Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol 6, No.1, March 2023, pp. 123 – 130
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Data Sharing Technique for Electronic Health Record (EHR) Classification using Support Vector Machine Algorithm

Moh. Erkamim[1], Said Thaufik Rizaldi[2], Sepriano[3], Khoirun Nisa[4], Sulhatun[5], Zilrahmi[6], Winalia Agwil[7]

[1]Program Studi Sistem Informasi Kota Cerdas, Fakultas Teknik, Universitas Tunas Pembangunan Surakarta, Indonesia
[2]Puzzle Research Data Technology, Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim Riau, Indonesia
[3]Program Studi Sistem Informasi Fakultas Sains dan Teknologi, UIN Sulthan Thaha Saifuddin Jambi, Indonesia
[4]Program Studi Informatika, Fakultas Sains dan Teknologi, Universitas Harapan Bangsa, Indonesia
[5]Program Studi Teknik Kimia, Fakultas Teknik Universitas Malikussaleh, Indonesia
[6]Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Padang, Indonesia
[7]Prodi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Bengkulu, Indonesia

Email: [1]erkamim@lecture.utp.ac.id, [2]saidthaufik@uin-suska.ac.id, [3]sepriano@uinjambi.ac.id, [4]khoirunnisa@uhb.ac.id, [5]sulhatun@unimal.ac.id, [6]zilrahmi@fmipa.unp.ac.id, [7]winaliaagwil@unib.ac.id

Article history:
Received May 15th, 2023
Revised Jun 23th, 2023
Accepted Jul 20th, 2023

Keyword:
Classification
Data Sharing
Electronic Health Record
Radial Basis Function
Support Vector Machine

ABSTRACT

The Electronic Health Record (EHR) integrates information about medical history in patients, complications, and history of drug use efficiently, which demands optimality and speed of service for efficiency and effectiveness of services, especially in determining outpatient and inpatient services on accurate patient history data. In efforts to improve data accuracy, this study combined the c, γ, and degree kernels in the Linear, Polynomial, and Radial Basis Function (RBF) kernels as well as data sharing techniques 10-fold cross-validation, k-medoids, and Hold- out (70 % 30%) resulted in superior K-Medoids data sharing techniques for each Polynomial kernel with an accuracy of 75.76% and a Radial Basis Function (RBF) kernel with an accuracy of 75.56% so that it can be said that the combination of K-Medoids and Polynominal kernel in the algorithm Support Vector Machine (SVM) can be used in this research case.

Corresponding Author:
Moh. Erkamim
Program Studi Sistem Informasi Kota Cerdas, Fakultas Teknik,
Universitas Tunas Pembangunan Surakarta, Indonesia
Jl. Walanda Maramis No.31, Nusukan, Kec. Banjarsari, Kota Surakarta, Jawa Tengah, Indonesia.
Email: erkamim@lecture.utp.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v6i2.xxxxx

1. INTRODUCTION

Based on the Regulation of the Minister of Health of the Republic of Indonesia Number 269/MENKES/PER/2008 concerning Medical Records, it is explained that medical record data is very necessary for the maintenance and treatment of patients, health services, and health statistical data [1]. The Electronic Health Record (EHR) is a health information system that provides information about the patient's medical history, complications, and history of efficient drug use [2].

EHR implementation requires optimality, data accuracy, and service speed for service efficiency and effectiveness, especially in determining outpatient and inpatient services, as long as the use of this new system has never been evaluated (Dipura & Rahmadin, 2018). Based on the explanation above, it is necessary to evaluate data related to the implementation of the EHR system using Classification techniques in Data Mining. Classification is modeling in data mining that aims to form a function on an object based on the formation of its class [3].

According to Janardhanan et al (2015), Support Vector Machine (SVM) is an effective algorithm for processing classification using health datasets [4]. Previous research was conducted by Zhang (2019)

regarding the classification of Electronic Health Record (EHR) data for classifying cancer cases using the SVM algorithm, which resulted in an accuracy of 97.33% for determining the type of cancer with 400 data [5]. However, this study only used one main Radial Basis Function (RBF) kernel with a single data-sharing technique 10- fold cross-validation. Research by Mustakim et al (2016) uses more than one parameter and kernel optimization experiment, namely Linear, Polynomial, and Radial Basis Function (RBF), to maximize the performance results of the Support Vector algorithm. [6].

However, this study did not specifically mention the data-sharing techniques used in the study. Research by Mustakim et al (2019) compared the Hold-out data division technique (70% 30%), K-Means Clustering, and 10- fold cross validation in the classification algorithm resulting in 10-fold cross-validation with an accuracy of 97.40% [7]. This study compared the 10-fold cross-validation, k- medoids, and hold-out (70% 30%) data-sharing techniques with a combination of C, γ, and degree parameters in the Linear, Polynomial, and Radial Basis Function (RBF) kernels.

2. MATERIAL AND METHOD

This research starts from the Planning Stage, where the problem identification is carried out, determines the research objectives explicitly, and the boundaries applied. The data collection stage was carried out by observing Electronic Health Record (EHR) data by Sujiono (2020) for cases of classification of determining inpatients and outpatients in private hospitals in Indonesia [8]. The attributes in this dataset include Haematocrit, Haemoglobins, Erythrocyte, Leucocyte, Thrombocyte, MCH, HCHC, MCV, Age, Gender, and Source as inpatient class attributes (in) or outpatient (out). The following is the dataset used in this study in Table 1. The research dataset is as follows:

Table 1. Research Datasets

NO	HCT	HB	RBC	WBC	TB	MCH	MCHC	MCV	AGE	SEX	SOURCE
1	35,10	11.80	4.65	6.30	310.00	25,40	33,60	75.50	1	F	out
2	43.50	14.80	5,39	12.70	334.00	27.50	34.00	80,70	1	F	out
3	33.50	11.30	4.74	13,20	305.00	23.80	33,70	70,70	1	F	out
4	39,10	13.70	4.98	10.50	366.00	27.50	35.00	78.50	1	F	out
5	30,90	9.90	4,23	22,10	333.00	23,40	32.00	73.00	1	M	out
6	34,30	11.60	4.53	6,60	185.00	25,60	33.80	75,70	1	M	out
7	31,10	8.70	5.06	11,10	416.00	17,20	28.00	61.50	1	F	out
8	40,30	13.30	4.73	8,10	257.00	28,10	33.00	85,20	1	F	out
9	33,60	11.50	4.54	11.40	262.00	25.30	34,20	74.00	1	F	out
10	35,40	11.40	4.80	2.60	183.00	23.80	32,20	73,80	1	F	out
…	…	…	…	…	…	…	…	…	…	…	…
4408	32.80	10,40	3.49	8,10	72.00	29.80	31.70	94.00	92	F	in
4409	33,70	10.80	3.67	6,70	70.00	29,40	32.00	91.80	92	F	in
4410	33,20	11.20	3,47	7,20	235.00	32,30	33,70	95.70	93	F	out
4411	31.50	10,40	3,15	9,10	187.00	33.00	33.00	100.00	98	F	in
4412	33.50	10.90	3,44	5.80	275.00	31.70	32.50	97.40	99	F	out

Preprocessing process and the distribution of training and test data, as shown in Figure 1, and carried out experiments on the c, gamma, and degree parameters on the Support Vector Machine (SVM) kernel to evaluate the accuracy of the classification. The following is the research methodology contained in Figure 1.

2.1 Electronic Health Record (EHR)

Electronic Health Record (EHR) is an integrated health system that contains patient information and history, gender and laboratory results, as well as medical history that is managed and issued by official medical parties or other health care units to be accessed instantly and safely [5]. EHR is used by medical parties such as doctors for clinical purposes to assist patients in managing individual patient care and monitoring in hospitals [9].

2.3 K-Medoids

K-Medoids or Partitioning Around Medoids (PAM) are possible representative or non-representative grouping combinations of the best iteration trial sets that form a new medoid. Determination of the central value in the cluster is calculated using the Euclidean Distance equation with the following equation [14] .

$$d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - x_{kj})^2} \quad (4)$$

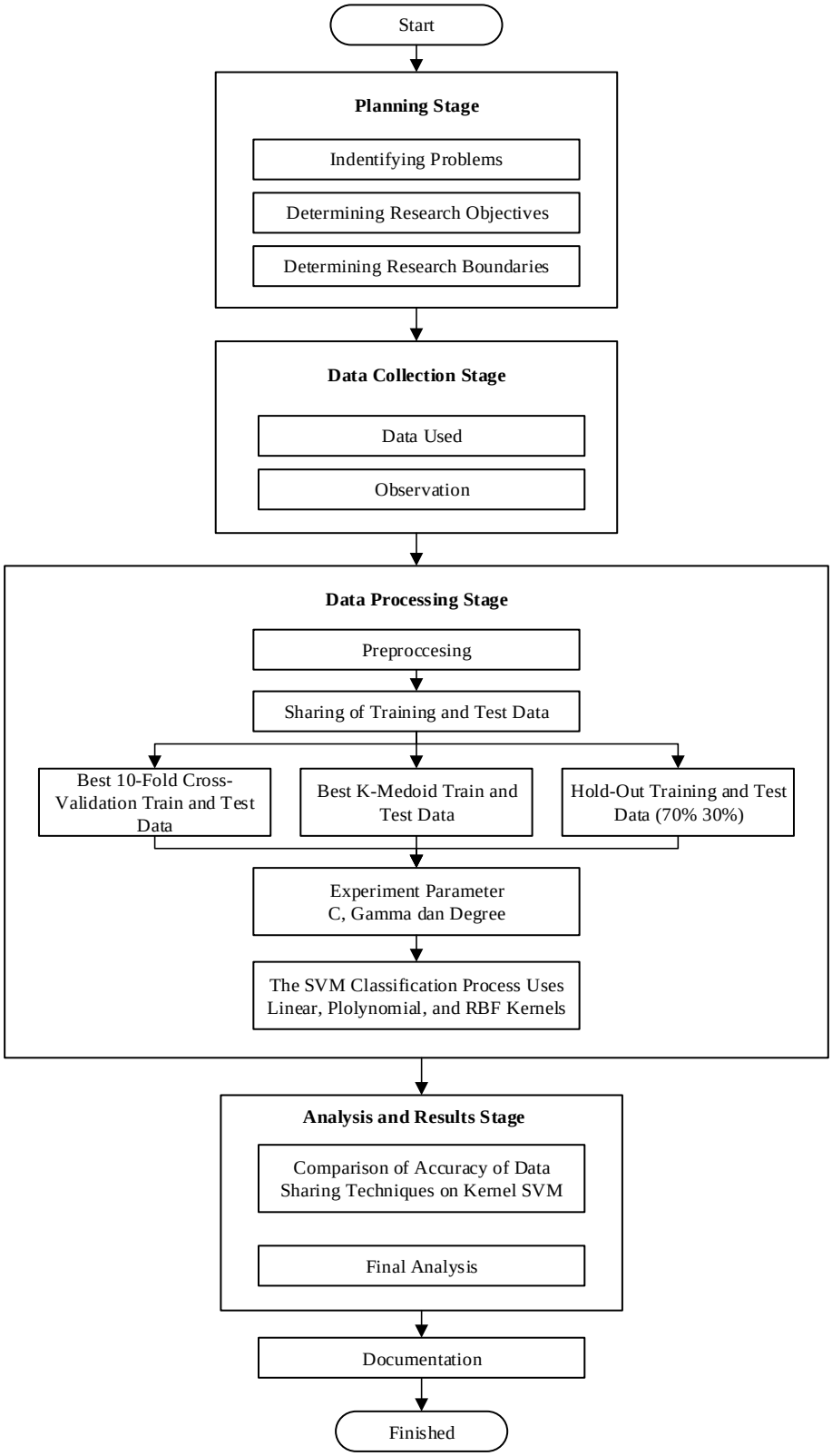

Figure 1. Research Methodology

2.2 Support Vector Machine (SVM)

SVM is a method in Machine Learning that is very good at recognizing subtle patterns in complex data formations compared to other Machine Learning methods [10]. Support Vector Machine (SVM) was introduced by Vapnik (1992) to solve the Lagrangian equation in quadratic programming, where class

separation is determined based on the hyperplane in the input space by measuring the hyperplane margins to find the maximum point. SVM can get a hyperplane to separate the points into different classes [5]. In non-linear cases, SVM is non-probabilistic and kernel-based [11]. There are several kernel functions stated in K(x_i,x_j) which are used in SVM, including the following [12] :

Linear Kernels

$$K(x_i, x_j) = x_i' x_i + c \quad (2)$$

Kernel Radial Basis Function

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2_{a^2}}\right) \quad (3)$$

Polynomial Kernels

$$K(x_i, x_j) = (x_i' x_i + c)^P \quad (4)$$

2.4 Hold-Outs

Hold-out is a technique for dividing data into two parts, namely training data and test data based on percentage [15]. Hold-out data-sharing techniques are commonly used in test cases in classification algorithms [16].

2.5 K-Fold Cross Validation

K- Fold Cross Validation is a method that randomly divides the desired subsets, which will produce a subset of the training data and test data [13]. The following is an illustration of the distribution of k-fold cross-validation with a value of k = 4 in Figure 2.

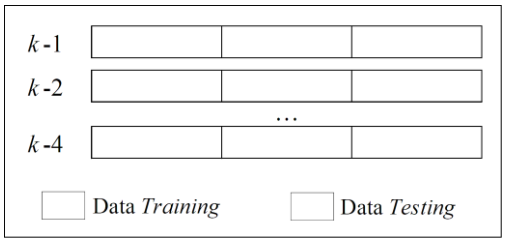

Figure 2. Division k-4 k- Fold Cross Validation

2.6 Confusion Matrix

Confusion Matrix is a technique for determining the performance and performance of an algorithm, especially a classification where this evaluation technique consists of accuracy, precision, and recall. [17] . Accuracy is a measure of the performance and effectiveness of the algorithm. Precision is the potential for positive results from a comparison of the original data amount. Recall is the positive equivalent of the original data amount to the correct predicted algorithm model used [18].

2 RESULT AND ANALYSIS

3.1 Linear Kernels

The parameter used in the Linear Kernel is the C value, where a combination of experiments is carried out on the C parameter values to obtain the most optimal accuracy results, as shown in Table 2.

Table 2. Linear Kernel Experiment

C	Data Sharing		
	10-Fold Cross Validation	K-Medoids	Hold-outs (70% 30%)
1	70.55 %	70.37 %	67.89 %
10	71.23 %	70.77 %	69.22 %
100	71.38 %	71.08 %	69.33 %
1000	71.32 %	70.98 %	69.63 %

Based on the above experiments conducted on the Linear Kernel, the 10-Fold Cross Validation data sharing technique with a parameter value of C = 100 was the best experiment with the highest accuracy of 71.38%. In comparison, the Hold-out data-sharing technique is 70% training data and 30% test data as the lowest accuracy with a parameter value of C = 1.

3.2 Polynomial Kernels

The Polynomial kernel uses C, γ (gamma), and degrees in determining parameter values in the Support Vector Machine classification. The degree value used in this experiment by default is determined by the degree value = 2. To optimize the accuracy results in the Kernel Polynomial experiment, an experiment was carried out on a combination of these parameters based on research conducted by Ridwan (2019) in Table 3.

Table 3. Polynomial Kernel Experiments

C	γ	Degrees	Data Sharing		
			10-Fold Cross Validation	K-Medoids	Hold-outs (70% 30%)
1	0.01	2	60,25	60,18	60,22
10	0.01	2	60,25	60,18	60,22
100	0.01	2	66,26	64,66	63.50
1000	0.01	2	70.55	70,26	68.00
1	0.1	2	66,26	64,66	63.50
10	0.1	2	70.58	70,26	68.00
100	0.1	2	74,29	74,44	72,60
1000	0.1	2	74.85	74,44	73,52
1	1	2	74,29	74,44	72,60
10	1	2	74.85	74,44	73,52
100	1	2	74,66	75.25	73.01
1000	1	2	75.09	75,76	73,31

Based on the above experiments carried out on the Polynomial Kernel, the K-Medoids data division technique with parameter values C = 1000, γ = 1**,** and degree = 2 was the best experiment with the highest accuracy of 75.76%. While the Hold-out data sharing technique is 70% training data and 30% test data as the lowest accuracy with parameter values C = 1, γ = 0.01**,** and degree = 2 and C = 10, γ = 0.01**,** and the value of degree = 2 each of 60.22%

3.3 Kernel Radial Basis Function (RBF)

Kernel Radial Basis Function (RBF) using parameters C and γ (gamma) is commonly used in determining the performance of a Support Vector Machine (SVM) [19]. The combination of these parameters produces optimal accuracy in the RBF Kernel Support Vector Machine (SVM) algorithm, which is shown in Table 4.

Table 3. Kernel Radial Basis Function (RBF) Experiment

C	γ	Data Sharing		
		10-Fold Cross Validation	K-Medoids	Hold-outs (70% 30%)
1	0.01	60,34	60,18	60,12
10	0.01	68,53	68,53	67,28
100	0.01	70,89	70,67	68,81
1000	0.01	72,76	72,40	70,86
1	0.1	68,44	68,53	66,97
10	0.1	72.06	71.49	69,84
100	0.1	74.75	75,15	69,84
1000	0.1	74,69	75.25	73,93
1	1	74,51	73,32	72,60
10	1	75,28	75,56	74,44
100	1	75,40	75,36	74,23
1000	1	75.06	74,44	73,52

Based on the above experiments carried out on the Kernel Radial Basis Function (RBF), the K-Medoids data division technique was produced with a parameter value of C = 10 and a value of γ = 1 as the best trial with the highest accuracy of 75.56%. In comparison, the Hold-out data sharing technique is 70% training data and 30% test data as the lowest accuracy with a parameter value of C = 1 and a value of γ = 0.01 of 60.12%.

The experiments were carried out using the Kernel Linear, Polynomial, and Radial Basis Function (RBF), respectively the best experiments from a combination of data sharing techniques with C, γ, and degree parameters on the Support Vector Machine (SVM) can be seen in Figure 3.

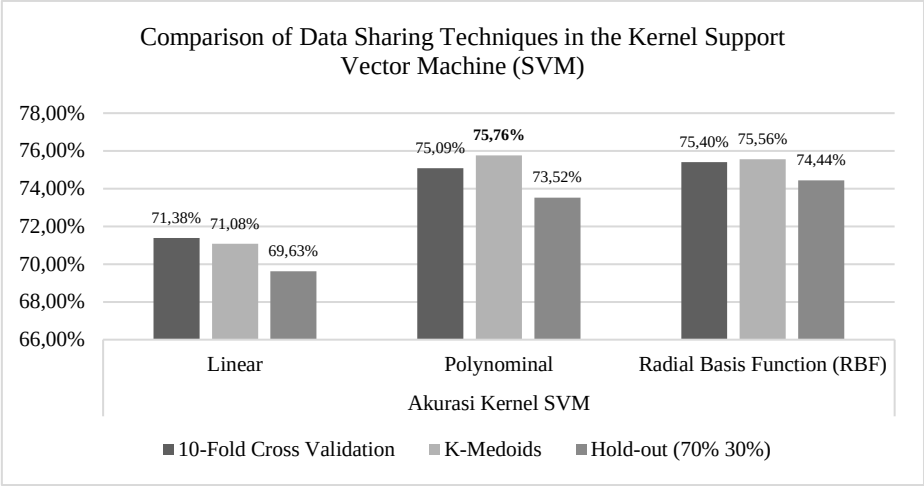

Figure 3. Comparison of Data Sharing Techniques in the SVM Kernel

Evaluation of the Support Vector Machine (SVM) algorithm using the Linear kernel produces 10-Fold Cross Validation with an accuracy of 71.38%, Polynominal produces K-Medoids with an accuracy of 75.76%, and Radial Basis Function (RBF) with K-Medoids with an accuracy of 75 .56% as the best data sharing technique trial in the Electronic Health Record (EHR) case in this study.

3 CONCLUSION

Based on the results of the research and explanation above for the case of comparison of data sharing techniques in the Support Vector Machine (SVM) algorithm, K-Medoids \\are superior in Polynominal kernels of 75.76% and Radial Basis Function (RBF) kernels of 75.56%, so it can be concluded that using the K-Medoids data sharing technique with the Polynomial kernel the SVM algorithm can be used in this research case.

REFERENCES

[1] N. Munsir, N. Yuniar, F. Nirmala, and Suhadi, "Analysis of Completion of Medical Record Documents for Bpjs Inpatients at Dewi Sartika General Hospital in 2017," Jimkesmas , *vol* . VOL. 3/NO., pp. 1–7, 2018.

[2] B. Shickel, PJ Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE J. Biomed. Heal. Informatics* , vol. 22, no. 5, pp. 1589–1604, 2018, doi: 10.1109/JBHI.2017.2767063.

[3] Y. Cheng, K. Chen, H. Sun, Y. Zhang, and F. Tao, "Data and knowledge mining with big data towards smart production," *J. Ind. inf. Integr.* , vol. 9, no. October, pp. 1–13, 2018, doi: 10.1016/j.jii.2017.08.001.

[4] P. Janardhanan, L. Heena, and F. Sabika, "Effectiveness of support vector machines in medical data mining," *J. Commun. softw. syst.* , vol. 11, no. 1, pp. 25–30, 2015, doi: 10.24138/jcomss.v11i1.114.

[5] X. Zhang, J. Xiao, and F. Gu, "Applying support vector machine to electronic health records for cancer classification," *Simul. Ser.* , vol. 51, no. 5, 2019, doi: 10.23919/SpringSim.2019.8732906.

[6] M. Mustakim, A. Buono, and I. Hermadi, "Performance Comparison Between Support Vector Regression and Artificial Neural Network for Prediction of Oil Palm Production," *J. Computer Science. and Inf.* , vol. 9, no. 1, p. 1, Feb. 2016, doi: 10.21609/jiki.v9i1.287.

[7] Mustakim *et al.* , "Data Sharing Technique Modeling for Naive Bayes Classifier for Eligibility Classification of Recipient Students in the Smart Indonesia Program," *J. Phys. Conf. Ser.* , vol. 1424, no. 1, 2019, doi: 10.1088/1742-6596/1424/1/012009.

[8] S. Mujiono, "EHR Dataset for Patient Treatment Classification," 2020.

[9] KR Pradeep and NC Naveen, "Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics," *Procedia Comput. sci.* , vol. 132, pp. 412–420, 2018, doi: 10.1016/j.procs.2018.05.162.

[10] S. Huang, CAI Nianguang, P. Penzuti Pacheco, S. Narandes, Y. Wang, and XU Wayne, "Applications of support vector machine (SVM) learning in cancer genomics," *Cancer Genomics and Proteomics* , vol. 15, no. 1, pp. 41–51, 2018, doi: 10.21873/cgp.20063.

[11] M. Bernardini, L. Romeo, P. Misericordia, and E. Frontoni, "Discovering the Type 2 Diabetes in

Electronic Health Records Using the Sparse Balanced Support Vector Machine," *IEEE J. Biomed. Heal. Informatics* , vol. 24, no. 1, pp. 235–246, 2020, doi: 10.1109/JBHI.2019.2899218.

[12] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing , *no* . xxxx, 2020, doi: 10.1016/j.neucom.2019.10.118.

[13] SD Agustina, Mustakim, Okfalisa, C. Bella, and MA Ramadhan, "Support Vector Regression Algorithm Modeling to Predict the Availability of Foodstuff in Indonesia to Face the Demographic Bonus," J. Phys . *Conf. Ser.* , vol. 1028, no. 1, 2018, doi: 10.1088/1742-6596/1028/1/012240.

[14] X. Jin and J. Han, "K-Medoids Clustering," *Encyclopedia. Mach. Learn. Data Min.* , pp. 697–700, 2017, doi: 10.1007/978-1-4899-7687-1_432.

[15] ST Rizaldi and M. Mustakim, "Comparison of Data Sharing Techniques for Classifying Water Access Facilities in the K-Nearest Neighbor Algorithm and Naïve Bayes Classifier," in *National Seminar on Information Technology, Communication and Industry (SNTIKI) 12* , 2020, pp. 130–137.

[16] A. Ghazvini, J. Awwalu, and A. Abu Bakar, "Comparative Analysis of Algorithms in Supervised Classification: A Case study of Bank Notes Dataset," *Int. J.Comput. Trends Technol.* , vol. 17, no. 1, pp. 39–43, 2014, doi: 10.14445/22312803/ijctt-v17p109.

[17] MS Pervez and DM Farid, "Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs," *Ski. 2014 - 8th Int. Conf. Software, Knowledge, Info. Manag. appl.* , 2014, doi: 10.1109/SKIMA.2014.7083539.

[18] S. Dewi, "Comparison of 5 Data Mining Classification Algorithm Methods in Predicting the Success of Banking Service Product Marketing," *None* , vol. 13, no. 1, pp. 60–66, 2016.

[19] Y. Xie, YLM B, and DS Kochhar, *SVM Parameter Optimization Using Swarm Intelligence for Learning from Big Data* , vol. 1. Springer International Publishing, 2018.

BIBLIOGRAPHY OF AUTHORS

Moh. Erkamim, S.Kom., M.Kom, Lecturer in the Smart City Information Systems Study Program, Faculty of Engineering, Tunas Pembangunan University, Surakarta. The author teaches web programming, data structures, information technology governance, and frameworks.

Said Thaufik Rizaldi, Member Researcher in Puzzle Research Data Technology, Faculty of Science and Technology, State Islamic University of Sultan Syarif Kasim Riau. This researcher has a research focus and has also published deep learning research publications for health care.

Sepriano, M.Kom. Lecturer in the Information Systems Study Program, Faculty of Science and Technology, Sulthan Thaha Saifuddin State Islamic University, Jambi. The author teaches web programming and technopreneurship courses

Khoirun Nisa, Lecturer in Informatics Study Program, Faculty of Science and Technology, Harapan Bangsa University, Surakarta. She focuses on the research field of Artificial Intelligence, SPK, and Computer Vision

Dr. Sulhatun, ST, MT, Lecturer in the Chemical Engeneering Study Program, Faculty of Engeneering, Malikussaleh University, Lhokseumawe, Aceh.

Zilrahmi is a lecturer in the Statistics Department at Universitas Negeri Padang. Her research focuses on machine learning and big data analysis.

Winalia Agwil is a lecturer in the Statistics Study Program at Bengkulu University. Her research focuses on machine learning and big data analysis.