PAPER • OPEN ACCESS

Data Mining Optimization uses C4.5 Classification and Particle Swarm Optimization (PSO) in the location selection of Student Boardinghouses

To cite this article: Ari Waluyo *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **874** 012024

View the article online for updates and enhancements.

This content was downloaded from IP address 180.244.232.47 on 30/12/2020 at 16:04

Data Mining Optimization uses C4.5 Classification and Particle Swarm Optimization (PSO) in the location selection of Student Boardinghouses

Ari Waluyo[1], Hendra Jatnika[2], Marlina Rahmi Shinta Permatasari[3], T Tuslaela[4], Indah Purnamasari[4], Agus Perdana Windarto[5*]

[1]Politeknik Dharma Patria, Indonesia
[2]Institut Teknologi PLN, Indonesia
[3]Universitas Bina Sarana Informatika, Indonesia
[4]STMIK Nusa Mandiri, Indonesia
[5]STIKOM Tunas Bangsa, Indonesia

Email: agus.perdana@amiktunasbangsa.ac.id *

Abstract. The purpose of this study is to select the location of student boarding houses using Particle Swarm Optimization (PSO) and C4.5 optimization techniques. The source of the data was obtained by observing and giving questionnaires to 150 respondents who were lodging in the Pematangsiantar-Simalungun area. from the data set of 81 records and using 5 parameters of assessment ((C1) water cleanliness, (C2) Facilities, (C3) Transportation, (C4) Security, and (C5) Conditions) obtained the results of modeling using the C4.5 + PSO algorithm has better accuracy is 97.78% compared to the C4.5 model whose accuracy is 97.53%. Thus, it is evident that the PSO applied to the weighting of the C4.5 attribute increases the value of accuracy..

1. Introduction

Ideal residence is a place that can protect and become a place of rest for us and a place to grow a healthy life spiritually and physically [1]. There are various types of residences that are permanent or temporary, temporary residences are rented houses and boarding houses. Boarding houses become one of the places to stay for students, students, workers, employees who are far from their homes. Boarding houses are an option for them, especially students, to become temporary residences while carrying out work. The choice of boarding houses is an alternative that must be thought wisely because it considers many aspects such as affordable prices, convenient location, transportation facilities and others. Many boarding houses are an obstacle for students to choose eligibility according to their wants and needs. Boarding houses that are safe, secure, comfortable become one of the dreams of students who are far from where they live. In choosing a boarding house dream there are criteria in choosing it. The number of criteria in the selection of boarding houses is an obstacle for students. So it is necessary to classify the criteria in selecting the location of student boarding houses. The method used for the selection of the location of the boarding houses of students uses Datamining [2]–[4] with C4.5 classification techniques [5], [6] and Particle Swarm Optimization (PSO) [7], [8]. The method often used for classification is C4.5 [5], [6]. This method is also called a very strong and well-known decision tree for classification and prediction [9]. The ability of this C4.5 method can produce decision trees that are easily interpreted [10], have an acceptable level of accuracy [11], are efficient in handling discrete type attributes and can handle discrete and numeric type attributes [7]. However, the

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

C4.5 method has several disadvantages including overlapping often when classes and criteria used are very numerous [8] and overfitting occurs because there is noise training data, which is irrelevant data so that the tree has a long and unbalanced subtree. In this problem can be overcome by optimized using the Particle Swarm Optimization (PSO) algorithm. The PSO algorithm is used to optimize accuracy on C4.5 to get maximum results. Research using PSO optimization for prediction has been done a lot before, namely for the prediction of sea tides where PSO is used to optimize the minimum error value in the network in order to obtain the ideal neural network weights. PSO and artificial neural networks [12]–[16] have several input parameters such as, the number of input neurons, learning rate, swarm, c1, c2 min inertia, max inertia. The data used are 1000 which are divided into 700 training data and 300 testing data. The test results showed that the prediction accuracy was 91.56% using 90 swarm, learning rate 0.9 and iteration 20 times [8]. Subsequent research on the classification of credit analysis using the c4.5 algorithm and PSO [7]. From the results of experiments conducted a C4.5 algorithm model based on Particle Swarm Optimization (PSO) got the best results at 70%, while the C4.5 algorithm model without Particle Swarm Optimization (PSO) was only 68.6%. Based on this, it is expected that the research results can classify the location of student boarding houses.

2. Methodology

2.1. Data Mining

Data mining is in the form of observational analysis of data sets to find unexpected relationships and to summarize data in new ways that can be understood and useful for data owners. There are several settlement techniques used in data mining, including: clustering, classification, estimation and association [2].

2.2 Classification

Data classification is a process that finds the same properties in a set of objects in a database and classifies them into different classes according to the specified classification model [17]. The purpose of classification is to find a model of a training set that distinguishes attributes into appropriate categories or classes, the model is then used to classify attributes whose classes have not been known before. Classification techniques are divided into several techniques including ID3, CART, and C4.5 [9].

2.3 Decision Tree C4.5

C4.5 method is to change the tree generated in several rules. The number of rules is equal to the number of paths that might be built from the root to the leaf node [6]. In general, C4.5 algorithm to build a decision tree with the following general steps:

a) Select the attribute as the root
b) Create a branch for each value
c) Divide cases in branches
d) Repeat the process for each branch until all cases in the branch have the same class [4], [5].

2.4 Research Method

The research data were obtained by conducting direct observations and giving questionnaires to students as many as 150 respondents who were randomly assigned. Data of 150 respondents consisted of student data from five private tertiary institutions in the area of Simalungun Regency and Pematangsiantar City. Respondents are those who live in boarding-houses. The results of observation data and questionnaires were pre-processed using Microsoft Excel software. From the results of the questionnaire obtained several criteria for the selection of the location of student boarding houses, among others: (C1) water cleanliness, (C2) Facilities, (C3) Transportation, (C4) Security, and (C5) Conditions. For water hygiene criteria (C1) have sub criteria {Good, Enough, Bad}, facility criteria (C2) have sub criteria {Luxury, Simple, Standard}, transportation criteria (C3) have sub criteria {Far, Medium, Near}, security criteria (C4) have sub criteria {Strict, Normal, Free} and condition criteria (C5) have sub criteria {Eligible, Not Eligible}.

3. Results and Discussion

The following is an example dataset used as research material (the dataset is randomly generated as many as 81 data samples / training).

Table 1. Learning Dataset

No	Water cleanliness (C1)	Facilities (C2)	Transportation (C3)	Security (C4)	Conditions (C5)
1	Good	Luxury	Far	Strict	Not Eligible
2	Good	Luxury	Far	Normal	Not Eligible
3	Good	Luxury	Far	Free	Not Eligible
4	Good	Simple	Far	Strict	Not Eligible
5	Good	Simple	Far	Normal	Not Eligible
6	Good	Simple	Far	Free	Not Eligible
7	Good	Standard	Far	Strict	Not Eligible
8	Good	Standard	Far	Normal	Not Eligible
9	Good	Standard	Far	Free	Not Eligible
10	Good	Luxury	Medium	Strict	Eligible
11	Good	Luxury	Medium	Normal	Eligible
12	Good	Luxury	Medium	Free	Not Eligible
13	Good	Simple	Medium	Strict	Eligible
14	Good	Simple	Medium	Normal	Eligible
15	Good	Simple	Medium	Free	Not Eligible
16	Good	Standard	Medium	Strict	Eligible
17	Good	Standard	Medium	Normal	Eligible
18	Good	Standard	Medium	Free	Not Eligible
19	Good	Luxury	Near	Strict	Eligible
20	Good	Luxury	Near	Normal	Eligible
21	Good	Luxury	Near	Free	Not Eligible
22	Good	Simple	Near	Strict	Eligible
23	Good	Simple	Near	Normal	Eligible
24	Good	Simple	Near	Free	Not Eligible
25	Good	Standard	Near	Strict	Not Eligible
26	Good	Standard	Near	Normal	Eligible
27	Good	Standard	Near	Free	Not Eligible
28	Enough	Luxury	Far	Strict	Not Eligible
29	Enough	Luxury	Far	Normal	Not Eligible
30	Enough	Luxury	Far	Free	Not Eligible
31	Enough	Simple	Far	Strict	Not Eligible
32	Enough	Simple	Far	Normal	Not Eligible
33	Enough	Simple	Far	Free	Not Eligible
34	Enough	Standard	Far	Strict	Not Eligible
35	Enough	Standard	Far	Normal	Not Eligible
36	Enough	Standard	Far	Free	Not Eligible
37	Enough	Luxury	Medium	Strict	Eligible
38	Enough	Luxury	Medium	Normal	Eligible
39	Enough	Luxury	Medium	Free	Eligible
40	Enough	Simple	Medium	Strict	Eligible
41	Enough	Simple	Medium	Normal	Eligible
42	Enough	Simple	Medium	Free	Not Eligible
43	Enough	Standard	Medium	Strict	Eligible
44	Enough	Standard	Medium	Normal	Eligible
45	Enough	Standard	Medium	Free	Not Eligible
46	Enough	Luxury	Near	Strict	Eligible
47	Enough	Luxury	Near	Normal	Eligible
48	Enough	Luxury	Near	Free	Eligible
49	Enough	Simple	Near	Strict	Eligible
50	Enough	Simple	Near	Normal	Eligible
51	Enough	Simple	Near	Free	Eligible
52	Enough	Standard	Near	Strict	Eligible
53	Enough	Standard	Near	Normal	Eligible
54	Enough	Standard	Near	Free	Eligible
55	Bad	Luxury	Far	Strict	Not Eligible
56	Bad	Luxury	Far	Normal	Not Eligible
57	Bad	Luxury	Far	Free	Not Eligible

58	Bad	Simple	Far	Strict	Not Eligible
59	Bad	Simple	Far	Normal	Not Eligible
60	Bad	Simple	Far	Free	Not Eligible
61	Bad	Standard	Far	Strict	Not Eligible
62	Bad	Standard	Far	Normal	Not Eligible
63	Bad	Standard	Far	Free	Not Eligible
64	Bad	Luxury	Medium	Strict	Eligible
65	Bad	Luxury	Medium	Normal	Eligible
66	Bad	Luxury	Medium	Free	Not Eligible
67	Bad	Simple	Medium	Strict	Eligible
68	Bad	Simple	Medium	Normal	Eligible
69	Bad	Simple	Medium	Free	Not Eligible
70	Bad	Standard	Medium	Strict	Eligible
71	Bad	Standard	Medium	Normal	Eligible
72	Bad	Standard	Medium	Free	Not Eligible
73	Bad	Luxury	Near	Strict	Eligible
74	Bad	Luxury	Near	Normal	Eligible
75	Bad	Luxury	Near	Free	Not Eligible
76	Bad	Simple	Near	Strict	Eligible
77	Bad	Simple	Near	Normal	Eligible
78	Bad	Simple	Near	Free	Not Eligible
79	Bad	Standard	Near	Strict	Eligible
80	Bad	Standard	Near	Normal	Eligible
81	Bad	Standard	Near	Free	Not Eligible

The Dataset Learning Summary data in table 1 will then be processed to obtain a decision tree using the help of RapidMiner 5.3 software.

3.1 The results of the C4.5 method with the RapidMiner software

The Dataset Learning Summary data in table 1 will then be processed to obtain a decision tree using the help of RapidMiner 5.3 software. Following is the C4.5 model using the RapidMiner software as shown in the following image:

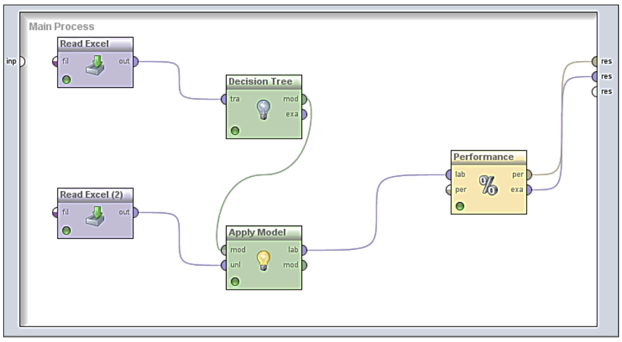

Figure 1. C4.5 model using RapidMiner software

In figure 1 it can be explained that the parameters used in the C4.5 method include: criterion: gain_ratio; minimum size for split: 4; minimum leaf size: 2; minimum gain: 0.1; minimum depth: 20; confidence: 0.25. From this model the accuracy of the results is obtained as shown below:

accuracy: 97.53%

	true Not Eligible	true Eligible	class precision
pred. Not Eligible	41	1	97.62%
pred. Eligible	1	38	97.44%
class recall	97.62%	97.44%	

Figure 2. Model C4.5 accuracy

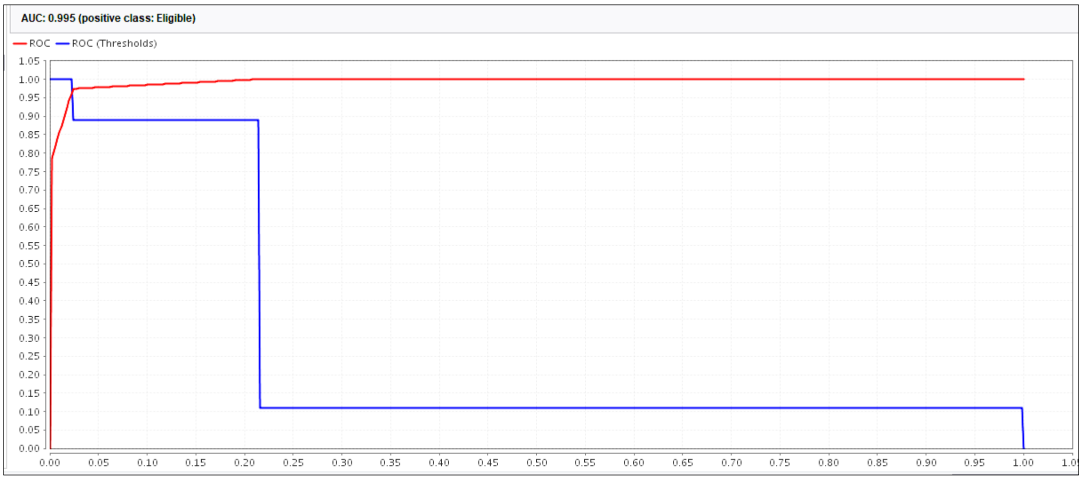

Figure 3. Model C4.5 AUC (Area Under Curve)

```
PerformanceVector
accuracy: 97.53%
ConfusionMatrix:
True:   Not Eligible      Eligible
Not Eligible:    41         1
Eligible:        1          38
precision: 97.44% (positive class: Eligible)
ConfusionMatrix:
True:   Not Eligible      Eligible
Not Eligible:    41         1
Eligible:        1          38
recall: 97.44% (positive class: Eligible)
ConfusionMatrix:
True:   Not Eligible      Eligible
Not Eligible:    41         1
Eligible:        1          38
AUC (optimistic): 0.999 (positive class: Eligible)
AUC: 0.995 (positive class: Eligible)
AUC (pessimistic): 0.990 (positive class: Eligible)
```

3.2 The results of the C4.5 + Particle Swarm Optimization (PSO) method with the RapidMiner software

Following is the C4.5 + Particle Swarm Optimization (PSO) model using the RapidMiner software as shown in the following image:

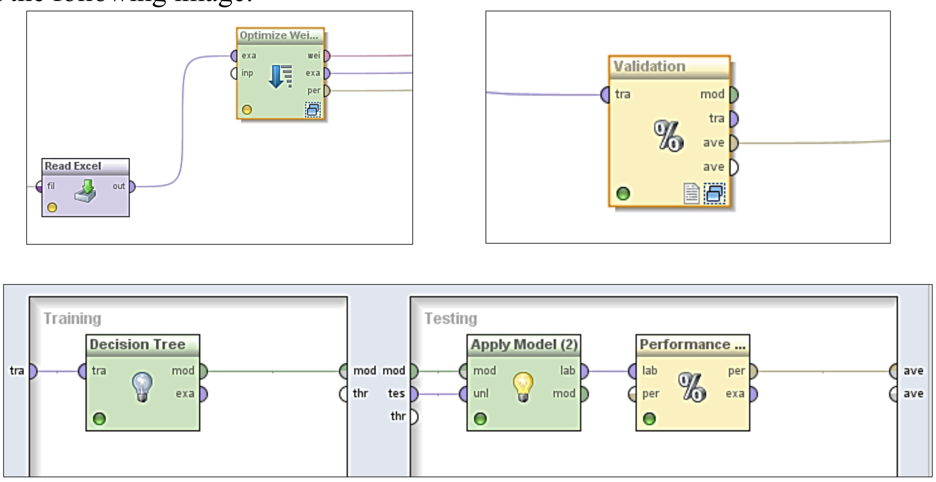

Figure 4. C4.5 + Particle Swarm Optimization (PSO) model using RapidMiner software

Using the same parameters for the C4.5 model (figure 1), optimization with Particle Swarm Optimization (PSO) uses different population sizes and maximum number of generations as shown in the following table:

Table 2. The results of the experiment used population size and maximum number of generations in the PSO + C4.5 method

No	PSO parameter	Accuracy	AUC (Area Under Curve)
1	Posize=10; generate=30	97.64%	0.836
2	Posize=20; generate=30	97.64%	0.962
3	Posize=10; generate=40	97.64%	0.971
4	Posize=15; generate=40	97.64%	0.883
5	Posize=30; generate=50	97.64%	0.825
6	Posize=50; generate=50	97.78%	0.912

Based on table 2, the selection of student boarding locations obtained results that the accuracy value of C4.5 + PSO (97.78%) is better than C4.5 without PSO (97.53%). Up around 0.25%. While the accuracy of classification using AUC, C4.5 method without PSO (0.995) is better than C4.5 + PSO method (0.912). So based on these results, the C4.5 + PSO method can improve the results of prediction accuracy compared to the C4.5 method without PSO. The results of the optimization of C4.5 + PSO states that of the 4 criteria (Water cleanliness (C1), Facilities (C2), Transportation (C3), Security (C4)) used, Water cleanliness (C1) is the most influential alternative with weight = 1.0 and Security (C4) with a weight = 0.5380 as shown in the following image:

accuracy: 97.64% +/- 4.73% (mikro: 97.53%)

	true Not Eligible	true Eligible	class precision
pred. Not Eligible	41	1	97.62%
pred. Eligible	1	38	97.44%
class recall	97.62%	97.44%	

Figure 5. C4.5 + Particle Swarm Optimization (PSO) model of accuracy

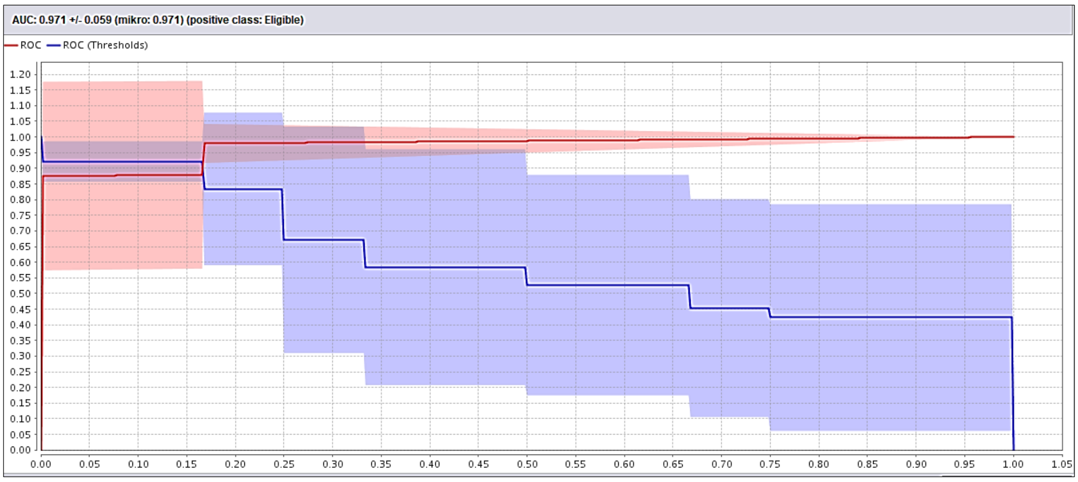

Figure 6. C4.5 + Particle Swarm Optimization (PSO) model of AUC (Area Under Curve)

```
PerformanceVector
accuracy: 97.64% +/- 4.73% (mikro: 97.53%)
ConfusionMatrix:
True:   Not Eligible      Eligible
Not Eligible:    41         1
Eligible:        1          38
precision: 97.50% +/- 7.50% (mikro: 97.44%) (positive class:
Eligible)
ConfusionMatrix:
True:   Not Eligible      Eligible
```

```
Not Eligible:   41       1
Eligible:       1        38
recall: 97.50% +/- 7.50% (mikro: 97.44%) (positive class:
Eligible)
ConfusionMatrix:
True:  Not Eligible      Eligible
Not Eligible:   41       1
Eligible:       1        38
AUC (optimistic): 1.000 +/- 0.000 (mikro: 1.000) (positive
class: Eligible)
AUC: 0.971 +/- 0.059 (mikro: 0.971) (positive class: Eligible)
AUC (pessimistic): 0.958 +/- 0.085 (mikro: 0.958) (positive
class: Eligible
```

attribute	
Water cleanliness (C1)	1
Facilities (C2)	0
Transportation (C3)	0.486
Security (C4)	0.539

Figure 7. The results of the evaluation criteria using C4.5 + PSO Method

4. Conclusion

In this study modeling was carried out using the C4.5 and C4.5 + PSO algorithms with the focus of the research being the application of the PSO algorithm in weighting the attributes of the C4.5 data mining classification technique using the help of RapidMiner 5.3 software. Model validation uses 10 fold cross-validation and model evaluation uses confusion matrix and ROC curves. The results showed that the C4.5 + PSO model had a better accuracy of 97.78% compared to the C4.5 model whose accuracy was 97.53%. Thus, it is evident that the PSO applied to the weighting of the C4.5 attribute increases the value of accuracy.

References

[1] D. L. Fithri, "Model Data Mining Dalam Penentuan Kelayakan Pemilihan Tempat Tinggal Menggunakan Metode Naive Bayes," *J. SIMETRIS*, vol. 7, no. 2, pp. 725–730, 2016.

[2] A. P. Windarto *et al.*, "Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012001.

[3] Sudirman, A. P. Windarto, and A. Wanto, "Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, 2018, doi: 10.1088/1757-899X/420/1/012089.

[4] D. Hartama, A. Perdana Windarto, and A. Wanto, "The Application of Data Mining in Determining Patterns of Interest of High School Graduates," *J. Phys. Conf. Ser.*, vol. 1339, no. 1, 2019, doi: 10.1088/1742-6596/1339/1/012042.

[5] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto, and A. Wanto, "C.45 Classification Rules Model for Determining Students Level of Understanding of the Subject," *J. Phys. Conf. Ser.*, vol. 1255, no. 012005, pp. 1–7, 2019, doi: 10.1088/1742-6596/1255/1/012005.

[6] M. Widyastuti, A. G. Fepdiani Simanjuntak, D. Hartama, A. P. Windarto, and A. Wanto, "Classification Model C.45 on Determining the Quality of Custumer Service in Bank BTN Pematangsiantar Branch," *J. Phys. Conf. Ser.*, vol. 1255, no. 012002, pp. 1–6, 2019, doi: 10.1088/1742-6596/1255/1/012002.

[7] S. Saprudin, "Penerapan Particle Swarm Optimization (PSO) untuk Klasifikasi dan Analisis Kredit dengan Menggunakan Algoritma C4.5," *J. Inform. Univ. Pamulang*, vol. 2, no. 4, p. 214, 2017, doi: 10.32493/informatika.v2i4.1488.

[8] N. Nikentari, H. Kurniawan, N. Ritha, D. Kurniawan, U. Maritim, and R. Ali, "Particle Swarm Optimization Untuk Prediksi Pasang Surut Air Optimization of Backpropagation Artificial Neural Network With Particle Swarm Optimization To Predict Tide Level," *J. Teknol. Inf. dan*

Ilmu Komput., vol. 5, no. 5, pp. 605–612, 2018, doi: 10.25126/jtiik2018551055.

[9] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha, and V. Honrao, “Predicting Students’ Performance Using Id3 and C4.5 Classification Algorithms,” *Int. J. Data Min. Knowl. Manag. Process*, vol. 3, no. 5, pp. 39–52, 2013, doi: 10.5121/ijdkp.2013.3504.

[10] H. Siahaan, H. Mawengkang, S. Efendi, A. Wanto, and A. P. Windarto, “Application of Classification Method C4 . 5 on Selection of Exemplary Teachers,” in *IOP Conference Series*, 2018, pp. 1–6.

[11] I. S. Damanik, A. P. Windarto, A. Wanto, Poningsih, S. R. Andani, and W. Saputra, “Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm,” *J. Phys. Conf. Ser.*, vol. 1255, no. 012012, pp. 1–7, 2019, doi: 10.1088/1742-6596/1255/1/012012.

[12] Sumijan, A. P. Windarto, A. Muhammad, and Budiharjo, “Implementation of Neural Networks in Predicting the Understanding Level of Students Subject,” *Int. J. Softw. Eng. Its Appl.*, vol. 10, no. 10, pp. 189–204, 2016.

[13] Budiharjo, T. Soemartono, A. P. Windarto, and T. Herawan, “Predicting School Participation in Indonesia using Back-Propagation Algorithm Model,” *Int. J. Control Autom.*, vol. 11, no. 11, pp. 57–68, 2018.

[14] Budiharjo, T. Soemartono, A. P. Windarto, and T. Herawan, “Predicting tuition fee payment problem using backpropagation neural network model,” *Int. J. Adv. Sci. Technol.*, vol. 120, pp. 85–96, 2018, doi: 10.14257/ijast.2018.120.07.

[15] A. P. Windarto, M. R. Lubis, and Solikhun, “Implementasi Jst Pada Prediksi Total Laba Rugi Komprehensif Bank Umum Konvensional Dengan Backpropagation,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 4, pp. 411–418, 2018, doi: 10.25126/jtiik.201854767.

[16] A. P. Windarto, M. R. Lubis, and Solikhun, “Model Arsitektur Neural Network Dengan Backpropogation Pada Prediksi Total Laba Rugi Komprehensif Bank Umum Konvensional,” *Kumpul. J. Ilmu Komput.*, vol. 05, no. 02, pp. 147–158, 2018.

[17] L. N. Rani, “Klasifikasi Nasabah Menggunakan Algoritma C4.5 Sebagai Dasar Pemberian Kredit,” *J. KomTekInfo Fak. Ilmu Komput.*, vol. 2, no. 2, pp. 33–38, 2015.