Proceeding of 2nd International Conference on Education, Society and Humanity
e-ISSN : 2986-5832
p-ISSN : 2986-6979
Vol. 02 No. 01 (2024)
Available online at https://ejournal.unuja.ac.id/index.php/icesh

ANALYSIS AND IDENTIFICATION OF NEW STUDENT ACCEPTANCE SYSTEM WITH EQUAL WIDTH INTERVAL DISCRETIZATION TECHNIQUE IN K-MEANS CLUSTERING METHOD (CASE STUDY: SMA NEGERI 9 MEDAN)

Sofyan Rahmad[1], Muhammad Zarlis[2], Zakarias Situmorang[3]
[1,2,3]Potensi Utama University, North Sumatra, Indonesia
Email: sofyanrahmad99@gmail.com[1], m.zarlis@yahoo.com[2], zakarias65@yahoo.com[3]

Abstract:

The admission of new students is the first gate that students and schools must pass through in screening educational objects. It is an important event for a school because it is the starting point that determines the smooth running of a school's work. The new student admission system (PPDB) has been implemented by many schools. One of these schools is SMA Negeri 9 Medan. The Equal-width interval Discretisation technique is the simplest discretization method that divides the range of observed values on each feature/attribute, variable k is a parameter provided by the user. This study will calculate what criteria are used to select new students at the school. The criteria used such as the distance of the new student's residence area to the school, academic achievement, and others are then calculated and as a support, Rapidminer 10.1 software is used. The results of data testing and cluster data will be processed to be considered as recommendations for schools and new students at the school.

Keywords: *Equal Width Interval Description Technique, New Student Reception, K-Means Clustering Method.*

INTRODUCTION

The admission of new students is the first gate that students and schools must pass through in screening educational objects. An important event for a school, because this event is the starting point that determines the smooth running of a school's tasks. Errors in admitting new students can determine the success or failure of the educational efforts at the school concerned (Andini et al., 2016). The new student admission system (PPDB) has been implemented by many schools. One of these schools is SMA Negeri 9 Medan. The school follows the minister's regulations. SMA Negeri 9 Medan is a high school in the city of Medan, North Sumatra, where the school applies the system so that the community around the school can pursue education close to the school. The system has been used since 2021, precisely at the beginning of the new school year. The system is considered very efficient for the surrounding community. For this reason, the author wants to help the school in implementing the system so that it will be good by applying the Equal-width interval Discretization Technique and classifying data using K-Means Clustering (Ayuni & Fitrianah, 2019).

K-means clustering is a data analysis method or Data Mining method that performs an unsupervised modeling process and is one of the methods that classify data with a partition system (Lubis et al., 2021). Two types of data

clustering are often used in the process K-Means algorithm is a clustering algorithm that groups data based on the nearest cluster centroid to the data (Hendrickx et al., 2015). The purpose of K-Means is to group data by maximizing the similarity of data within a cluster and minimizing the similarity of data between clusters. The similarity measure used in clustering is a distance function. So that the maximization of data similarity is obtained based on the shortest distance between the data and the centroid point (Kamal & Ilyas, 2017). The initial stage carried out in the data clustering process using the K-Means algorithm is the formation of the initial centroid point cj in general, the formation of the initial centroid point is generated randomly. The number of centroids cj generated corresponds to the number of clusters determined at the beginning (Lumbantoruan, 2017). After k centroids are formed, the distance of each data xi is calculated with jth to kth centroid, denoted by d (xi,cj) (Nofriansyah et al., 2016). There are several distance measures used as a measure of the similarity of a data instance, one of which is the Euclid distance of data clustering, namely Hierarchical, and Non-Hierarchical, and K-Means is one of the non-hierarchical or Partitional Clustering data clustering methods (Nugraha et al., 2016). The K-Means Clustering method tries to group existing data into several groups, where the data in one group has the same characteristics as each other and has different characteristics from the data in other groups (Sulardi et al., 2017). Equal-width interval Discretisation technique is the simplest discretization method that divides the range of observed values on each feature/attribute, the variable k is a parameter provided by the user (Zunaidhi et al., 2012). The process involves sorting the observed values of continuous features/attributes and finding the minimum (V-min) and maximum (V-max) values (Gustientiedina et al., 2019). The interval can be calculated by dividing the range of observed values for the variable into K equal-sized ones using the following formula, variable k is a user-provided parameter (Muin, 2016).

RESEARCH METHODS

The data used in the study were taken from SMA Negeri 9 Medan for the 2021-2022 Entry Year, which totaled 115 new students. The data is processed through several stages starting from calculations using the equal width interval discretization technique and then grouped using the k-means clustering method. After processing the data, testing, and data accuracy are carried out using the Rapidminer 10.1 application compared with the original data and then measured for accuracy. Here are some things that are done.

A. Data Collection

The data collection process is carried out based on the criteria used by the school in admitting new students for T.A. 2021-2022. Then data processing or cleaning is carried out by determining the attributes used in data processing.

B. Data Testing

Data testing is done by entering initial data in the Rapidminer 10.1 application to get 90-100% accuracy results with existing data.

C. Calculation of Equal Width Interval Discretisation

Before testing, the existing data is calculated with the equal width interval discretization technique with the results of data similarity accuracy with the initial data is 90 - 100%. Data calculation is done by determining the interval from cluster 0 to cluster 1 onwards. The data calculation is iterated 3 times until it is felt that the data gets good results. The formula used in the equal-width interval discretization technique is as follows:

$$\text{Interval} = \frac{Vmax - Vmin}{k}$$

Where:
Vmax = maximum value
Vmin = minimum value
K = Interval Width

D. Data Classification Model

The data classification model is carried out by selecting the main attributes for calculation. Based on the results of data classification attributes used such as name, report card score, and distance from the student's home to school. Attribute values will be calculated and produce data on the chances of new students entering using zoning, affirmation, achievement, or moving parents. The data will be clustered with the k-means clustering method.

E. Testing Method Using Rapidminer 10.1 Application

After manual calculation, data testing will be carried out using the help of the Rapidminer 10.1 application with the initial data and then the results using the equal width interval discretization technique.

1. Test data in the form of Excel files that will be imported into the Rapidminer 10.1 application.
2. The clustering process is done by selecting the k-means clustering operator on the Rapidminer 10.1 menu. After setting the number of clusters to the variables used, the data is run.
3. The results will show the number of each cluster which will be compared with the calculation results using the equal width interval discretization technique.

F. Evaluation and Validation

This research will get 2 different results, namely the results using the calculation with the equal width interval discretization technique and the results without the equal width interval discretization technique through the Rapidminer 10.1 tool. The data obtained will be evaluated and calculated for accuracy or data similarity with the initial data. After the results obtained have more than 90% data similarity, the data is validated with accurate results.

RESULTS AND DISCUSSION

This research data was obtained from SMA Negeri 9 Medan school in the student affairs and public relations department about the selection of new student admissions based on criteria from the Ministry of Education and Culture. The new student admission pathways owned by the school are zoning pathways (where students live), achievement pathway, and affirmation pathways (low economy), and transfer pathways. Students who will register first fill out the form and then the school will record and enter it into the new student admission data form. The new student admission pathways are a reference for the school in selecting new students. The data that has been obtained will be managed manually and the results announced. This study uses data for the 2021-2022 school year with data for 115 new students. The data will be taken through the new student admission form from the ministry that has been managed by the school and data transformation will be carried out with supporting attributes such as NISN, student name, gender, date of birth, address, parent's name, junior high school diploma number, KIP recipient, NIK, Information filled in with Zoning or Affirmation options or Achievement or Moving parents' workplace using the K-Mens Clustering method with the

Equal-width interval Discretization technique by researchers. Data on new student admissions at SMA Negeri 9 Medan:

Table 1. New student enrolment school data

No.	No. Registration	NISN	Participant Name	Address	Average Report Card	Junior High Accreditation	Distance (Km)	Entry Path				Number of Opportunity Pathways
								Zoning	Affirmation	Achievement	Moving Parents	
1	426300	71424224	NANDA PAULINA SITOMPUL	Jl, Sei Mati Lk 2 Medan Labuhan	87,9	A	13	1	0	1	1	3
2	42481	8253816	SHAFA MAHARANI IMANTI POHAN	Jl, Sei Mati Lk 2 Medan Labuhan	88,7	A	11	1	0	1	1	3
3	504552	76619185	M, FATHAHILLAH AJI CAKRANINGRAT	Jl, Sei Mati Lk 2 Medan Labuhan	71,9	A	14	1	0	0	1	2
4	456832	6892216	WINDA SYAHRANI	Jl, Sei Mati Lk 2 Medan Labuhan	89,74	A	16	0	1	1	1	3
5	330074	73520246	NAILA FADILA	Jl, Sei Mati Lk 2 Medan Labuhan	88,74	A	18	0	1	1	1	3
6	396401	74009412	SALMA ULIMA BAHY	Jl, Sei Mati Lk 10 Medan Labuhan	88,03	A	19	0	1	1	1	3
7	509420	71833886	AMANDA CHAIRUNNISA	Jl, Sei Mati Lk 10 Medan Labuhan	86,63	A	19	0	1	1	1	3
8	394392	3071196438	MANTASYA MAGFIRAH	Jl, Sei Mati Lk 10 Medan Labuhan	86,31	A	22	0	1	1	1	3
9	41185	76946405	GABRIEL PANDAPOTAN SIMANGUNSONG	Jl, Sei Mati Lk 10 Medan Labuhan	86,03	A	30	0	1	1	1	3
10	495243	66815296	JIHAN SALSABILLA	Jl, Sei Mati Lk 10 Medan Labuhan	85,86	A	39	0	1	1	1	3
11	366057	75345149	YULIA CITRA	Jl, Sei Mati Lk 10 Medan Labuhan	85,8	A	5	1	0	1	1	3
12	484220	76989442	DWI NABILAH	Jl, Sei Mati Lk 10 Medan Labuhan	89,11	B	13	0	1	1	0	2
13	451217	71389578	NOVA APRILIA	Jl, Sei Mati Lk 10 Medan Labuhan	84,66	A	2	1	0	0	1	2
14	414421	78771498	ARLINA INDRI MEYLISA BR,SIMANJUNTAK	Jl, Sei Mati Lk 10 Medan Labuhan	84,51	A	9	1	0	0	1	2
15	545408	72429478	ARNETTA MEI SUCI LESTARI	Jl, Sei Mati Lk 10 Medan Labuhan	84,43	A	14	1	0	0	1	2
16	394486	73634684	AULIA SHAWQI	Jl, Sei Mati Lk 10 Medan Labuhan	84,34	A	5	1	0	0	1	2
17	42846	3078879485	MUHAMMAD GENTA SUDIRJA	Jl, Sei Mati Lk 10 Medan Labuhan	84,14	A	16	1	0	0	1	2
18	540646	71945547	FANI FITRI ANA	Jl, Sei Mati Lk 10 Medan Labuhan	83,86	A	21	1	0	0	1	2
19	418350	72163519	CHELSEA NABABAN	Jl, Sei Mati Lk 10 Medan Labuhan	88,09	B	33	0	1	1	0	2
20	595013	74079176	CHRISTINE RIANA SIMANJUNTAK	Jl, Sei Mati Lk 10 Medan Labuhan	83,74	A	37	0	1	0	1	2
21	356736	72103740	CHARLY YOSEP ARDANA SILALAHI	Jl, Sei Mati Lk 10 Medan Labuhan	83,46	A	40	0	1	0	1	2
22	382141	76352927	YEHEZKIEL DEO SUTANSHA HUTAGALUNG	Jl, Sei Mati Lk 10 Medan Labuhan	83,37	A	50	0	1	0	1	2
23	48287	73897136	NAILAH MAHDIAH PUTRI	Jl, Sei Mati Lk 10 Medan Labuhan	87,66	B	54	0	1	1	0	2
24	587208	76271540	PASKAH HAMONANGAN SIANTURI	Jl, Sei Mati Lk 10 Medan Labuhan	83,11	B	55	0	1	0	0	1
25	418587	79439455	NADA SORAYA	Jl, Yos sudarso medan labuhan lk 21	86,94	B	61	0	1	1	0	2
26	547259	7682565	NAYLA SYAHIDA	Jl, Yos sudarso medan labuhan lk 21	82,54	B	64	0	1	0	0	1
27	350850	71635799	TOGA ANDREAS MANUELA	Jl, Yos sudarso medan labuhan lk 21	82,23	B	64	0	1	0	0	1
28	331366	77141204	RANGGA JULION PRADITIA MANALU	Jl, Yos sudarso medan labuhan lk 21	82,23	B	44	0	1	0	0	1
29	504123	78865286	SYIFA NAMIRA NASUTION	Jl, Yos sudarso medan labuhan	81,69	A	47	0	1	0	1	2
30	421278	78740446	APRILIA STEVANI SIBURIAN	Jl, Yos sudarso medan labuhan	81,63	A	56	0	1	0	1	2

Data cleaning is the process of selecting attributes that are important to use in the study. Existing data attributes such as registration number, NISN, and others are selected using only 4 attributes, namely the participant's name, average report card score, accreditation of the school of origin, and distance from home to school, the following are the results of cleaning the data attributes of new student admissions to be managed.

Table 2. Data Cleaning

NO	Participant Name	Average Report Card	Junior High Accreditation Origin	Distance (Km)
1	NANDA PAULINA SITOMPUL	87,9	A	13
2	SHAFA MAHARANI IMANTI POHAN	88,7	A	11
3	M, FATHAHILLAH AJI CAKRANINGRAT	71,9	A	14
4	WINDA SYAHRANI	89,74	A	16
5	NAILA FADILA	88,74	A	18
6	SALMA ULIMA BAHY	88,03	A	19
7	AMANDA CHAIRUNNISA	86,63	A	19
8	MANTASYA MAGFIRAH	86,31	A	22
9	GABRIEL PANDAPOTAN SIMANGUNSONG	86,03	A	30
10	JIHAN SALSABILLA	85,86	A	39
11	YULIA CITRA	85,8	A	5
12	DWI NABILAH	89,11	B	13
13	NOVA APRILIA	84,66	A	2
14	ARLINA INDRI MEYLISA BR,SIMANJUNTAK	84,51	A	9
15	ARNETTA MEI SUCI LESTARI	84,43	A	14
16	AULIA SHAWQI	84,34	A	5
17	MUHAMMAD GENTA SUDIRJA	84,14	A	16
18	FANI FITRI ANA	83,86	A	21
19	CHELSEA NABABAN	88,09	B	33
20	CHRISTINE RIANA SIMANJUNTAK	83,74	A	37
21	CHARLY YOSEP ARDANA SILALAHI	83,46	A	40
22	YEHEZKIEL DEO SUTANSHA HUTAGALUNG	83,37	A	50
23	NAILAH MAHDIAH PUTRI	87,66	B	54
24	PASKAH HAMONANGAN SIANTURI	83,11	B	55
25	NADA SORAYA	86,94	B	61
26	NAYLA SYAHIDA	82,54	B	64
27	TOGA ANDREAS MANUELA	82,23	B	64
28	RANGGA JULION PRADITIA MANALU	82,23	B	44
29	SYIFA NAMIRA NASUTION	81,69	A	47
30	APRILIA STEVANI SIBURIAN	81,63	A	56

This stage will be calculated to determine the optimal accuracy value in determining the amount of test data and training data. Then the calculation will be carried out by determining how many test data and training data values are made in the form of random data. The distribution of data as a whole is divided into 9 forces which can be seen in the table.

Table 3. Distribution of Test Data and Training Data

No	Participant Name	Report Card Score	Home Distance
1	NANDA PAULINA SITOMPUL	87,9	13
2	SHAFA MAHARANI IMANTI POHAN	88,7	11
3	M, FATHAHILLAH AJI CAKRANINGRAT	71,9	14
4	WINDA SYAHRANI	89,74	16
5	NAILA FADILA	88,74	18
6	SALMA ULIMA BAHY	88,03	19
7	AMANDA CHAIRUNNISA	86,63	19
8	MANTASYA MAGFIRAH	86,31	22
9	GABRIEL PANDAPOTAN SIMANGUNSONG	86,03	30
10	JIHAN SALSABILLA	85,86	39
11	YULIA CITRA	85,8	5
12	DWI NABILAH	89,11	13
13	NOVA APRILIA	84,66	2
14	ARLINA INDRI MEYLISA BR, SIMANJUNTAK	84,51	9
15	ARNETTA MEI SUCI LESTARI	84,43	14
16	AULIA SHAWQI	84,34	5
17	MUHAMMAD GENTA SUDIRJA	84,14	16
18	FANI FITRI ANA	83,86	21
19	CHELSEA NABABAN	88,09	33
20	CHRISTINE RIANA SIMANJUNTAK	83,74	37
21	CHARLY YOSEP ARDANA SILALAHI	83,46	40
22	YEHEZKIEL DEO SUTANSHA HUTAGALUNG	83,37	50
23	NAILAH MAHDIAH PUTRI	87,66	54
24	PASKAH HAMONANGAN SIANTURI	83,11	55
25	NADA SORAYA	86,94	61
26	NAYLA SYAHIDA	82,54	64
27	TOGA ANDREAS MANUELA	82,23	64
28	RANGGA JULION PRADITIA MANALU	82,23	44
29	SYIFA NAMIRA NASUTION	81,69	47
30	APRILIA STEVANI SIBURIAN	81,63	56
31	GIZELA TESALONIKA LIMBONG	85,57	57
32	KAISYA AURA SYABILLAH	80,8	60
33	ANGGI RASMINI	84,83	65

The calculation is done by applying the Equal-width interval Discretisation technique to determine the interval used. The result is the number of correct data as many as 115 students with the formula.

$$Intervas = \frac{Value(max) - Value(min)}{|}$$

$$Intervas = \frac{89.7 - 71.9}{4}$$

$$Intervas = 4.4$$

After these calculations, the data is generated by grouping into k-mens clustering with the following results:

Table 4. 10% Testing Data and 90% Training Data

No	Entry Path	Total
1	Zoning	38 Students
2	Affirmation	28 Students
3	Achievement	22 Students
4	Moving Parents	27 Students

At this stage, data testing is carried out where the initial data will be entered into the Rapidminer10.1 application which will be tested using the k-means clustering method:

1. Rapidminer 10.1 Testing Results

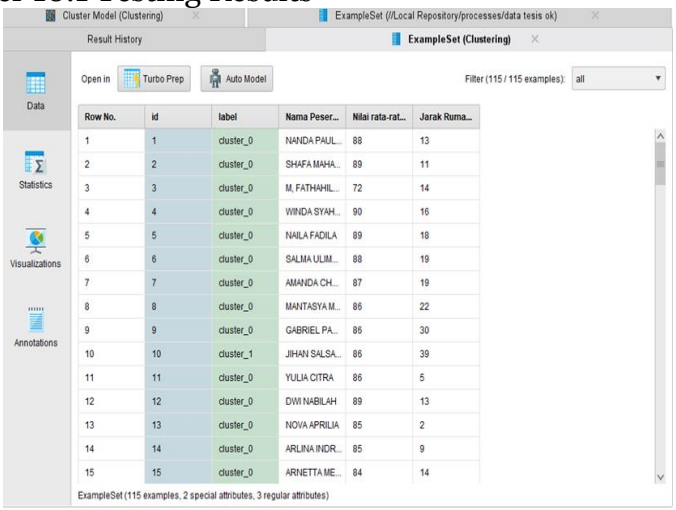

Figure 1. Testing Results

2. Comparison of Results

The next stage compares the results obtained with the calculation using the Equal-width interval Discretisation technique with the results using Rapidminer 10.1.

Table 5. Comparison of Results

Results	Y	X
Co	38	37
C1	28	29
C2	22	27
C3	27	22

In Table V the confusion matrix for value "Y" means "Yes" and value "T" means "No". Manual calculations to determine precision, recall, accuracy, weight average prediction, and weight average recall are as follows:

$$X = \frac{Same\ Amount\ of\ Data}{Number\ of\ Students}\ x\ 100\%$$

$$X = \frac{106\ Students}{115\ Students} x\ 100\%$$
$$X = 92{,}2\%$$

3. Accuracy Results

The results obtained by using the Equal-width interval Discretization technique and then tested using the Rapidminer 10.1 tool, the accuracy of student data is as follows:

Table 6. Result Accuracy

Results	Same Amount of Data	Accuracy
Equal-Width Interval Technique	106 New Students	92,2%
Tools Rapidminer 10.1	104 New Students	90,4%

With these results, it is concluded from the implementation that the calculation with the Equal-width interval Discretisation technique and tested with the help of Rapidminer 10.1 tools.

CONCLUSION

Based on the results obtained from the grouping of new students at SMA Negeri 9 Medan using the Equal Width Interval Discretization technique, the results of iterating student data that has the opportunity to enter from 4 groups, namely C0 (Zoning) as many as 38 students, C1 (Affirmation) as many as 28 students, C2 (Achievement) as many as 22 students, and C3 (Moving parents' duties) as many as 27 students. The results obtained from calculations with the Equal Width Interval Discretization technique have an accuracy (data similarity) with data from the school of 92.2% which is classified as accurate in the use of these techniques. The test results using Rapidminer 10.1 tools obtained an accuracy of 90.4% so that these results are still classified as accurate and can be used in the acceptance of new students in the future.

REFERENCES

Andini, T. I., Witanti, W., & Renaldi, F. (2016). Prediksi Potensi Pemasaran Produk Baru dengan Metode Naïve Bayes Classifier dan Regresi Linear. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*.

Ayuni, G. N., & Fitrianah, D. (2019). Penerapan metode Regresi Linear untuk prediksi penjualan properti pada PT XYZ. *Jurnal Telematika*, *14*(2), 79–86.

Gustientiedina, G., Siddik, M., & Deselinta, Y. (2019). Penerapan Naïve Bayes untuk Memprediksi Tingkat Kepuasan Mahasiswa Terhadap Pelayanan Akademis. *Jurnal Infomedia: Teknik Informatika, Multimedia, Dan Jaringan*, *4*(2), 89–93.

Hendrickx, T., Cule, B., Meysman, P., Naulaerts, S., Laukens, K., & Goethals, B. (2015). Mining association rules in graphs based on frequent cohesive itemsets. *Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part II 19*, 637–648.

Kamal, I. M., & Ilyas, R. (2017). Prediksi penjualan buku menggunakan data mining di pt. niaga swadaya. *Semnasteknomedia Online*, *5*(1), 1–2.

Lubis, C. P., Rosnelly, R., Roslina, R., Situmorang, Z., & Wanayumini, W. (2021). Penerapan Metode Naïve Bayes dan C4. 5 Pada Penerimaan

Pegawai di Universitas Potensi Utama. *CSRID (Computer Science Research and Its Development Journal)*, *12*(1), 51–63.
Lumbantoruan, R. (2017). Buletin Ekonomi ISSN: 1410-3842. *Buletin Ekonomi*, *21*(1).
Muin, A. A. (2016). Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi). *Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar*, *2*(1), 22–26.
Nofriansyah, D., Erwansyah, K., & Ramadhan, M. (2016). Penerapan Data Mining dengan Algoritma Naive Bayes Clasifier untuk Mengetahui Minat Beli Pelanggan terhadap Kartu Internet XL (Studi Kasus di CV. Sumber Utama Telekomunikasi). *J. Saintikom*, *15*(2), 81–92.
Nugraha, P., Aribawa, I. W., Priyana, I. P. O., & Indrawan, G. (2016). Penerapan Metode Decision Tree (Data Mining) Untuk Memprediksi Tingkat Kelulusan Siswa Smpn1 Kintamani. *Semin. Nas. Vokasi Dan Teknol*, 35–44.
Sulardi, P., Hendro, T., & Umbara, F. R. (2017). Prediksi Kebutuhan Obat Menggunakan Regresi Linier. *Prosiding SNATIF*, 57–62.
Zunaidhi, R., Saputra, W. S. J., & Sari, N. K. (2012). Aplikasi Peramalan Penjualan Menggunakan Metode Regresi Linier. *SCAN VOL. VII NOMOR 3, ISSN: 1978*, *87*.