JurnalMantik
Volume 4, Number 4, Feb 2021, pp. 2499-2507 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik

Performance Comparison of Data Mining Algorithms Which Occupy the Top:C4.5 and SVM

Indri Tri Julianto [1], Ricky Rohmanto [2], Ujang Sarifudin [3],Fanji Fakhru Zaman [4]

Jurusan Sistem Informasi Bisnis, Sekolah Tinggi Manajemen Ilmu Komputer LIKMI
Jl. Ir. H. Juanda No.96, Lebakgede, Kecamatan Coblong, Kota Bandung, Indonesia, (+62)222502121

E-mail: indritrijulianto@gmail.com[1], rickyrohmanto@gmail.com[2], ujangedosarifudin@gmail.com[3], fanjifakhru@gmail.com[4]

ARTICLEINFO

Article history:
Received: 01/01/2021
Revised: 10/01/2021
Accepted: 15/01/2021

Keywords:
Algorithm, C4.5, Data Mining, SVM

ABSTRACT

Data is a collection of various kinds of facts that are stored but do not have meaning. Mining is a mining process. So data mining can be interpreted as the process of mining large and complex amounts of data for new knowledge or information that can be useful for data owners. There is a sequence of systematic ways to solve problems in Data Mining, known as Data Mining algorithms. The IEEE International Conference on data mining which was conducted in 2006 produced the 10 most frequently used data mining algorithms by the research community around the world. Two of the ten most commonly used algorithms are the C4.5 algorithm and the Support Vector Modeling (SVM) algorithm. The methodology used in this research is The Knowledge Discovery in Database (KDD) stage. This study aims to compare the C4.5 with the SVM in terms of performance where what will be seen is the value of Area Under Curve (AUC), Receiver Operating Characteristic (ROC), Accuracy, Error, Precison, and Recall.

1. Introduction

Data mining is the mining process of data sets that aims to produce knowledge output [1]. Data mining is inseparable from several disciplines that support the process, such as statistics, visualization, algorithms, pattern recognition, machine learning, and database technology [1]. Algorithms have a role to solve problems in data mining, so there are 10 algorithms that are often used by researchers in data mining, namely the C4.5 Algorithm, K-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, and CART [2] Based on the 10 ranking algorithms, then the one who is in the top rank will be compared with one another, namely the C4.5 Algorithm with the SVM Algorithm.

Several studies discussing the application of the Data Mining algorithm, namely research on the application of the C4.5 algorithm in identifying the factors that cause accidents in a construction company [3], some are comparing the C4.5 algorithm with Id3 and Random Forest [4], then some are discussing the advantages and disadvantages of the SVM algorithm [5], then some use the SVM algorithm in the selection of a Vocational High School [6] and finally the application of SVM in the field of schizophrenia psychiatric diseases [7].

Based on previous studies, this study fills the research gap by comparing the performance of the two classification algorithms that have the highest ranking, namely C4.5 with SVM. The goal is to see the performance of each based on AUC, ROC performance, accuracy, error, precision, and recall. When the performance value is out, it will be known which algorithm has a better performance.

2. Literature Review

This study uses the Performance Comparison of Data Mining Algorithms Which Occupy the Top: C4.5 And SVM.

2.1 Literature Method

The method of collecting library data is done by collecting data from sources or books that are relevantto the research.

2.2 Interview Method

The interview method is carried out by asking directly to the data manager and Performance Comparison of Data Mining Algorithms C4.5 And SVM

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

3. Research Methodology

The research method used is Knowledge Discovery in Database (KDD). KDD is a data mining method for extracting knowledge that can be following size specifications and limits, which uses a database together with the necessary preprocessing, sampling, and transformation of the database [8]. There are 5 stages in this methodology as shown in Fig 1.

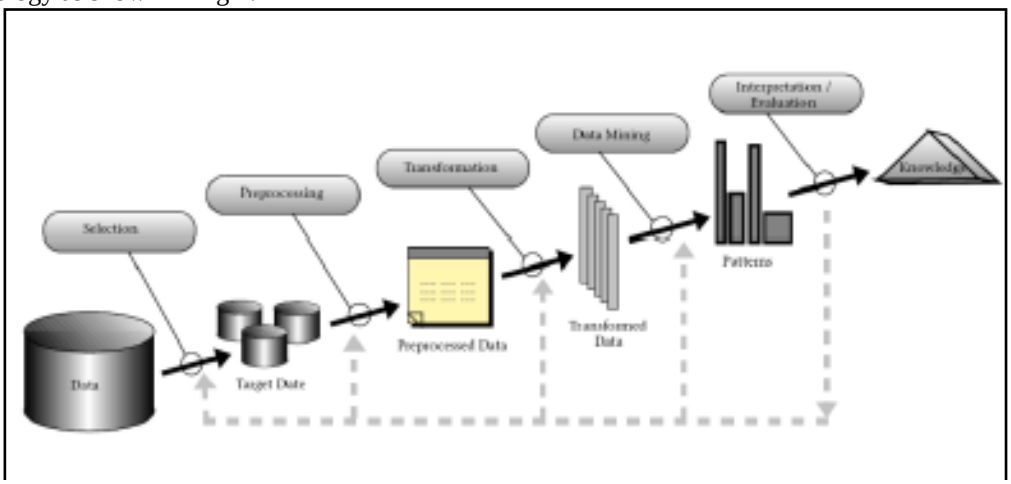

Fig 1. Stages of KDD [8]

Classification is a technique that looks at the behavior and attributes of a defined group, where this technique can provide classification for new data by manipulating classified data and by using results that provide a number of rules [9]. One easy and popular example is the decision tree, which is one of the most popular classification methods because it is easy to interpret. Decision tree is a prediction model using a tree structure or hierarchical structure.

The algorithm used in this research is the classification algorithm using C4.5 and SVM techniques. The two algorithms were chosen because they are based on a literature study conducted in the book The Top Ten Algorithms In Data Mining which states the ranking of the Data Mining algorithm. The ratings are as follows:

a. C4.5;
b. K.Means;
c. SVM.
d. Apriori;
e. EM;
f. PageRank;
g. AdaBoost;
h. kNN;
i. Naïve Bayes; and
j. CART

This study uses a classification technique so that C4.5 and SVM were chosen because they are data-mining classification algorithms. The explanation of the algorithm is as follows:

3.1 C4.5 Algorithm

The C4.5 algorithm is an algorithm used to classify data that has numeric and categorical attributes. The result of the classification process is in the form of rules that can be used to predict the value of the discrete type attribute from the new record. Algorithms C4.5 is a development of the ID3 algorithm [3].

In general, the C4.5 algorithm for building a decision tree is as follows:

a. Select attribute as root;
b. Create a branch for each value;
c. Divide cases into branches; and
d. Repeat the process for each branch until all cases on the branch have the same class.

To select the root attribute, based on the highest acquisition value of the existing attributes. To calculate gain, a formula is used as shown in equation 1 below [10]:

$$Gain\ (S, A) = Entropy\ (S) - \sum \frac{|Sv|}{|S|}\ Entropy\ (Sv) \quad (1)$$

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Where:

S = Case Set.
A =Attribute.
[Sv] = The number of samples for the value v.
[S] = The total sample data.

Entropy is the diversity of data, where the formula is as shown in equation (2) below:

$$Entropy\ (S) = -\sum pi\ log2\ pi\ ....\qquad (2)$$

Where:

pi = the portion or ratio between the number of samples of class i with the number of all samples in the data set

3.2 SVM Algorithm

This algorithm is a technique invented in (1995) which is intended as an algorithm for making predictions, both in a classification and regression classification [6]. SVM is included in the supervised learning class. In the real method, SVM is a method that uses a hyperplane which is used as a linear separator between data, therefore to solve data problems that are not linear (nonlinear), the kernel trick technique can be used [7]. The linear separation area in SVM is presented in the form of an image as shown in Fig 2.

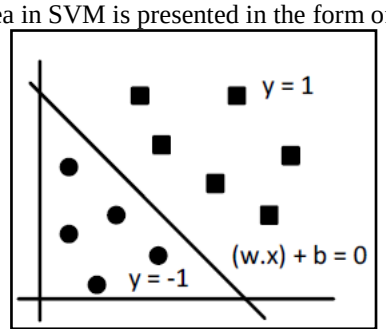

Fig 2. SVM Linear Separation Plane [10]

3.3 Performance Assessment

Binary classification is a statistical model and calculation that divides a data set into two groups, positive and negative. The Confusion Matrix is used to explain the performance measurement of the classification technique. The Confusion Matrix is presented in tabular form shown in Table 1.

Table 1.
Confusion matrix[10]

Actual	**Prediction**	
	Positive	**Negative**
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Performance assessment is included in the evaluation stage. Some of the performance values for the classification technique are Area Under Curve (AUC), Receiver Operating Characteristics (ROC) Accuracy, Error, Precision, and Recall. Accuracy is a measure of the ratio of the correct prediction to the total number of samples evaluated. Error is a measure of the ratio of incorrect predictions to the total number of samples evaluated. Precision is the level of accuracy between the information served by the user and the answers given by the system. A recall is the level of confidence in the system in retrieving information. The AUC is calculated to measure performance. Data Mining Classification, AUC values can be divided into several groups [10]:

a. 0.90-1.00 = Very Good Classification.
b. 0.80-0.90 = Good Classification
c. 0.70-0.80 = Sufficient Classification
d. 0.60-0.70 = Bad Classification
e. 0.50-0.60 = Incorrect classification

The formula is shown in equation 3 [10].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}\qquad (3)$$

$$\text{Error} = \frac{FP+FN}{TP+TN+FN+FP}\qquad (4)$$

Where:

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

TP : True Positive
TN : True Negative
FP : False Positive
FN : False Negative

The ROC is in the form of a curve that will show the accuracy and visually compare the classification with the false positive rate (specificity) as a horizontal line and the true positive rate (sensitivity) as a vertical line, lazy is to determine the intersection point on a continuous diagnostic test [10].

3.4 Rapidminer

Tests are carried out on the two algorithms using the Rapid Miner application. Rapid Miner is an open source software, which is an application for analyzing data mining, text mining and predictive analysis [9].

3.5 Research Stages

The stages of this research are based on the Knowledge Discovery methodology in the Database, where these stages are presented in the form of an image as shown in Fig 3.

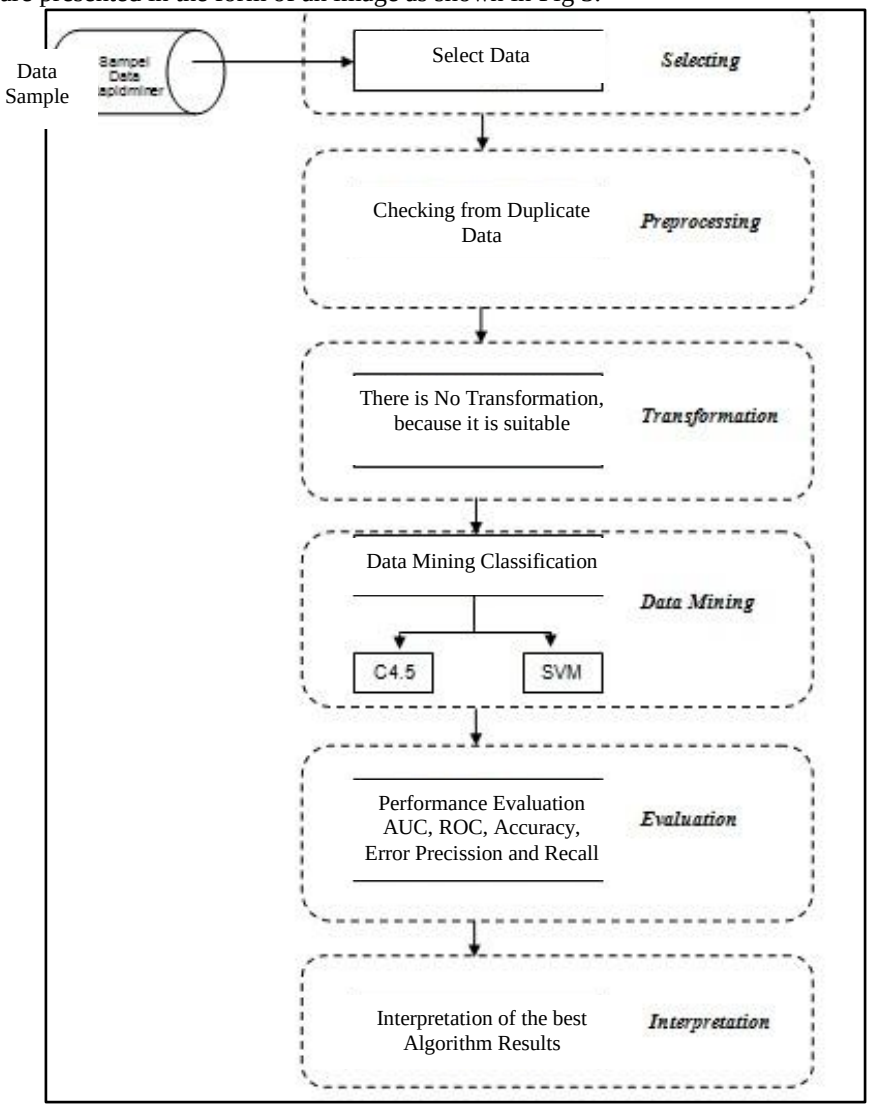

Fig 3. Research Stages [10]

The data set provided is sample data available in the RapiMiner application, namely golf testing data and golf training data. Both algorithms were tested with the same data to produce AUC, ROC, accuracy, error, precision, and recall values.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

4. Results and Discussion

4.1 Data Selection

The data used is sample data sourced from the Rapidminer application where the name of the data source is Play Golf Data. The data is presented in the form of an image shown in Fig 4.

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	yes	sunny	85	85	false
2	no	overcast	80	90	true
3	yes	overcast	83	78	false
4	yes	rain	70	96	false
5	yes	rain	68	80	true
6	no	rain	65	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	no	sunny	75	80	false
11	yes	sunny	68	70	true
12	yes	overcast	72	90	true
13	no	overcast	81	75	true
14	yes	rain	71	80	true

Fig 4. *Dataset Play Golf*

4.2 Preprocessing Data

Checks are carried out to avoid redundant or duplicate data. The results of the check show that the golf play data is ready for use.

4.3 Transformation

Golf play data is sample data that has been provided by the Rapidminer application, so that this stage does not carry out a transformation of the data to be processed because it is already being used.

4.4 Data Mining

The Data Mining process is carried out by doing the modeling stage for the C4.5 and SVM algorithms as follows:

a. Testing the C4.5 Algorithm

The testing phase is carried out by entering a golf data set in the Rapidminer process area. The model used is a decision tree for this C4.5 algorithm. The model is shown in picture form, shown in Fig 5.

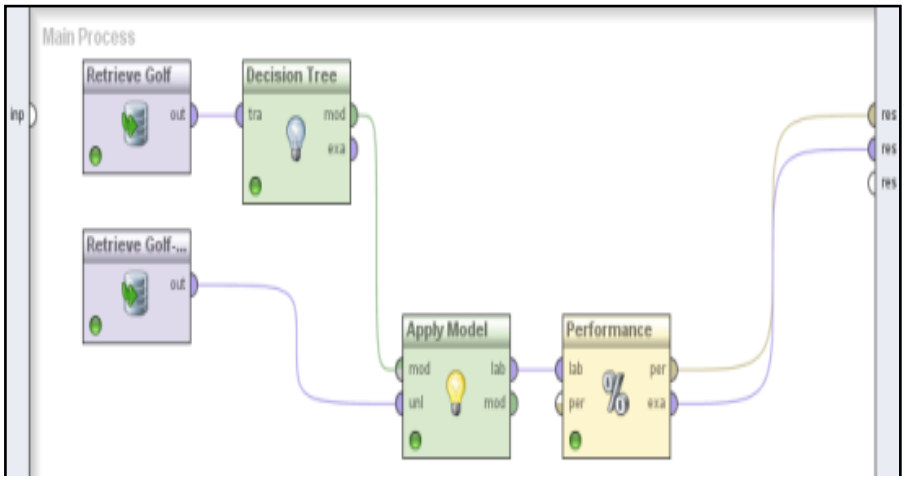

Fig 5. C4.5 Process Model

The results of C4.5 testing with classification techniques can be seen from the Confusion Matrix. The Confusion Matrix is presented in tabular form shown in Table 2.

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

Table 2.
C4.5 Confusion Matrix

Actual	Prediction	
	YES	**NO**
YES	3	3
NO	2	6

Information:

1) The actual number of data that MATCH and predictably ACCORD is 3.
2) The amount of data that actually DOES NOT MATCH & is predicted to NOT MATCH is 6.
3) The actual amount of data that DOES NOT MATCH and predicted to MATCH is 2.
4) The actual amount of data that MATCH and predictably DOES NOT MATCH is 3.

b. Testing on the SVM Algorithm

The testing phase at SVM begins by changing the data from nominal to numeric form. The SVM Process Model is shown in Fig 6.

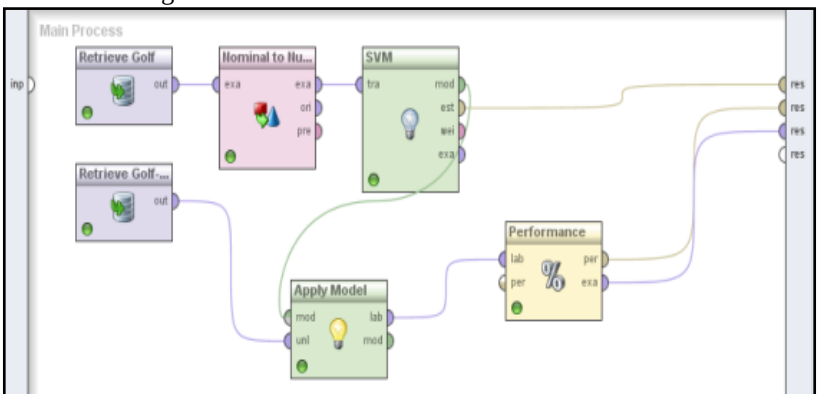

Fig 6. SVM Process Model

The results of SVM testing with classification techniques can be seen from the Confusion Matrix. The Confusion Matrix is presented in tabular form shown in Table 3.

Table 3.
SVM Confusion Matrix

Actual	Prediction	
	NO	**YES**
NO	0	0
YES	5	9

Information:

1) The actual amount of data that MATCH and predictably MATCH is 0.
2) The amount of data that actually DOES NOT MATCH and is predicted to NOT MATCH is 9
3) The actual amount of data that DOES NOT MATCH and predicted to MATCH is 5.
4) The number of actual data that MATCH and predictably DOES NOT MATCH is 0

4.5 Evaluation

The evaluation was carried out using the AUC, ROC, Accuracy, Precision, Recall values. The evaluation phase for the two algorithms is as follows:

a. Evaluation of C4.5

Then the performance value calculated using Rapidminer is as follows:

$$\text{Accuracy} = \frac{3+6}{3+3+2+6} = 64.29\%$$

$$\text{Error} = \frac{2+3}{3+3+2+6} = 35.71\%$$

Precision = 75,00%

Recall = 66,67%

AUC = 0,867

ROC is presented in the form of an image as shown in Fig 7

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

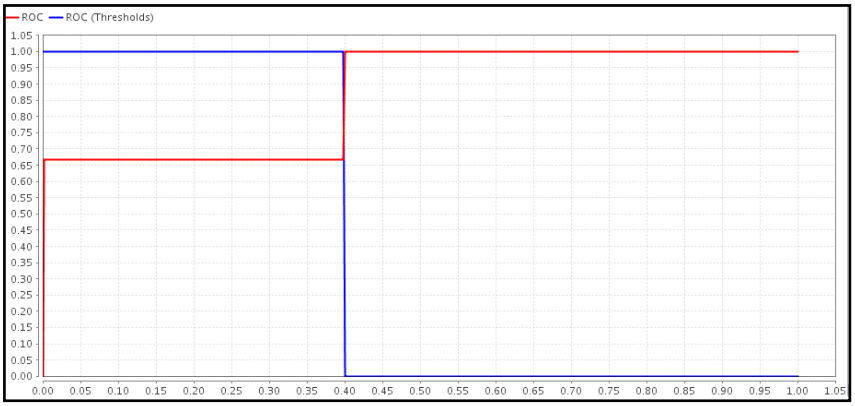

Fig 7. C4.5 ROC Curve

b. Evaluation of SVM

Then the performance value calculated using Rapidminer is as follows:

$$\text{Accuracy} = \frac{0+9}{0+0+5+9} = 64.29\%$$

$$\text{Accuracy} = \frac{5+0}{0+0+5+9} = 35.71\%$$

Precision = 64,29%

Recall = 100%

AUC = 100

ROC is presented in the form of an image as shown in Fig 8

Fig 8. SVM ROC Curve

Apart from the performance testing results, this algorithm has advantages and disadvantages based on the literature study conducted. The advantages and disadvantages of the C4.5 and SVM algorithms are as follows:

a. Advantages of C4.5

The advantage of this algorithm is that it can solve continuous and discrete attributes. In the attribute attribute, this algorithm creates a threshold value and then divides it into a list of values whose attribute can be above, less than or equal to it [4].

b. Disadvantages of C4.5

The C4.5 algorithm builds empty branches which is the most important step in rule creation in C4.5. There are many nodes with zero values or serving zero values. These values do not contribute in generating rules or in constructing any class for classification [4].

c. Advantages of SVM

The advantage of this algorithm is, if C and Parameters are chosen correctly (in the Gaussian kernel), the SVM algorithm is able to generalize very well for new samples [5].

d. Disadvantages of SVM

The drawback of the algorithm is that since the SVM dimension becomes high, in general the transparency of the results is reduced. For example, SVM cannot display a company's score as a parametric function based on financial ratios or other functional forms [5].

Accredited "Rank 4"(Sinta 4), DIKTI, No. 36/E/KPT/2019, December 13th 2019.

4.6 Interpretation

The results of testing the performance values of the C4.5 and SVM algorithms from the AUC value, ROA accuracy, error, precision and recall will then be compared to see which one is better. The following table is presented in tabular form shown in Table 4.

Table 4.
Comparison of Performance Values

Algoritma	AUC	*Accuracy*	*Error*	*Precision*	*Recall*
C4.5	0,867	64,29%	35,71%	75,00%	66,67%
SVM	100	64,29%	35,71%	64,29%	100%

5. Conclusions

The conclusions obtained from this study illustrate that in general, the two algorithms have their respective advantages and disadvantages in Data Mining ,. If seen in Table 4, from the AUC value, SVM is superior with a value of 100 compared to C4.5 with 0.867, then the accuracy and balance error, from the precision value of C4.5 is better, namely with a value of 75.00%, but from the recall value, SVM is better with a value of 100%. So thus it can change in terms of performance that the SVM Data Mining classification algorithm is superior in processing the Playing Golf dataset compared to the C4.5 algorithm.Suggestions for further research are to test other algorithms, both classification, clustering and association algorithms.

6. References

[1] D. Nofriansyah and G. W. Nurcahyo, *Algoritma Data Mining dan Pengujian*. Yogyakarta: CV Budi Utama, 2015.

[2] A. R. L. Francisco, *The Top Ten Algorithms in Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery*, vol. 53, no. 9. 2013.

[3] E. Elisa, "Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Kontruksi PT.Arupadhatu Adisesanti," *J. Online Inform.*, vol. 2, no. 1, p. 36, 2017, doi: 10.15575/join.v2i1.71.

[4] S. Sathyadevan and R. R. Nair, "Comparative analysis of decision tree algorithms: Id3, c4.5 and random forest," *Smart Innov. Syst. Technol.*, vol. 31, no. 7, pp. 549–562, 2015, doi: 10.1007/978-81-322-2205-7_51.

[5] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan, and M. J. Rajabi, "Advantage and drawback of support vector machine functionality," *I4CT 2014 - 1st Int. Conf. Comput. Commun. Control Technol. Proc.*, no. I4ct, pp. 63–65, 2014, doi: 10.1109/I4CT.2014.6914146.

[6] Lukman, "Penerapan Algoritma Support Vector Machine (SVM) Dalam Pemilihan Beasiswa : Studi Kasus SMK Yapimda," *Fakt. Exacta*, vol. 9, no. 1, pp. 49–57, 2016.

[7] A. Perdana and M. T. Furqon, "Penerapan Algoritma Support Vector Machine (SVM) Pada Pengklasifikasian Penyakit Kejiwaan Skizofrenia (Studi Kasus : RSJ . Radjiman Wediodiningrat , Lawang)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. 9, pp. 3162–3167, 2018.

[8] I. Budiman, "28. 29. 1," Iniversitas Diponegoro Semarang, 2012.

[9] D. Aprilia, D. Aji Baskoro, L. Ambarwati, and I. W. S. Wicaksana, "Belajar Data Mining Dengan Rapid Minner," p. 139, 2013, [Online]. Available: https://www.academia.edu/7712860/Belajar_Data_Mining_dengan_RapidMiner.

[10] S. Widaningsih, "Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm," *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: 10.36787/jti.v13i1.78.

[11] Sinambela, Y., Herman, S., Takwim, A., & Widianto, S. (2020). A STUDY OF COMPARING CONCEPTUAL AND PERFORMANCE OF K-MEANS AND FUZZY C MEANS ALGORITHMS (CLUSTERING METHOD OF DATA MINING) OF CONSUMER SEGMENTATION. Jurnal Riset Informatika, 2(2), 49-54. https://doi.org/10.34288/jri.v2i2.116.

[12] Abdullah, Thoip & Qidri, Sulhan & Nuryadi, Wadi &Widianto, Septian Rheno. (2020) Failover Cluster Nodes and ISCSI Storage Area Network on virtualization Windows Server 2016. JOIN (Jurnal Online Informatika) Volume 5 No.1. Juni 2020: 89-96. DOI: 10.15575/join. v5IL.564. p-ISSN: 2528-1682. E-issn: 2527-9165.

[13] Utami, Amalia & Pratama, Bayu & Widianto, Septian. (2020). DATA MART DESIGN IN BKPP BANDUNG USING FROM ENTERPRISE MODELS TO DIMENSIONAL MODELS METHOD. JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer). 5. 279-284. 10.33480/jitk.v5i2.1219.

[14] Aditya, Adhisyanda M & Mulyana, Dicky R & Widianto, Septian Rheno (2020). Penggabungan Teknologi Untuk Analisa Data Berbasis Data Science. Seminar Nasional Teknologi Komputer & Sains (SAINTEKS). Hal. 51-56. ISBN: 978-602-52720-7-3.

[15] Tohirin &Widianto, Septian Rheno. (2020). Peran Trello dalam Adopsi Agile Scrum pada Pengembangan Sistem Informasi Kesehatan. Jurnal Multinetics. Vol 6. No.1. pg.32-39. https://doi.org/10.32722/multinetics.vol6i.2765.

[16] Gunadi, Faustina & Widianto, Septian Rheno. (2020). Efektifitas Pelaporan Pajak Online di Indonesia Berbasis Cobit 5.0 pada Domain MEA (Monitor, Evaluate, Assess). Seminar Nasional Teknologi Komputer & Sains (SAINTEKS). Hal. 82-85. ISBN: 978-602-52720.-7-3.

[17] Gunadi, Faustina &Widianto, Septian Rheno. (2020). Evaluasi Kualitas Pelaporan Manajemen pada Sistem Epicor Perusahaan Manufaktur Berbasis McCall.Jurnal Multinetics. Vol 6. No.1. pg.21-31. https://doi.org/10.32722/multinetics.vol6i.2765.
[18] Widianto, Septian Rheno. (2015). Perancangan Jaringan WLAN di PT. Gemopia Jewellery Indonesia. Jurnal Multinetics. Vol.1, No. 2. https://doi.org/10.32722/multinetics.Vol1.No.2.2015.pp.50-53.
[19] Mahardi, Sandi & Kuncoro, Adi M & Widianto, Septian Rheno. Integrasi Data Sektoral Pemerintah. (2020). Seminar Nasional Teknologi Komputer & Sains (SAINTEKS). Hal. 615-617. ISBN: 978-602-52720.-7-3.
[20] Widianto, Septian Rheno & Azzam, Abdullah Izzudin (2018). Analisis Upaya Peretasan Web Application Firewall dan Notifikasi Serangan Menggunakan Bot Telegram pada Layanan Web Server. Jurnal Elektra. Vol. 3, No.2, Juli 2018. Hal. 19-28. ISSN: 2503-0221.
[21] Utami, Sri Farida (2020). Penerapan Data Mining Algoritma Decision Tree BerbasisPSO. Seminar Nasional Teknologi Komputer & Sains (SAINTEKS). Hal. 677-681.ISBN: 978-602-52720.-7-3.
[22] Widianto, S., S.B.K, F. and Purwanto, A. (2020) "Analysis of Mobile Based Software Development Model: Systematic Review", *Jurnal Mantik*, 4(3, Nov), pp. 1703-1711. doi: 10.35335/mantik.Vol4.2020.973.pp1705-1713.
[23] Widianto, S. and Magdalena, M. (2020) "Online Disposition Data Based Management System", *Jurnal Mantik*, 4(3, Nov), pp. 1641-1648. doi: 10.35335/mantik.Vol4.2020.971.pp1641-1648.
[24] Widianto, S. and Warmayudha, I. P. (2020) "HSQL Database", *Jurnal Mantik*, 4(3, Nov), pp. 1717-1721. doi: 10.35335/mantik.Vol4.2020.982.pp1717-1721.
[25] Widianto, S., Sudiro, S., Suwandi, I. and Leiliawati, L. (2020) "Database Management System on Raw Material Transaction System Case Study : Sabana Fried Chicken", *Jurnal Mantik*, 4(3, Nov), pp. 1722-1727. doi: 10.35335/mantik.Vol4.2020.983.pp1722-1727.