Volume 2, Issue 1(Page 18-29)
ISSN: 2709-0248

Comparative Analysis of the C4.5 Algorithm and the Nearest Neighbor for the Number of Prospective New Student Registrants

Nursetia Wati[1], Irawan Ibrahim[1]

*Correspondence: nursetiawati@umgo.ac.id

[1]Faculty of Engineering, Information System Department, Muhammadiyah University of Gorontalo, Indonesia

Abstract

In 2015, the number of registrants for new student candidates at Muhammadiyah University of Gorontalo, has increased about 20% - 50% from the last year in 2014, but when it starts from 2017/2018 of the academic year the number of new student candidates who registered was only around 4,713 students for bachelor's and there is 1,256 students for Bachelor's Degree, while in the academic year of 2018/2019 bachelor's degree students were only 765 and bachelor's students were around 4,187, it is known as a decline from the previous year. This study, aims to help to predict the number of prospective of the new students who will enroll in the following of the academic year by analyzing the comparison of the C4.5 and Nearest Neighbor Algorithms with comparing two of algorithms to get the best results. In the C4.5 and Nearest Neighbor Algorithms, it is necessary to be able to see some patterns from the data about the prospective students, then, they can produce the predictions of the number of prospective students who can help in increasing the number of prospective students that is according to the target achievements of Muhammadiyah University of Gorontalo (UMG) itself.

Keywords: Prediction, C4.5, Nearest Neighbor

Received : January 12, 2021
Received in Revised: February 8, 2021
Accepted: February 19, 2021

Introduction

Muhammadiyah University of Gorontalo is one of the private universities in Gorontalo, which has been established for more than one decade and, in 2018 with the number of students in the 2015/2016 of the academic year it is experienced a very significant increase that around six hundred (600) students who were accepted. In addition, in every new academic year, the University regularly holds the New Student Admissions activities. This new student admission activity has been routinely carried out, then it will indirectly collect a lot of data from the prospective new students themselves, therefore, in this case what the University needs to do is process of the data that came from all prospective students, then later it will becomes an important information. Based on this information, one of the data that can be generated is the target of the number of new student candidates itself.

The main problem that always continues and becomes an obstacle is, the large of the number of students who register, but, somehow the university cannot predict the number of prospective students who will re-register. This is can be seen in the number of students who entered in the 2017/2018 run into a drastic decline, this is because the private universities are not always become the first choices from prospective students that has been accepted at State Universities.

DOI: https://doi.org/10.47667/ijpasr.v2i1.74

Data Mining is a sequence of several processes to find an added value from a data set in the form of knowledge that is currently unknown manually (Retnosari & Jananto, 2013; Kurgan & Musilek, 2006). Whereas, prediction itself is the process of forecasting future events that based on certain parameters to reduce uncertainty of a condition and create a benchmark to predict future events based on patterns that have occurred in the past (Hartatik, 2015; Mariscal et al, 2010).

In this study, one of the methods that is contained in Data Mining science will be applied, it is a prediction by analyzing the comparison of the C4.5 and Nearest Neighbor algorithms, which is expected by doing this forecasting or prediction model to obtain the best value by looking at the highest level of accuracy of the two algorithms which later can be use by Muhammadiyah University of Gorontalo in predicting the number of new student candidates who re-register and determine policies for the upcoming events of the admission of new students.

Methods

In this method of research, the researcher uses an analysis method that is based on CRISP-DM *(CRoss-Industry Standard Process for Data Mining)* (Kusrini, 2009)

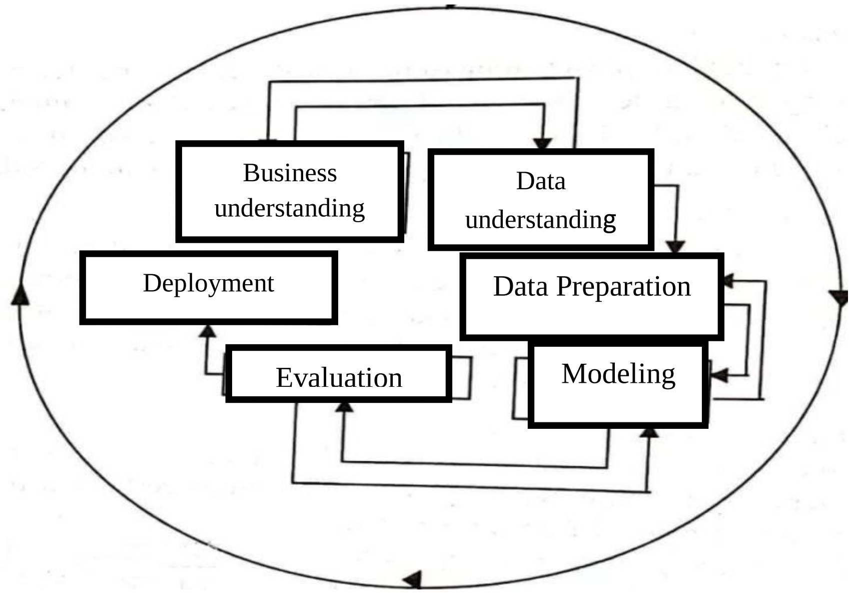


Figure 1. Data Mining Process According to CRISP-DM

The following sentence below is an explanation of the six stages of the *data mining* development life cycle based on the CRISP-DM *(CRoss-Industry Standard Process for Data Mining)*:

The Business Understanding

At this stage the research aims to identify needs in detail, namely by identifying the patterns of the previous student admissions dataset based on the variables that have been selected, as shown in the table below:

DOI: https://doi.org/10.47667/ijpasr.v2i1.74

Gender	Address	Major	……
Female	Luwo'o Villace, sub-district of Telaga Jaya	Language	
Female	Pantungo Village, sub-district of Telaga Biru	Nursing	
Male	Bulota Village, sub-district of Telaga Jaya	Social Sciences (IPS)	

Data Understanding

In collecting the data in this study, the researcher gets the data from prospective new students of Muhammadiyah University of Gorontalo that from 2015-2017 were obtained from the Admission of New Students in Muhammadiyah University of Gorontalo team, as in the table below:

Enthusiasts

Year	1st Sequence	2nd Sequence	3rd Sequence	4rt Sequence	5th Sequence	Sum
2015	415	79	105	130	319	1048
2016	675	269	250	177	15	1386
2017	239	201	771	259	-	1470

Year	Number of the enthusiasts	Number of the registered
2015	1048	1009
2016	1386	1370
2017	1470	1442

Data Processing (Data Preparation Phase)

This part presents, the data that it is includes data processing that is data for New Student Admissions for 2015th to 2017th. This part is includes some several processes, including:

Data Selection

In this process, several attributes were selected to be used for modeling, namely, Gender, School, Department, Wave, Department Choice (2 choices), notes, Religion, Department of graduation, Address and Parents' Occupation.

Data Cleaning

Initial data does not all contain complete or complete data. So it is necessary to do a cleaning process, where the blank data is removed so that the remaining data is ready to be processed.

Data Change

At this part, there are several attributes in the data that also can be simplified into a new attributes. Then the value of the conversion for the required attributes is also carried out.

Modeling

DOI: https://doi.org/10.47667/ijpasr.v2i1.74

This part will be explain about the use of the *Data Mining* technique with the prediction method by using the C45 algorithm and Nearest Neighbor, which produces the prediction rules and the most influential variables in predicting the entry of prospective new students.

Evaluation

In this part, it is about an evaluation that is carried out to obtain the quality and effectiveness of the model used, then the prediction results are obtained for each prediction algorithm. The prediction results are then tested for the level of accuracy with the help of the confusion matrix method. After that, the process of comparing the level of accuracy of each algorithm is carried out to determine which algorithm has the highest accuracy.

Deployment Phase

From the existing results, at this stage there will be a dissemination in the form of making reports and it can be implemented to the Muhammadiyah University of Gorontalo as a reference in predicting the number of incoming new student candidates in the future.

Results and Discussion

The C4.5 Algorithm Calculation Results

The C4.5 algoritma is one the most effective decision algorithms for classification. The dicision tree is built by recursively dividing the data until each part consists of data from the same class (Iskandar & Suprapto, 2016)

Specifically, the C4.5 Decision Tree algorithm uses a modified split criterion called Gain Ration in the split attribute selection process (Jovanovic et al., 2012; Dongming et al., 2016; Mishra et al., 2016; Wang et al., 2019)**.** In this algorithm, it is an algorithm that is used to form a *decision tree* which is, that consists of a set of rules to divide into a number of populations into a smaller ones.

Attribute		**The Number**	**Pass**	**Year of pass**	**Entropy**
Total		**4431**	**3748**	**683**	**0.62011**
Gender	Male	1228	058	70	0.58013
	Female	3203	2690	13	0.63470
The Choices of Department	Agribusiness (AGB)	86	1	5	0.66771
	Accountancy	87	67	0	0.77781
	Primary of Education Teacher (PGSD)	477	438	9	0.40835
	Information System (SI)	179	155	4	0.56852
	….	**….**	**….**	**….**	**….**
	English Literature (ELITE)	93	4	9	0.73045
Last Education	Senior High School (MA)	305	259	6	0.61189
	Vocational High School (SMK)	729	07	2	0.65161
	Senior High School	3397	882	5	0.61384
Major of School	Office administration	94	9	5	0.83560
	Religion	4	2	2	1
	Agribusiness Processing of	11	8	3	0.84535

DOI: https://doi.org/10.47667/ijpasr.v2i1.74

	Agricultural Products				
	…	…	…	…	…
	Agricultural Product Processing Technology	21	4	7	0.70247
	Travel agent	2	2	0	0
Total of Exam	Not Following the exam	77	0	0	0
	<70	631	96	19	0.94316
	<80	1453	1407	0	0
	<90	1679	657	2	0.10073

The total row of the Entropy column is calculated by the following equation below:

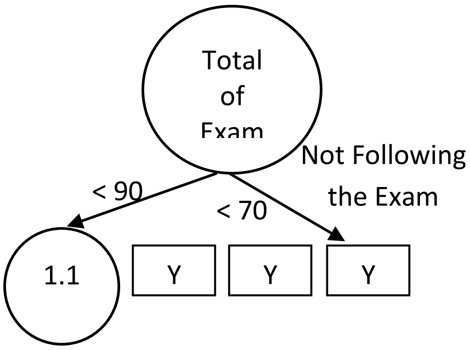
$$Entropy\ (total) = \left(-\frac{3748}{4431}\ \boldsymbol{xlog_2}\left(\frac{3748}{4431}\right)\right) + \left(\frac{683}{4431}\ \boldsymbol{xlog_2}\left(\frac{683}{4431}\right)\right) = 0.62011$$

Meanwhile, the gain value for gender is calculated by the following equation below:

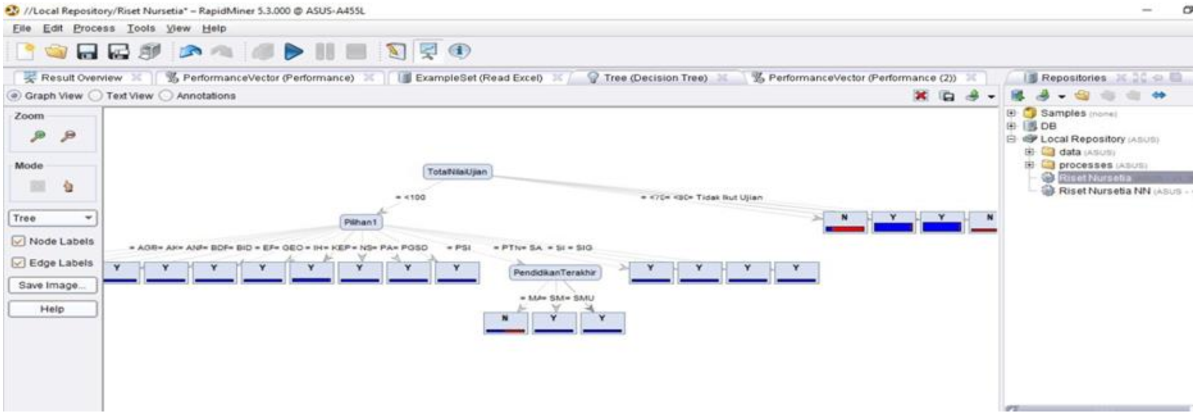
$$Gain\ (total, gender) = Entropy\ (total) - \sum_{i=1}^{n} \frac{gender}{total}\ x\ Entropy\ (gender)$$

$$Gain\ (total, gender) = 0.62011 - \left(\left(\frac{1228}{4431} x0.58013\right) + \left(\frac{3203}{4431} x0.63470\right)\right) = 0.00053$$

From the calculation results contained in the node calculation table, it can be seen that the highest gain is found in the total test attribute, which is 0.44763. Thus the test total can be the root node. And in the total test there are 4 attribute values from which the results can be described as a temporary decision tree as follows:



The results also look like the same, when we test using the Rapid-miner tools which are shown in the following image below:



Nearest Neighbor Algorithm Calculation Results

This algorithm uses an approach to searching for cases by calculating the closeness between a new cases and old cases which is, it is based on matching the weights from a number of an existing features.

Table of the Old Case

Gender	Last Education	Religion	Pass Examination
Male	Vocational High School (SMK)	Hindu	Y
Female	Senior High School	Islam	N
Female	Senior High School	Islam	Y

The attribute of passing the test is an attribute of the goal. The weights between one attribute and another on non-goal attributes can be defined with different values.

Attribute Table of Weight Definition

Attribute	Weight
Gender	0.5
Last Education	1
Religion	0.75

Proximity Table of Gender Attribute Value

Score 1	Score 2	Proximity
Female	Female	1
Male	Male	1
Female	Male	0.5
Male	Female	0.5

Table of Last Education Attribute Value Proximity

Value 1	Value 2	Proximity
Vocational high School (SMK)	Vocational high School (SMK)	1
Senior High School (SMU)	Senior High School (SMU)	1
Senior High School (SMU)	Vocational high School	0.4

DOI: https://doi.org/10.47667/ijpasr.v2i1.74

	(SMK)	
Vocational high School (SMK)	Senior High School (SMU)	0.4

Proximity Table of Religious Attribute Values

Score 1	Score 2	Proximity
Hindu	Hindu	1
Islam	Islam	1
Hindu	Islam	0.75
Islam	Hindu	0.75

New Case

Gender : Male

Last Education : Vocational High School (SMK)

Religion : Hindu

To predict whether the prospective new student will pass the exam or not, use the following steps:

Calculate the closeness of the new case to the old case. Known as: (a) Proximity of gender attribute values (Male to Male) 1 (b) Weight Attribute Gender 0.5 (c) The closeness of the last educational attribute values (SMK and SMU) 0.4 (d) Last educational attribute weights (e) Proximity of religious attribute values (Hinduism and Hinduism) 1 (f) The Weight of Religious Attributes 0.75

Counted:

$$Range = \frac{(a*b)+(c*d)+(e*f)}{b+d+f}$$

$$Range = \frac{(1*0.5)+(0.4*1)+(1*0.75)}{0.5+1+0.75}$$

$$Range = \frac{1.65}{2.25}$$

$$Range = 0.73$$

Calculating the closeness of the new case with the case of number 2. Known as: (a) Proximity of gender attribute values (Female to Male) 0. (b) Weight Attribute Gender 0.5 (c) The closeness of the last educational attribute values (SMK and SMU) 0.4 (d) Last educational attribute weights 1 (e) The proximity of the value of religious attributes (Islam and Islam) 1 (f) The Weight of Religious Attributes 0.75

Counted:

$$Range = \frac{(a*b)+(c*d)+(e*f)}{b+d+f}$$

$$Range = \frac{(0.5*0.5)+(0.4*1)+(1*0.75)}{0.5+1+0.75}$$

$$Range = \frac{1.4}{2.25}$$

$$Range = 0.62$$

Calculating the closeness of the new case to case number 3

DOI: https://doi.org/10.47667/ijpasr.v2i1.74

Known: (a) Proximity of gender attribute values (Female to Male) 0.5 (b) Weight Attribute Gender 0.5 (c) The closeness of the last educational attribute values (SMK and SMU) 0.4 (d) Last educational attribute weights 1 (e) The closeness of the religious attribute values (Hinduism and Islam) 0.75 (f) The Weight of Religious Attributes 0.75

Counted:

$$Range = \frac{(a*b)+(c*d)+(e*f)}{b+d+f}$$

$$Range = \frac{(0.5*0.5)+(0.4*1)+(0.75*0.75)}{0.5+1+0.75}$$

$$Range = \frac{2.15}{2.25}$$

$$Range = 0.95$$

Selecting the case with the closest proximity.

From steps 1st, 2nd and 3rd it can be seen that the highest value is in the case number 3, therefore the closest case to new case is case 3.

Uses the classification of cases with the closest proximity.

Based on the results of step 4, the case of number 3 will be used to predict new cases with the possibility that new students will **pass the exam**. The following explanation below are the results of the Nearest Neighbor algorithm using real data using Excel:

Proximity of the Gender Attribute Values

Value 1	Value 2	Proximity
Male	Male	1
Female	Female	1
Female	Female	0.5
Female	Male	0.5

Proximity of Religious Attribute Values

Nilai 1	Nilai 2	Bobot
Hindu	Kristen	0.75
Islam	Hindu	0.75
Kristen	Islam	0.75
Hindu	Hindu	1
Islam	Islam	1
Kristen	Kristen	1

Proximity of the Educational Attribute Values

Value 1	Value 2	Weight
Senior High School (MA)	Vocational high School	0.4

DOI: https://doi.org/10.47667/ijpasr.v2i1.74

	(SMK)	
Senior High School	Senior High School	0.4
Vocational high School (SMK)	Senior High School	0.4
Senior High School (MA)	Senior High School (MA)	1
Senior High School	Senior High School	1
Vocational high School (SMK)	Vocational high School (SMK)	1

Gender	**Religion**	**Last Education**
Female	I	Senior High School (SMU)

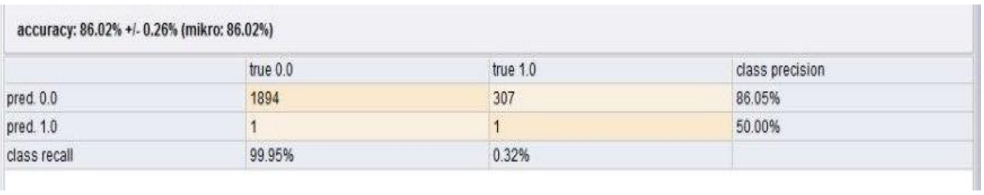
Table of Case

Gender	**Religion**	**Last Education**
Female	I	Senior High School (SMU)
Female	I	Senior High School (SMU)
Female ….	I ….	Senior High School (SMU) ….
Male	I	Senior High School (MA)
Female	I	Senior High School (MA)
Male	Christ	Senior High School (SMU)
Female	I	Vocational High School (SMK)
Male	I	Senior High School (SMU)
Male	I	Senior High School (SMU)
Male	I	Senior High School (MA)
Female	I	Senior High School (SMU)
Female	I	Senior High School (SMU)
Male	I	Senior High School (SMU)
Male	Christ	Senior High School (SMU)
Male	I	Senior High School (SMU)
Female	I	Senior High School (MA)
Male	I	Senior High School (SMU)
Male	I	Senior High School (SMU)

DOI: https://doi.org/10.47667/ijpasr.v2i1.74

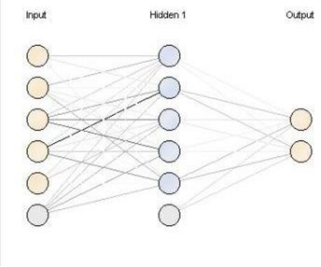
Number	Proximity of Gender Attribute Value	Gender Attribute Weights	Proximity of Religious Attribute Values	Weight of Religious Attributes	Proximity of Educational Attribute Values	Education Attribute Weights	The Formula of Range
1	1	0.5	1	0.75	1	1	1
2	1	0.5	1	0.75	1	1	1
3	1	0.5	1	0.75	1	1	1
4	1	0.5	1	0.75	1	1	1
5	0.5	0.5	1	0.75	1	1	0.888888889
6	1	0.5	1	0.75	0.4	1	0.733333333
…	…	…	…	…	…	…	…
333	0.5	0.5	1	0.75	0.4	1	0.433333333
334	1	0.5	1	0.75	1	1	1
335	0.5	0.5	0.75	0.75	0.4	1	0.433333333
336	1	0.5	1	0.75	0.4	1	0.6
337	0.5	0.5	0.75	0.75	0.4	1	0.433333333
338	0.5	0.5	1	0.75	0.4	1	0.433333333
339	0.5	0.5	1	0.75	0.4	1	0.433333333
340	1	0.5	1	0.75	0.4	1	0.6
341	0.5	0.5	0.75	0.75	1	1	0.833333333
342	0.5	0.5	1	0.75	0.4	1	0.433333333
343	0.5	0.5	1	0.75	1	1	0.833333333
344	0.5	0.5	1	0.75	1	1	0.833333333
345	1	0.5	1	0.75	0.4	1	0.6
346	1	0.5	1	0.75	1	1	1
347	1	0.5	1	0.75	1	1	1
348	0.5	0.5	1	0.75	0.4	1	0.433333333
349	0.5	0.5	0.75	0.75	1	1	0.833333333
350	0.5	0.5	1	0.75	1	1	0.833333333
351	1	0.5		0.75	0.4	1	0.6

The Following table below is the results of using the Rapid-Miner Tool:



accuracy: 86.02% +/- 0.26% (mikro: 86.02%)

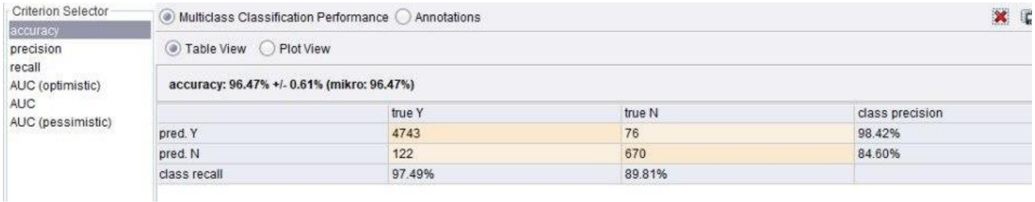
	true 0.0	true 1.0	class precision
pred. 0.0	1894	307	86.05%
pred. 1.0	1	1	50.00%
class recall	99.95%	0.32%	



DOI: https://doi.org/10.47667/ijpasr.v2i1.74

Algorithm C4.5

The conclusion of the *C4.5 algorithm* will be show up in the picture below; the accuracy that has been obtained is 96.47% with the Yes Prediction (Y) as *true yes 4743* and *true no 76* and also, in 98.42% as Precision class and Prediction No (N) yes itself as 122 and True no is 670 with the Class Precision 84.60%.



Criterion Selector
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)

● Multiclass Classification Performance ○ Annotations

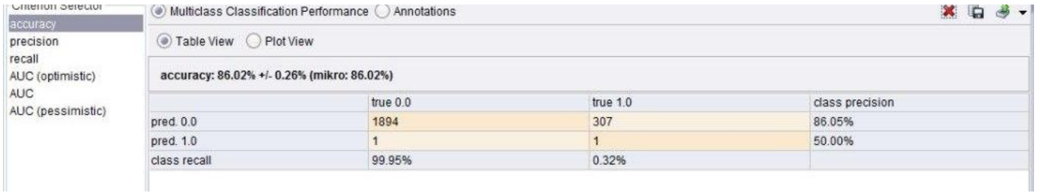
● Table View ○ Plot View

accuracy: 96.47% +/- 0.61% (mikro: 96.47%)

	true Y	true N	class precision
pred. Y	4743	76	98.42%
pred. N	122	670	84.60%
class recall	97.49%	89.81%	

Nearest Neighbor Algorithm

In the picture below, the accuracy of value of the *Nearest Neighbor Algorithm* is 86.02% with the result of precision class is 86.05% and a precision class is 50.00%.



Criterion Selector
accuracy
precision
recall
AUC (optimistic)
AUC
AUC (pessimistic)

● Multiclass Classification Performance ○ Annotations

● Table View ○ Plot View

accuracy: 86.02% +/- 0.26% (mikro: 86.02%)

	true 0.0	true 1.0	class precision
pred. 0.0	1894	307	86.05%
pred. 1.0	1	1	50.00%
class recall	99.95%	0.32%	

Based from both images above, it is very clear that the highest accuracy is in the *C4.5 Algorithm* itself. Because, according to the researchers' analysis that has been explained on the finding and discussion on this research, the *C4.5 algorithm* is in processing of the calculations, through the *Rapid Miner* tools in particular, it does not require to the process of changing the original data itself. Meanwhile, from data that is containing letters and numbers, while for the Nearest Neighbor algorithm, the process of changing the data is very necessary because, when processing the algorithm then it must be numerical and for other reasons, that is the Nearest Neighbor algorithm is more widely used for the classification process according to the proximity of values.

Conclusion

It is clear when the comparison between both of the algorithms itself is superior to the C4.5 algorithm, which in the decision tree is the top node in the Total test of attribute, meanwhile, it is seen in real data when researchers and the researcher’s assistant is perform the data collection in a very differently way, with what has been expected, it is the amount of blank data that is redundant their data when new students fill in their data, especially on the attributes of the origin of the department during school and their Pure Eptanas Score (NEM) or the final scores, which are important attributes in determining the best results.

References

Dongming, L., Yan, L., Chao, Y., Chaoran, L., Huan, L., & Lijuan, Z. (2016, October). The Dongming, L., Yan, L., Chao, Y., Chaoran, L., Huan, L., & Lijuan, Z. (2016, October). The application of decision tree C4. 5 algorithm to soil quality grade forecasting model. In *2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)* (pp. 552-555). IEEE.

DOI: https://doi.org/10.47667/ijpasr.v2i1.74

Hartatik, H. (2015). Penerapan Algoritma Learning Vector Quantization Untuk Prediksi Nilai Akademis Menggunakan Instrumen Ams (Academic Motivation Scale). *Data Manajemen dan Teknologi Informasi (DASI)*, *16*(3), 53.

Iskandar, D., & Suprapto, Y. K. (2016). Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan Antara Algoritma C 4.5 Dan Naïve Bayes. *Network Engineering Research Operation*, *2*(1).

Jovanovic, M., Vukicevic, M., Milovanovic, M., & Minovic, M. (2012). Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study. *International Journal of Computational Intelligence Systems*, *5*(3), 597-610.

Kurgan, L. A., & Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *Knowledge Engineering Review*, *21*(1), 1-24.

Kusrini, E. T. L. (2009). Algoritma data mining. *Yogyakarta: Andi Offset*.

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2), 137.

Mishra, T., Kumar, D., & Gupta, S. (2016). Students' employability prediction model through data mining. *International Journal of Applied Engineering Research*, *11*(4), 2275-2282.

Retnosari, P., & Jananto, A. (2013). Implementasi Data Mining Untuk Menemukan Hubungan Antara Kota Kelahiran Mahasiswa Dengan Tingkat Kelulusan Mahasiswa Pada Fakultas Teknologi Informasi Unisbank. *Journal of Dinamika Informatika, 5*(2), 112–121.

Wang, X., Zhou, C., & Xu, X. (2019). Application of C4. 5 decision tree for scholarship evaluations. *Procedia Computer Science*, *151*, 179-184.

DOI: https://doi.org/10.47667/ijpasr.v2i1.74