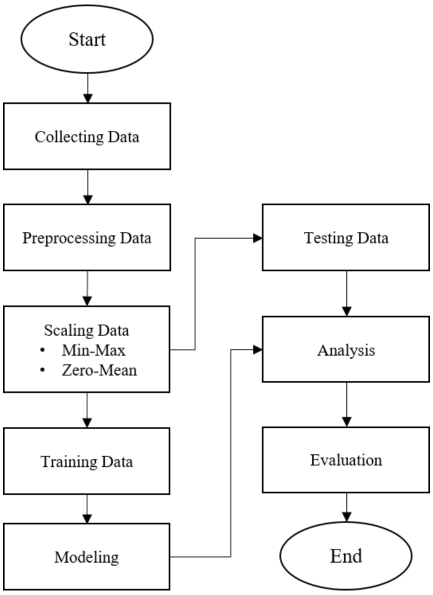

Figure 1. Research stages

$$z_{std} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(z_i - z_{mean} \right)^2} \quad (2)$$

$$z_i^{'} = \frac{z_i - z_{mean}}{z_{std}} \quad (3)$$

C. Classification

Data sharing was carried out in the dataset before and after normalization, with a composition of 90% training data and 10% test data. The model evaluation was carried out from the training data using k-fold cross-validation with k of 10. The model evaluation aimed to measure the accuracy of the model during training and to obtain a learning model. Validation is carried out from the model that has been made using test data.

The model evaluation and validation results are used to analyze the machine learning algorithm's performance. Four classifiers are used at the model evaluation stage and the making of a learning model, including the Random Forest, Naïve Bayesian, Decision Tree, and SVM. Tests were carried out using the programming language R 3.6.0 using the machine learning algorithm's random forest, party, and e1071 libraries.. Random Forest, Naïve Bayesian, and Decision Tree used the default configurations. SVM uses two kernels (linear and RBF), with values of C 1000 and gamma 0.25.

III. Results and Discussion

The abalones dataset consists of three classes (female 1307, infant 1342, and male 1528) and eight features (ring, length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight). Based on these detailed data, the data for each sex is not balanced. Thus, data from the female sex tends to be

[1] https://archive.ics.uci.edu/ml/datasets/abalone

Table 1. SMOTE result details

Sex	Total	
	Before	**After**
Female	1307	1528
Male	1342	1528
Infant	1528	1528

classified as noise or outlier when the classification process is carried out, damaging the modeling results. Therefore, to overcome this unbalanced data, the SMOTE method is used. SMOTE results can be seen in Table 1. It shows that the dataset of each sex class is balanced (1528). These results are obtained based on the workings of the SMOTE method, which finds the number of samples in one class and then equates them with samples in other classes.

A. Data scaling

Data scaling is required in machine learning only when features have different ranges, as in the abalone feature dataset. The abalone feature dataset has a different scale for ring features and seven features: length, diameter, height, whole weight, shucked weight, Viscera weight, and shell weight. The ring feature has a value range of 1 to 30, while the seven features range from 0 to 2.

Visualization of eight features from the Abalone feature dataset before data normalization is shown in Figure 2. It shows the scale difference between the ring and seven features. The ring features range from 1-30, while the seven features are in the 0-2 range. This difference can be minimized by normalizing the data. Data normalization is needed to optimize the performance of machine learning algorithms [14]. The smaller the value range of each feature, the faster the computation process will be performed [13].

The results of the min-max normalization on the Abalone feature dataset are shown in Figure 3. Min-max normalization changes the data into intervals of 0 to 1. Unlike the min-max normalization, at zero-mean normalization or standardization, the scale change is done by changing the average value (mean) to 0 and the standard deviation to 1. The scale of each feature is still different (it has different intervals). The visualization of the dataset of standardized abalone features is shown in Figure 4. From Figure 2 through Figure 4, it can be seen that the data pattern has not changed. Changes are visible only at the scale of each feature. It follows [13] that the data obtained before and after normalization does not change.

B. Classification

Before testing the normalized data sets, we first tested the data sets that had not been normalized on four machine learning algorithms (Random Forest, Naïve Bayesian, Decision Tree, and SVM). The evaluation uses 90% of the training data (4094) in 10-fold cross-validation. The model evaluation of the abalone feature dataset before normalization is shown in Figure 5.

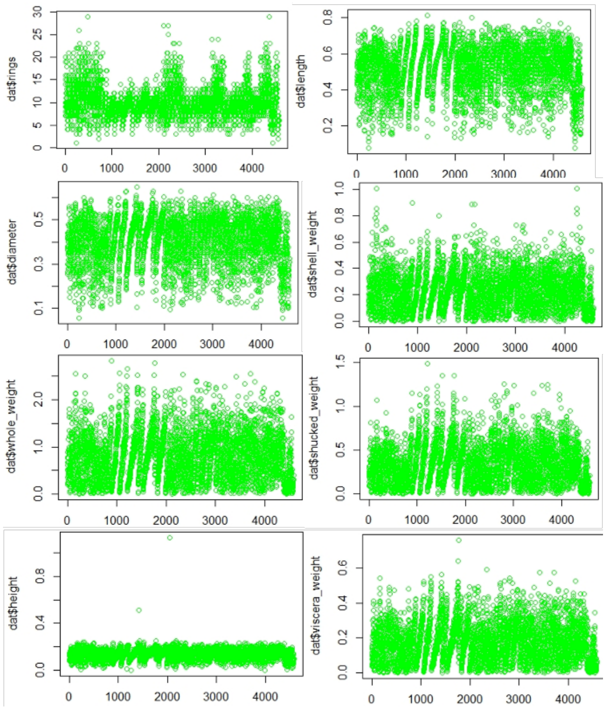

Figure 2. The distribution pattern of feature data before data scaling

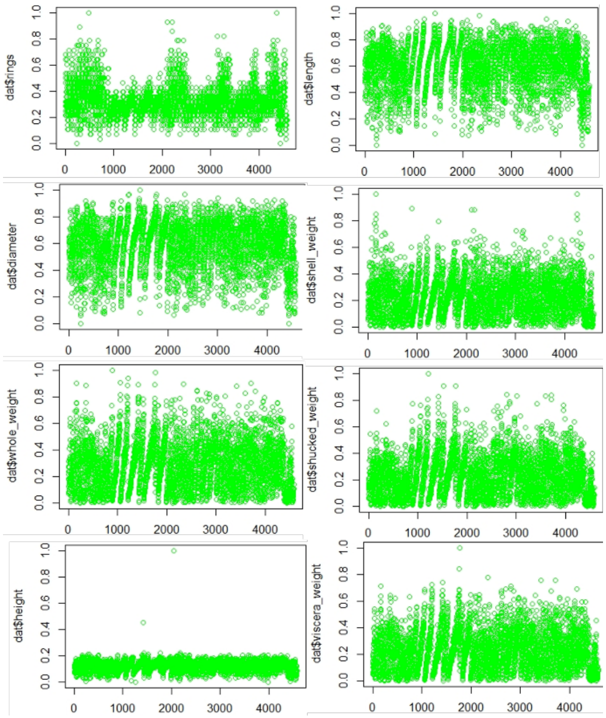

Figure 3. The distribution pattern of feature data using Min-Max normalization

In the abalone dataset before normalization, the highest to lowest average accuracy was obtained,

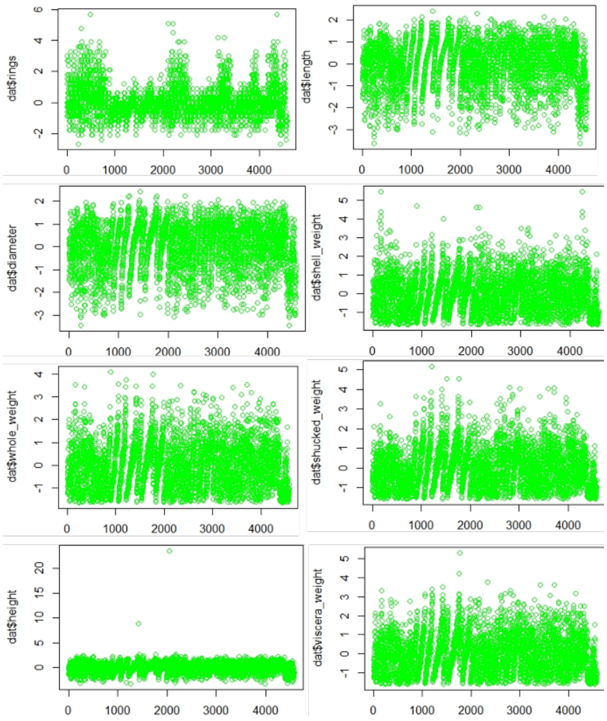

Figure 4. The distribution pattern of feature data using Zero-Mean standardization

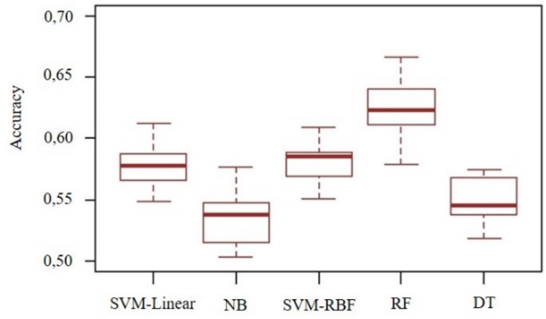

Figure 5. Models evaluation before data scaling

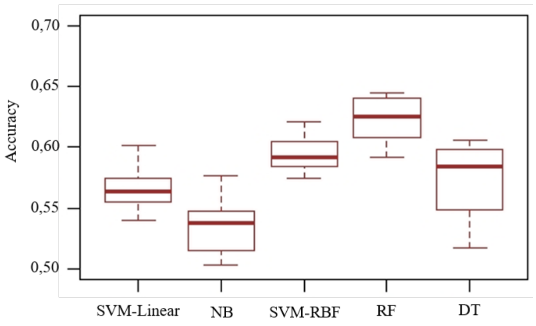

Figure 6. Models evaluation using Min-Max

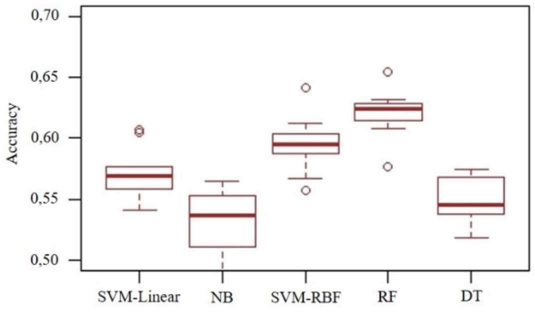

Figure 7. Models evaluation using Zero-Mean

Table 2. Performance comparison using dataset before data scaling

Algorithm	Balanced Accuracy (%)	Specificity (%)	Sensitivity (%)
SVM-RBF	68.45	78.35	58.56
SVM-Linear	67.96	78.74	57.20
Naïve Bayesian	66.31	77.25	55.38
Random Forest	**74.87**	**83.31**	**66.43**
Decision Tree	67.63	78.36	56.89

respectively, the Random Forest (62.53%), SVM kernel RBF (58.08%), SVM linear kernel (57.81%), Decision Tree (54.99%), and Naïve Bayesian (53.39). The model evaluation results with min-max and zero mean normalization are shown in Figure 6 and Figure 7.

The abalone dataset after min-max normalization has the highest to lowest average accuracy, respectively 62.37% (Random Forest), 59.49% (SVM kernel RBF), 57.20% (Decision Tree), 56.59% (SVM linear kernel), and 53.39% (Naïve Bayesian). It shows an increased performance on the SVM and Decision Tree algorithms. However, the increase in the accuracy value is not impactful by only 1-3%.

The abalone dataset after zero-mean normalization has the highest to lowest average accuracy: 62.08% (Random Forest), 59.50% (SVM kernel RBF), 57.12% (SVM linear kernel), 54.99% (Decision Tree), and 53.18% (Naïve Bayesian). It shows an increased performance on the SVM algorithms. However, the increase in the accuracy value is also not impactful by only 1-3%. The decrease in random forest accuracy when using min-max and zero-mean normalization is due to the inability to predict the range of response values in the training data [19].

The second trial results calculated each algorithm's average balanced accuracy, sensitivity, and specificity using 10% validation data (479 data). 90% of training data (4105 data) is used for learning on the five machine learning algorithms. It then validated using test data. The validation results on the abalone feature dataset before normalization are shown in Table 2.

The best performance is the Random Forest algorithm with a balanced accuracy of 74.87%, sensitivity of 66.43%, and specificity of 83.31%. Liu et al. [20] also found that Random Forest has better accuracy than other algorithms because Random Forest has a feature selection process. The process can take the best features to improve the performance of the classification model. This selection feature enables Random Forest to effectively work on large data with complex parameters. In addition, Random Forest can also work in parallel, known as multiple random forests [21].

Table 3. Performance comparison using min-max

Algorithm	Balanced Accuracy (%)	Specificity (%)	Sensitivity (%)
SVM-RBF	71.29	81.16	61.41
SVM-Linear	67.12	78.01	56.22
Naïve Bayesian	65.11	76.71	53.52
Random Forest	**73.29**	**82.15**	**64.42**
Decision Tree	64.63	76.27	53.00

Table 4. Performance comparison using zero-mean

Algorithm	Balanced Accuracy (%)	Specificity (%)	Sensitivity (%)
SVM-RBF	69.40	80.10	58.70
SVM-Linear	69.50	79.24	59.76
Naïve Bayesian	67.44	78.36	56.52
Random Forest	**71.67**	**81.22**	**62.12**
Decision Tree	68.90	78.87	58.93

The abalone feature dataset's validation results and the min-max normalization results are shown in Table 3. The Random Forest algorithm with an average balanced accuracy of 73.29%, 64.42% sensitivity, and 82.15% specificity. However, these results indicate that a decrease in performance occurs in the Random forest algorithm. Unlike without data scaling, the SVM kernel RBF algorithm shows performance improvement, where the average balanced accuracy is 71.29%, and the specificity is 81.16%. As stated in [14], the application of min-max normalization to the data to be tested using the SVM RBF kernel provides good performance in speed and accuracy.

The validation results on the abalone feature dataset and the zero-mean normalization results are shown in Table 4. The best performance is the Random Forest algorithm. However, like in min-max normalization, the Random Forest algorithm has decreased performance. The linear kernel SVM experienced an increase in performance from 67.21% using the min-max dataset to 69.50% using the Zero-mean dataset. The linear kernel SVM algorithm has increased when using a dataset of zero-mean normalization as indicated in [13].

IV. Conclusion

Naïve Bayes performs more more consistently than Random Forest, Decision Tree, and SVM when the abalone dataset is applied to both min-max and zero-mean normalization. There is an increase in the performance of SVM when the abalone dataset is applied to data normalization. SVM-Linear increased when using the zero-mean normalized dataset. SVM-RBF increased when using min-max normalization. On the other hand, in Random Forest, there is a decrease in performance when the abalone dataset is applied to both min-max and zero-mean normalization. However, Random Forest has the highest average balanced accuracy for all datasets.

References

[1] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, 2015. doi: 10.1126/science.aaa8415

[2] C. Beyan and H. I. Browman, “Setting the stage for the machine intelligence era in marine science,” *ICES Journal of Marine Science*, vol. 77, no. 4, pp. 1267–1273, 2020. doi: 10.1093/icesjms/fsaa084

[3] K. Malde, N. O. Handegard, L. Eikvil, and A. B. Salberg, “Machine intelligence and the data-driven future of marine science,” *ICES Journal of Marine Science*, vol. 77, no. 4, pp. 1274–1285, 2020. doi: 10.1093/icesjms/fsaa084

[4] L. Moitinho-Silva *et al.*, “Predicting the HMA-LMA status in marine sponges by machine learning,” *Frontiers in Microbiology*, vol. 8, no. 5, pp. 1–14, 2017. doi: 10.3389/fmicb.2017.00752

[5] Y. Shiu *et al.*, “Deep neural networks for automated detection of marine mammal species,” *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020. doi: 10.1038/s41598-020-57549-y

[6] L. Xu, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, *Deep learning for marine species recognition*, vol. 136. Springer International Publishing, 2019. doi: 10.1007/978-3-030-11479-4_7

[7] K. J. Wang, H. L. Ren, D. D. Xu, L. Cai, and M. Yang, “Identification of the up-regulated expression genes in hemocytes of variously colored abalone (Haliotis diversicolor Reeve, 1846) challenged with bacteria,” *Developmental and Comparative Immunology*, vol. 32, no. 11, pp. 1326–1347, 2008. doi: 10.1016/j.dci.2008.04.007

[8] A. B. A. Graf and S. Borer, “Normalization in support vector machines,” in *in Radig B., Florczyk S. (eds) Pattern Recognition*, 2001, pp. 277–278. doi: 10.1007/3-540-45404-7_37

[9] I. Ariawan, Y. Herdiyeni, and I. Z. Siregar, “Geometric morphometric analysis of leaf venation in four shorea species for identification using digital image processing,” *Biodiversitas*, vol. 21, no. 7, pp. 3303–3309, 2020. doi: 10.13057/biodiv/d210754

[10] I. Ariawan, Y. Herdiyeni, and I. Z. Siregar, “Geometry feature extraction of shorea leaf venation based on digital image and classification using random forest,” *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 1–10, 2021. doi: 10.12785/ijcds/110111

[11] A. Juneja and N. N. Das, “Big data quality framework: pre-processing data in weather monitoring application,” in *the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, Faridabad, India,* Oct. 2019, pp. 559–563. doi: 10.1109/COMITCon.2019.8862267

[12] A. Sahu, Z. Mao, K. Davis, and A. E. Goulart, “Data processing and model selection for machine learning-based network intrusion detection,” in *IEEE International Workshop Technical*

Committee on Communications Quality and Reliability, Stevenson, USA, May 2020, pp. 1-6. doi: 10.1109/CQR47547.2020.9101394

[13] A. Ambarwari, Q. Jafar Adrian, and Y. Herdiyeni, "Analisis pengaruh data scaling terhadap performa algoritme machine learning untuk identifikasi tanaman," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 1, pp. 117-122, 2020.

[14] W. Li and Z. Liu, "A method of SVM with normalization in intrusion detection," *Procedia Environmental Sciences*, vol. 11, no. PART A, pp. 256–262, 2011. doi: 10.1016/j.proenv.2011.12.040

[15] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford, *The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and the Islands of Bass Strait*. Tasmania: The Sea Fisheries Division, Marine Research Laboratories, 1994.

[16] A. T. Akbar, R. Husaini, B. M. Akbar, and S. Saifullah, "A proposed method for handling an imbalance data in classification of blood type based on Myers-Briggs type indicator," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 4, pp. 276–283, 2020. doi: 10.14710/jtsiskom.2020.13625

[17] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan teknik SMOTE untuk mengatasi imbalance class dalam klasifikasi objektivitas berita online menggunakan algoritma KNN," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 3, no. 2, pp. 196–201, 2019. doi: 10.29207/resti.v3i2.945

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. doi: 10.1613/jair.953

[19] P. Kashyap, *Machine Learning for Decision Makers*. Bangalore: Apress, 2017. doi: 10.1007/978-1-4842-2988-0

[20] M. Liu, M. Wang, J. Wang, and D. Li, "Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar," *Sensors and Actuators, B: Chemical*, vol. 177, pp. 970–980, 2013. doi: 10.1016/j.snb.2012.11.071

[21] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi random forest untuk klasifikasi motif songket palembang berdasarkan SIFT," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 7, no. 2, pp. 310–320, 2020. doi: 10.35957/jatisi.v7i2.289

©2022. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.