JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website: http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN: 2085-3688; e-ISSN: 2460-0997

Application of the k-means clustering method and simple linear regression to new student admissions as a promotion method

Taufik Rahmat Kurniawan[1,*], Endang Chumaidiyah[2], Luciana Andrawina[3]
[1,2,3]Department of Magister Industrial Engineering, Telkom University
[1,2,3] Jl. Telekomunikasi, No.1, Bandung 40257, Indonesia
[*]Corresponding email: taurahkur@gmail.com

Received 28 November 2022, Revised 13 December 2022, Accepted 30 December 2022

Abstract — At private-label universities in Indonesia, new students are still the main thing when reporting university operating income. This study intends to classify student data at the Institut Teknologi Telkom Surabaya by utilizing the data mining process of k-means clustering. The clustering results are predicted using simple linear regression to find out new students' achievements. The results of this study consist of a combination of the highest profiles of students and parents obtained by the province of East Java, the system information study program, the parents' income of 5–10 million per month, the occupation of other parents, and the ethnicity of students from Java Island. then the highest forecasting results are in the income variable of students' parents in cluster 3, with predictions of 1292 students in 2024. It is hoped that with clustering and forecasting based on this research, Institut Teknologi Telkom Surabaya can make the right decisions as a basis for decision-making. Other than that, it can be used to develop a campus promotion strategy.

Keywords – forecasting, k-means clustering, simple linear regression

I. Introduction

The admission process of new students to private universities in Indonesia has always been a challenge. This is contrary to public universities, where the number of enthusiasts will increase yearly, even for the new study program. Quoting data from LTMPT, in 2022, the number of SBMPTN registered participants reach 800,852 people from a capacity of 214,406 (ltmpt.ac.id) [1]. This number means there are only no more than 27% of those students who would fit in the public universities in Indonesia. In Indonesia, there are 2,990 private universities spread throughout the country, compared to only 125 public universities (Indonesia Statistics 2012). This data shows that public universities in Indonesia are only 4% as large as private universities [2]. Referring to the data for SBMPTN applicants who ultimately did not pass in 2022, there are still at least 586.000 prospective students who can properly apply to private tertiary institutions [3].

Institut Teknologi Telkom Surabaya (ITTS) is a private university located in Surabaya, East Java, Indonesia. If viewed from an internal aspect, ITTS is adequate in its teaching capacity for students. Following are the details, namely lecturers and implementers: ITTS has 37 TPA (academic support staff) and 75 lecturers with doctoral degree qualifications, six people, and 69 people with master's degrees [4]. Then the following field is research and publication, and ITTS 2021 will have 38 internationally reputable journals and proceedings and 34 nationally reputed journals [5]. In addition, in the external capacity of tertiary institutions, the Community Service Institute of ITTS has aims to realize ten productive ICT-based self-guided areas and 43 collaborative services [6]. ITTS has several recognitions, including good accreditation by BAN-PT in 2021, ISO 9001:2015 certification, IOS 21001:2018 certification, the excellent campus award, and rank three webometrics [7].

However, the achievements of new students are currently still minimal [8]. It can be said that they have not reached the target that has been agreed upon with the management. In this case, the Telkom Education Foundation, which is contained in the KM

Table 1. Achievements of ITTS Students

No.	Years	Registration Target	Target Achievement Registration	Achievement (%)
1.	2018	280	136	49%
2.	2019	720	380	53%
3.	2020	1080	681	63%
4.	2021	1300	702	54%

Table 2. Initial Center Point of Each Cluster

Centroid	Active	Province	Male	Female	Academic Year	Study Program
C0	9	90	69	21	4	7
C1	9	425	274	151	4	5
C2	9	208	121	87	4	6
C3	1	3	3	0	4	7

(Management Contract) for the last four years, along with the achievement table [9].

According to Aldino *et al.* [10], clustering is a process for grouping data into several clusters or groups so that data within one cluster has the maximum level of similarity and data between clusters has a minimum level of similarity. According to Fathi *et al.* [11], forecasting is the art and science of predicting future events.

ITTS student data classify by utilizing the data mining process using the k-means clustering method. The results of the clustering are predicted using simple linear regression. It used to predict student achievement as a consequence variable and year as a causative variable. So, ITTS can make the right decisions regarding campus marketing strategies.

II. Research Method

This section discusses k-means clustering and simple linear regression.

A. K-Means Clustering

The clustering part of this research is to group data using the k-means clustering algorithm. The variables used to group data consist of ”Student Province”, ”Student Study Program”, ”Income of Student Parents”, ”Occupation of Student Parents”, and ”Student Tribe” [12]. Each of the five variables above has an attribute (in brackets for each variable). And these attributes will be grouped into several clusters for each variable.

Several steps must be carried out to do clustering using the k-means method [13].

1) Data source

The primary data sources used in this study are local databases of students, student study programs, parents' income, parents' occupations, and student tribes, with a total of 6335 records from the 2018–2021 class on seven study programs. The data received is in the form of screenshots of images from i-Gracias, which the author then converts into Microsoft Excel form so that it is easier to do data cleaning or filtering. This data was obtained with permission from ITTS [14].

2) Data selection

In this stage, the data is filtered and several attributes are taken from the table for analysis, as follows:

1) Student Province
   Six attributes are used: province, active, male, female, academic year, and study program. We are retrieving data based on the points used, and selecting data using the delete column, sort, and filter in the Microsoft Excel feature.
2) Student Study Program
   Class year, active, leave, inactive, graduated, DO, resigned, others, total students, and study program are the ten attributes that are used. Microsoft Excel's sort and filter features uses to retrieve data based on attributes and select data.
3) Income of Student Parents
   There are seven attributes used: study program, 1 million, 1-3 million, 3-5 million, 5-10 million, $>$ 10 million, and academic year. Microsoft Excel's sort and filter features uses to retrieve data based on attributes and select data.
4) Occupation of Student Parents
   Six attributes used: work, student, father, mother, academic year, and study program. This research uses Microsoft Excel's sort, and filter features to retrieve data based on attributes and to select data.
5) Student Tribe
   Six attributes used: province, active, male, female, academic year, and study program. Microsoft Excels's sort, and filter features use to retrieve data based on attributes and to select data.

3) Pre-processing data

At this stage, the process of changing the data is carried out, so the data can be processed using the k-means clustering algorithm. Non-numeric data is initiated into a numeric form. However, no inititation is requeired. The initiation process is as follows [15]:

1) Province
   Attributes by province are initiated in alphabetical order (initials 1-34).
2) Study program
   Attributes on study programs are initiated based on the first study program sequence (initials 1-7).
3) Academic year
   Attributes in the academic year are initiated based on the order of the year (initials 1-4).
4) Occupation of Student Parents
   Attributes to the work of students' parents are initiated in alphabetical order (initials 1-9).
5) Student Tribe
   Attributes on the student tribe are initiated al-

phabetically (initials 1-85).

4) Data processing

New student data is processed after the transformation process using the k-mean clustering method. The k-mean clustering algorithm procedure is as follows [16]:

1) k is based on the number of new clusters. For example, there will be four clusters created.
2) Determine the starting point of each cluster. The determination of the initial center point in this study was done randomly, and the center point obtained is as shown in the following Table 2.
3) Calculate the distance of each data point from the cluster center to the nearest centroid. The closest centroid will be found in the cluster following the data. The Euclidean distance d calculation can be done with (1), where x_1, and y_1 are the coordinate of the first point, and x_2 and y_2 are the coordinate of the second point.

$$d = \sqrt{\left[\left(x_2 - x_1\right)^2 + \left(y_2 - y_1\right)^2\right]} \tag{1}$$

As an example, we will calculate the distance from the first student province data to the first cluster center with.

$$d(1,1) = \sqrt{(34-9)^2 + (1-90)^2 + (1-69)^2} \\ \overline{+(0-21)^2 + (3-4)^2 + (1-7)^2} = 116.82$$

From the calculation above, it was found that the distance between the first new student data and the first cluster is 116.82. The equation used to calculate the distance between the first new student data and the center of the second cluster:

$$d(1,2) = \sqrt{(34-9)^2 + (1-425)^2 + (1-274)^2} \\ \overline{+(0-151)^2 + (3-4)^2 + (1-5)^2} = 527,01$$

From calculation above, it was found that the data distance of the first new student to the second cluster is 527.01. Distance from the center of the third cluster to the first new student data using the equation:

$$d(1,3) = \sqrt{(34-9)^2 + (1-208)^2 + (1-121)^2} \\ \overline{+(0-87)^2 + (3-4)^2 + (1-6)^2} = 255,86$$

From calculation above, it was found that the data distance of the first new student to the third cluster is 255.86. The distance of the first new student data to the fourth cluster center with the equation:

$$d(1,4) = \sqrt{(34-1)^2 + (1-3)^2 + (1-3)^2} \\ \overline{+(0-0)^2 + (3-4)^2 + (1-7)^2} = 33,67$$

From calculation above, it was found that the data distance for the first new student to the fourth cluster is 33.67. Based on the results of the four calculations above, it can be concluded that the closest data distance from the province of the first student is cluster 2, so the province of origin of the first new student is included in cluster 2 [17].

1) After all the data is placed in the nearest cluster, recalculate the new cluster center based on the average number of members.
2) If the new centroid converges with the old centroid, stop the iteration. If not, iteration moves on to the next.
3) Next, group the cluster results from the first iteration that has not converged. To regenerate a new centroid, we use (2), where c is the data centroids, m is a data member that belongs to a certain centroid, and n is the number of data that is a member of a certain centroid.

$$c = \frac{\sum m}{n} \tag{2}$$

B. Simple Linear Regression

Simple linear regression (SLR) is a statistical method that tests the extent of a causal relationship between the causal factor variable (X) and the consequential variable (Y). The X also called the predictor, while Y also called the response. SLR is a statistical method used in manufacturing to forecast or predict quality and quantity characteristics [18]. The following are the steps for performing a simple linear regression analysis:

1) Purpose

Determine the purpose of carrying out a linear regression analysis, namely studying the relationships obtained and expressed in a mathematical equation that states the relationship between variables.

2) Determines X and Y

X is the numbers of period, and Y is number of student achievements.

3) Collect information

The primary data source used in this method is the result of clustering with the variables of student provinces, student study programs, parents' income, parent's occupations, and student tribes, with a total of 6272 records from the 2018-2021 class with seven study programs. The data received is in the form of Microsoft Excel, making it easier to clean or filter the data.

4) Pre-processing

The data is filtered first, and two attributes are taken, namely X (year) and Y (student/student parent) from the tables above, to be analyzed based on clustering results.

5) Calculation of X, Y, XY, and XX

In the tables below, the "Year" column is formed in numerical form in column X. Column Y is the total number of students. Column XY is the result of multiplying the contents of columns X and Y, and

column XX contains the squared result of the contents of column X, as an example of calculating the student province variable.

Table 3. The Number of Each Cluster Student Province

Cluster	Total
Cluster 0	297
Cluster 1	6
Cluster 2	2
Cluster 3	12

Table 4. Student Provincial Cluster Results

Attribute	Cluster			
	0	1	2	3
Province	25.92	9.00	9.00	9.00
Active	2.17	245.00	72.58	425.00
Men	1.36	150.50	58.08	274.00
Woman	0.60	94.00	14.41	151.00
Academic years	3.14	3.50	2.66	4.00
Study program	3.96	5.50	3.83	5.00

6) Calculations of a and b

Calculating constants:

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2} \quad (3)$$

Calculating coefficients :

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \quad (4)$$

7) Get the Linear Regression Equation

Calculating Y:

$$Y = a + b \cdot x \quad (5)$$

After obtaining the constant values a and b, calculations are performed using (5) to get the value of the regression equation, or Y.

8) Make predictions

The Y equation that has been obtained can also be used to calculate estimates of student achievement in the coming year. In Table 3, the forecasting results for each cluster and variable using (5).

9) MAPE test

$$MAPE = \frac{\sum_{t=1}^{n} \left| \left(\frac{A_t - F_t}{A_t} \right) 100 \right|}{n} \quad (6)$$

A_t is the actual value of period t, F_t is the forecasted value of period t, and n is the number of periods [19]. The accuracy predicted results of student achievement above, a calculation is performed using the Mean Absolute Percentage Error (MAPE). The smaller the deviation between the predicted results and the actual conditions, the better that the prediction method used is good [20].

III. Result

In this section, clustering results will be displayed per variable. These results will also be used for forecasting for the next three years.

A. K-Means Clustering

The following are the results of clustering based on the following variables [21].

1) Student province

The following results for generating clusters using the student province variables in Table 3. By determining four clusters in the province, students are looking for similar groups, and the centroid distance between clusters can be seen. Table 4 shows the amount of data in each cluster. Validate the number of clusters using SSE; the more significant cluster, the better because SSE is getting smaller, as shown in Fig. 1. The discussion of the tests carried out in Table 5, by looking at cluster groupings.

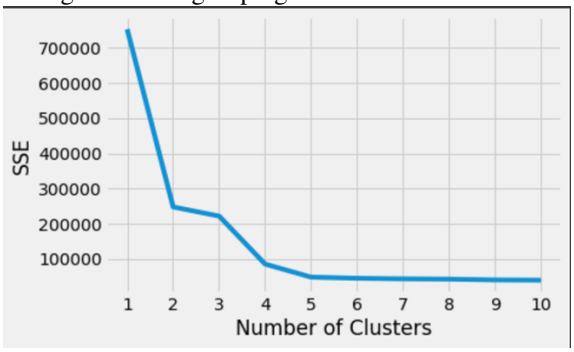

Fig. 1. Validating the number of clusters using SSE in student province.

2) Student study program

The following are the results for generating clusters on student study program variables in Table 6. By determining 4 clusters in the student study program data, we are looking for similar groups, and the centroid distance between clusters can be seen. Table 7 shows the amount of data in each cluster. Validate the number of clusters using SSE; the more significant cluster, the better because SSE is getting smaller, as shown in Fig. 2. The discussion of the tests carried out by looking at cluster groupings is presented in Table 8.

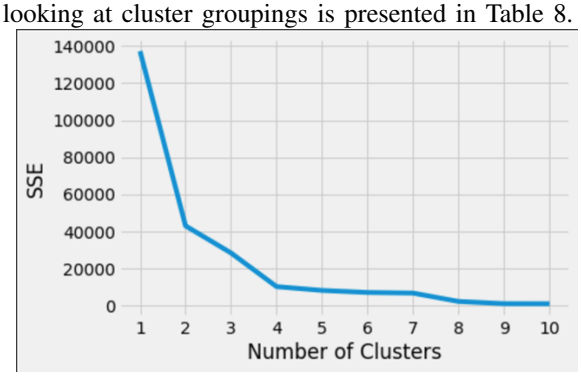

Fig. 2. Validating the number of clusters using SSE in study program.

3) Income of student parents

The following are the results for generating clusters on income variables for student parents in Table 9. By determining four clusters in the province, students are looking for similar groups, and the centroid distance between clusters can be seen. Table 10 shows the amount of data in each cluster. Validate the number of clusters using SSE; the more significant cluster,

Table 5. Profile per Province Cluster of Students

Cluster	Definition	Name
0	In non-East Java provinces (almost all provinces in Indonesia) with a total of 718 active students, a total of 480 males and a total of 177 females.	Variable 1: Student's Provincial Origin Cluster 0: non-East Java men and women equally.
1	In East Java province with the highest number of active students with a total of 950 people and the most academic year, namely the 2021/2022 class with 4 data records as well as information technology and Industrial Engineering study programs with each having 2 data records so that a total of 4 data records	Variable 1: Province of Origin Students Cluster 1: East Java, men and women are equal.
2	In the province of East Java, 353 students female gender are more dominant. The most academic years for the 2020/2021 and 2021/2022 batches are 2 data records and the information systems study program is the only study program with 2 data records.	Variable 1: Province of Student Cluster 2: In East Java, predominantly female.
3	In the province of East Java, students of a male gender were more dominant with 697 people and the most academic years, namely the 2019/2020 class with 6 data records and the most software engineering study programs with 3 data records. In the province of East Java, students of a male gender were more dominant with 697 people and the most academic years, namely the 2019/2020 class with 6 data records and the most software engineering study programs with 3 data records.	Variable 1: Province of Student Cluster 2: East Java, male dominance.

Table 6. The Number of Students per Study Program Cluster

Cluster	Total
Cluster 0	3
Cluster 1	11
Cluster 2	1
Cluster 3	13

Table 7. Cluster Results for Student Study Programs

Attribute	Cluster			
	0	1	2	3
Years	3.33	1.90	3.00	2.75
Active	1.22	1.98	2.19	5.35
Leave	1.33	9.09	1.00	4.61
Non-Active	4.00	9.09	7.00	1.23
Graduate	0.00	1.36	0.00	7.69
DO	0.00	0.00	0.00	0.00
Resign	2.33	1.72	4.00	2.30
Etc	0.00	1.38	1.00	7.69
Total Students	1.37	2.31	2.61	5.94
Study Program	4.33	4.18	5.00	3.69

the better because SSE is getting smaller, as shown in Fig. 3. The discussion of the tests carried out in Table 11, by looking at cluster groupings.

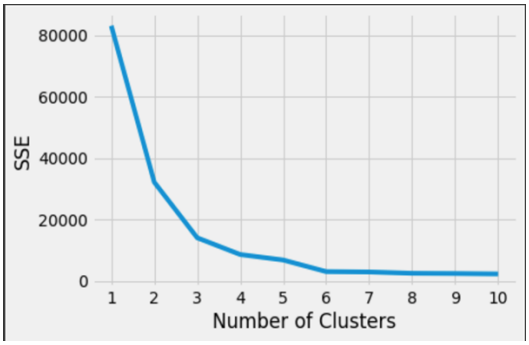

Fig. 3. Validating the number of clusters using SSE in income of student parents.

4) Occupation of student parents

The following are the results for generating clusters on occupation variables for student parents in Table 12. By determining four clusters in the province, students are looking for similar groups, and the centroid distance between clusters can be seen. Table 13 shows the amount of data in each cluster. Validate the number of clusters using SSE; the more significant cluster, the better because SSE is getting smaller, as shown in Fig. 4. The discussion of the tests carried out in Table 14, by looking at cluster groupings.

Fig. 4. Validating the number of clusters using SSE in occupation of student parents.

5) Student tribe

The following are the results for generating clusters on student tribe variables in Table 15. By determining four clusters in the province, students are looking for similar groups, and the centroid distance between clusters can be seen. Table 16 shows the amount of data in each cluster. Validate the number of clusters using SSE; the bigger the cluster, the better because SSE is getting smaller, as shown in Fig. 5. The discussion of the tests carried out by looking at cluster groupings in Table 17.

Fig. 5. Validating the number of clusters using SSE in student tribe.

B. Simple Linear Regression

At this stage, the data processed is the result of clustering, which consists of five variables, namely

Table 8. Profile per Study Program Cluster of Students

Cluster	Definition	Name
0	The total number of active students in the information technology study program, information systems, and industrial engineering is almost the same, namely around 100 people.	Variable 2 : Student study program Cluster 0: Informatics and industry study programs are dominant.
1	The electrical engineering study program is more dominant with 3 data records compared to other study programs with an average of 2 data records. The highest total number of graduating students is 15 people.	Variable 2: Student study program Cluster 1: Dominant in Electrical Engineering.
2	The information systems study program, is the only study program with 1 data record, and the highest number of active students in the study program is 261 people.	Variable 2: Student study program Cluster 2: Dominant information system.
3	The telecommunications engineering study program is more dominant, with 3 data records, compared to other study programs, with an average of 2 data records. The total number of active students in all study programs in the cluster is 696 people.	Variable 2: Student study program Cluster 3: Dominant in telecommunications engineering.

Table 9. The Variable Number of Parents' Income for Each Cluster

Cluster	Total
Cluster 0	11
Cluster 1	6
Cluster 2	1
Cluster 3	10

Table 10. The results of the Student Parent Income variable cluster

Attribute	Cluster			
	0	1	2	3
Study Program	3.90	4.16	5.00	3.90
<1 Million	8.36	16.50	24.00	2.80
1 - 3 Million	14.54	29.83	63.00	5.40
3 – 5 Million	36.27	72.00	163.00	7.20
5- 10 Million	30.36	72.16	156.00	6.80
>10 Million	12.00	29.66	75.00	2.20
Academic year	2.90	3.50	4.00	1.30

Table 11. Variable Income Profile of Student's Parents per Student Variable

Cluster	Definition	Name
0	The student parent's income is dominant in the range of 3-5 million as many as 399 people And in the range of 5-10 million, as many as 334 people.	Variable 3: Income of Student's Parents Cluster 0: dominant salary 3-5 million and 5-10 million
1	The income of student parents is evenly distributed throughout the salary range of 1321 people.	Variable 3: Income of Student's Parents Cluster 1: evenly distributed throughout the salary range
2	The income of parents of students with a salary range below 1 million is at least 24 people	Variable 3: Income of Student's Parents Cluster 2: salary below 1 million at least
3	The income of student parents with the highest salary range of 3-5 million is 72 people	Variable 3: Income of Student's Parents Cluster 3: the highest salary range is 3-5

Table 12. The Number of Each Cluster Occupation Student Parents Variables

Cluster	Total
Cluster 0	136
Cluster 1	38
Cluster 2	2
Cluster 3	10

Table 13. Occupation Student Parents Cluster Number

Attribute	Cluster			
	0	1	2	3
Job	5.47	4.18	2.00	3.20
Student	9.54	56.78	268.50	134.50
Father	7.36	35.34	67.00	68.40
Mother	3.72	32.55	248.00	92.10
Academic Year	2.41	3.23	3.50	3.70
Study Program	4.00	4.00	5.00	4.10

Table 14. Profile per Student Parents Occupation Variables of Students

Cluster	Definition	Name
0	The work of parents of students with all types of work is evenly distributed	Variable 4: Occupation of Student's Parents Cluster 0: all types of work are evenly distributed.
1	The work of parents of students with the dominant type of work as private employees is 730 people.	Variable 4: Occupation of Student's Parents Cluster 1: private employee jobs with 730 people.
2	The work of parents of students with dominant other types of work (mothers) is 537 people.	Variable 4: Occupation of Student's Parents Cluster 2: other jobs (mothers) with 537 people.
3	The work of parents of students with dominant other types of work (fathers and mothers) is 259 people.	Variable 4: Occupation of Student's Parents Cluster 3: other occupations (fathers and mothers) with 259 people.

Table 15. The Number of Each Cluster Student of Parents Variables

Cluster	Total
Cluster 0	186
Cluster 1	165
Cluster 2	2
Cluster 3	14

Table 16. Results of the Clustering of Student Parents

Attribute	Cluster			
Attribute	0	1	2	3
Job	6.97	1.82	3.20	3.20
Student	1.40	3.44	3.43	1.16
Father	1.08	2.46	2.19	8.70
Mother	3.22	9.87	1.23	2.92
Academic Year	2.23	3.02	3.50	3.28
Study Program	4.03	3.91	5.00	3.78

Table 17. Profile per Student Parents Variables of Students

Cluster	Definition	Name
0	In student tribes with all tribes evenly distributed	Variable 5 : Student Tribe Cluster 0: all tribes are evenly distributed
1	In the student tribe, the dominant tribe is Javanese compared to others.	Variable 5: Student Tribe Cluster 1: evenly dominated by Javanese
2	In the student tribe with Javanese ethnicity there are 686 people.	Variable 5: Student Tribe Cluster 2: Javanese
3	In the student tribe with the highest Javanese tribe, the total was 1628.	Variable 5: Student Tribe Cluster 3: the most Javanese

student provinces, student study programs, income of student parents, parental occupations, and student tribe, from seven majors at ITTS. First, the data is processed by removing inconsistent data, correcting errors, and enriching it with relevant external data. Then, the data is transformed to change it from its original form into a form suitable for grouping [22].

In Table 18, the "Year" column is formed in numerical form in column X. Column Y is the total number of students [23]. Column XY is the result of multiplying the contents of columns X and Y, and column XX contains the squared result of the contents of column X [24].

Table 18. Calculation results X, Y, XY, XX Student Province

Cluster 0					
Year (X)	**Total Student** (Y)	X	Y	XY	XX
2018	111	1	111	111	1
2019	105	2	105	210	4
2020	261	3	261	783	9
2021	241	4	241	964	16
Total		10	718	2068	30
Mean		2.05	180		
Cluster 1					
Year (X)	**Total Student** (Y)	X	Y	XY	XX
2018	0	1	0	0	1
2019	0	2	0	0	4
2020	305	3	305	915	9
2021	645	4	645	2580	16
Total		10	950	3495	30
Mean		2.05	238		
Cluster 2					
Year (X)	**Total Student** (Y)	X	Y	XY	XX
2018	0	1	0	0	1
2019	0	2	0	0	4
2020	282	3	282	846	9
2021	425	4	425	1700	16
Total		10	707	2546	30
Mean		2.05	177		
Cluster 3					
Year (X)	**Total Student** (Y)	X	Y	XY	XX
2018	0	1	0	0	1
2019	399	2	399	798	4
2020	328	3	328	984	9
2021	144	4	144	576	16
Total		10	871	2358	30
Mean		2.05	218		

Table 19 shows the result from calculating a and b for each variable [25]. Meanwhile, Table 20 is the predicted result of each variable for the next three years [26]. There is a possibility that the predicted results of student achievement above will be accurate. A calculation is performed using the MAPE test [27]. The smaller the deviation between the predicted results and the actual conditions, the better the prediction method used.

IV. Discussion

The k-means method for clustering data in this study is still a simple method. Many other methods are better for grouping data. However, the results obtained are still insufficient to classify the data according to the facts in the field. Moreover, the data obtained was separated by variables, so the researchers made five

Table 19. Calculation Results a and b

No.	**Variable**	**Cluster**	**Result**	
1	Student Province	Cluster 0	a	-40,5
			b	42,4
		Cluster 1	a	175,5
			b	-13,3
		Cluster 2	a	-206
			b	151
		Cluster 3	a	-345
			b	300,8
2	Student Study Program	Cluster 0	a	-85
			b	75,2
		Cluster 1	a	105
			b	-16,5
		Cluster 2	a	0
			b	26,1
		Cluster 3	a	91
			b	40,9
3	Income of Student Parents	Cluster 0	a	-62
			b	149,1
		Cluster 1	a	-159,5
			b	95,7
		Cluster 2	a	164
			b	-11,2
		Cluster 3	a	-306,5
			b	228,3
4	Occupation of Student Parents	Cluster 0	a	244,5
			b	32,0
		Cluster 1	a	-312
			b	340,6
		Cluster 2	a	-162,5
			b	118,7
		Cluster 3	a	-487
			b	329,3
5	Student Tribe	Cluster 0	a	-40,5
			b	42,4
		Cluster 1	a	175,5
			b	-13,3
		Cluster 2	a	-206
			b	151
		Cluster 3	a	-345
			b	300,8

Table 20. The Results of Forecasting Calculations for the Next Three Years

No	**Variable**	**Cluster**	**Year**		
			2022	2023	2024
1	Student Province	Cluster 0	316	371	425
		Cluster 1	798	1022	1246
		Cluster 2	566	722	877
		Cluster 3	308	344	380
2	Student Study Program	Cluster 0	291	366	441
		Cluster 1	23	6	-11
		Cluster 2	131	157	183
		Cluster 3	296	335	377
3	Income of Student Parents	Cluster 0	684	833	982
		Cluster 1	319	415	510
		Cluster 2	108	97	86
		Cluster 3	835	1063	1292
4	Occupation of Student Parents	Cluster 0	405	437	469
		Cluster 1	1391	1732	2072
		Cluster 2	431	550	668
		Cluster 3	1160	1489	1818
5	Student Tribe	Cluster 0	172	214	256
		Cluster 1	109	96	82
		Cluster 2	549	700	851
		Cluster 3	1159	1460	1761

Table 21. Range MAPE

MAPE	**Interpretation**
<10	Highly accurate forecasting
10-20	Good forecasting
20-50	Reasonable forecasting
>50	Inaccurate forecasting

Table 22. Accuracy Testing Results

No.	Variable	Cluster	MAPE
1	Student Province	0	22%
		1	6%
		2	3%
		3	32%
2	Student Study Program	0	17%
		1	84%
		2	18%
		3	20%
3	Income of Student Parents	0	16%
		1	8%
		2	27%
		3	4%
4	Occupation of Student Parents	0	28%
		1	25%
		2	3%
		3	12%
5	Student Tribe	0	13%
		1	33%
		2	3%
		3	6%

tests per variable. It is hoped that the data collected by the i-Gracias system will be better in the future. For example, one row of data already contains five variables or attributes that the author is currently researching, namely: student provinces, student study programs, the income of student parents, occupation of student parents, and student tribe.

From the simple linear regression carried out in this study and the MAPE test. At least the predictions are good; there are only seven clusters. Meaning that there are 13 other clusters whose accuracy is still not good. And even 1 out of 20 test results is not feasible to use. This research can be continued by testing different methods, and hopefully get better outcomes [29].

Clustering in this study is focused on identifying groups with certain characteristics so that they can easily be mapped into appropriate marketing strategies. Such as, for example, cluster 0, where student provinces are more dominant than non-East Java provinces, so the appropriate strategy is, for example, internet marketing. Then, linear regression to make forecasts for each cluster so that the management knows the achievements of each cluster in the following years and can make the right decisions regarding the marketing strategy to achieve these forecast numbers.

V. Conclusion

The results of this study consisted of 20 clusters, with each variable composed of four clusters. It can be concluded that the highest combination of student and parent profiles was obtained from East Java province, the information systems study program, parental income of 5–10 million per month, the work of other parents, and the ethnicity of students from Java Island. Each cluster is predicted for the next three years, with the highest forecasting results in the income variable of parents of students in cluster 3, with predictions of 1292 students in 2024. It is hoped that ITTS can make the right decision regarding marketing strategy.

References

[1] D. D. Fattoxovna, S. X. Xamimovna, and M. R. Ergasheva, "Promotion of practival trainings for the development of the creative abilities of students in special subjects using foreign methods of foreign education method," *European Journal of Research and Reflection in Educational Sciences*, vol. 8, no. 11, 2020, [Online]. Available: www.idpublications.org

[2] R. F. Siahaan, L. Simbolon, S. P. Nusantara, and J. Iskandar Muda No, "Promotion media recommendations on the acceptance of new students in private educations with the simple additive weighting method," *International Journal of Information System & Technology*, vol. 4, no. 1, 2020. [Online]. Available: http://forlap.ristekdikti.go.id

[3] E. Siswanto and A. W. Katili, "Implementation of decision support system for campus promotion management using fuzzy multiple analytic decision making (FMADM) method (case study: Universitas multimedia nusantara)," in *2017 4th International Conference on New Media Studies (CONMEDIA)*, Nov. 8–10, 2017.

[4] T. I. Oweis, "Effects of using a blended learning method on students' achievement and motivation to learn English in Jordan: A pilot case study," *Educ Res Int*, vol. 2018, 2018, doi: 10.1155/2018/7425924.

[5] A. H. Mirza, "Application of naive bayes classifier algorithm in determining new student admission promotion strategies," *Journal of Information Systems and Informatics*, vol. 1, no. 1, 2019, [Online]. Available: http://journal-isi.org/index.php/isi

[6] S. Abadi et al., "Application model of k-means clustering: Insights into promotion strategy of vocational high school," *International Journal of Engineering and Technology (UAE)*, vol. 7, no. 2.27 Special Issue 27, pp. 182–187, 2018, doi: 10.14419/ijet.v7i2.11491.

[7] E. A. Vetrova, E. E. Kabanova, N. v. Medvedeva, and E. E. Jukova, "Management of educational services promotion in the field of higher education (the example of 'Russian State Social University')," *European Journal of Contemporary Education*, vol. 8, no. 2, pp. 370–377, Jun. 2019, doi: 10.13187/ejced.2019.2.370.

[8] U. Kango, A. Kartiko, B. Zamawi, I. Pesantren, and K. A. Chalim, "The effect of service quality, facilities and promotion on the interest of new students," *Nidhomul Haq: Jurnal Manajemen Pendidikan Islam*, vol. 6, no. 1, 2021, doi: 10.31538/ndh.v6i2.1447.

[9] A. Heryati and M. I. Herdiansyah, "The application of data mining by using k means clustering method in determining new students' admission promotion strategy," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 3, pp. 824–833, Feb. 2020, doi: 10.35940/ijeat.C5414.029320.

[10] A. A. Aldino, D. Darwis, A. T. Prastowo, and C. Sujana, "Implementation of k-means algorithm for clustering corn planting feasibility area in South Lampung Regency," *J. Phys. Conf. Ser.*, vol. 1751, no. 1, 2021, doi: 10.1088/1742-6596/1751/1/012038.

[11] M. Fathi, M. H. Kashani, S. M. Jameii, and E. Mahdipour, "Correction to: Big data analytics in weather forecasting: A systematic review (Archives of computational methods in engineering, (2021), 10.1007/s11831-021-09616-4)," *Archives of Computational Methods in Engineering*, vol. 29, no. 1, p. 733, 2022, doi: 10.1007/s11831-021-09630-6.

[12] Y. Hozumi, R. Wang, C. Yin, and G. W. Wei, "UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets," *Comput. Biol. Med.*, vol. 131, Apr. 2021, doi: 10.1016/j.compbiomed.2021.104264.

[13] F. Yunita, "Penerapan data mining menggunkan algoritma k-means clustring pada penerimaan mahasiswa baru," *Sistemasi*, vol. 7, no. 3, p. 238, 2018, doi: 10.32520/stmsi.v7i3.388.

[14] M. Jahangoshai Rezaee, M. Eshkevari, M. Saberi, and O. Hussain, "GBK-means clustering algorithm: An improvement to the k-means algorithm based on the bargaining game," *Knowl Based Syst.*, vol. 213, Feb. 2021, doi: 10.1016/j.knosys.2020.106672.

[15] I. D. Borlea, R. E. Precup, A. B. Borlea, and D. Iercan, "A unified form of fuzzy c-means and k-means algorithms and its partitional implementation," *Knowl. Based Syst.*, vol. 214, Feb. 2021, doi: 10.1016/j.knosys.2020.106731.

[16] A. Sulistiyawati and E. Supriyanto, "Implementasi algoritma k-means clustering dalam penentuan siswa kelas unggulan," *Jurnal Tekno Kompak*, vol. 15, no. 2, 25–36, 2021.

[17] P. Garikapati, K. Balamurugan, T. P. Latchoumi, and R. Malkapuram, "A cluster-profile comparative study on machining AlSi7/63% of SiC hybrid composite using agglomerative hierarchical clustering and k-means," *Silicon*, vol. 13, no. 4, pp. 961–972, Apr. 2021, doi: 10.1007/s12633-020-00447-9.

[18] H. Hasanah, A. Farida, and P. P. Yoga, "Implementation of simple linear regression for predicting of students' academic performance in mathematics," *Jurnal Pendidikan Matematika (Kudus)*, vol. 5, no. 1, p. 38, Jun. 2022, doi: 10.21043/jpmk.v5i1.14430.

[19] W. A. L. Sinaga, S. Sumarno, and I. P. Sari, "The application of multiple linear regression method for population estimation Gunung Malela District," *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 1, pp. 55–64, Mar. 2022, doi: 10.55123/jomlai.v1i1.143.

[20] E. Rahayu, I. Parlina, and Z. A. Siregar, "Application of multiple linear regression algorithm for motorcycle sales estimation," *JOMLAI: Journal of Machine Learning and Artificial Intelligence*, vol. 1, no. 1, pp. 1–10, Mar. 2022, doi: 10.55123/jomlai.v1i1.142.

[21] Z. Nabila, A. Rahman Isnain, and Z. Abidin, "Analisis data mining untuk clustering kasus covid-19 di Provinsi Lampung dengan algoritma k-means," *Jurnal Teknologi dan Sistem Informasi (JTSI)*, vol. 2, no. 2, p. 100, 2021, [Online]. Available: http://jim.teknokrat.ac.id/index.php/JTSI

[22] W. Nurpadilah, I. M. Sumertajaya, and M. N. Aidi, "Geographically weighted regression with kernel weighted function on poverty cases in West Java Province," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 1, pp. 173–181, Mar. 2021, doi: 10.29244/ijsa.v5i1p173-181.

[23] D. R. S. Saputro, Y. K. Wardani, N. B. I. Pratiwi, and P. Widyaningsih, "Data simulation with markov chain monte carlo, gibbs sampling, and bayes (beta-binomial) methods as the parameter estimations of spatial bivariate probit regression model," in *AIP Conference Proceedings*, Feb. 2021, vol. 2326. doi: 10.1063/5.0040332.

[24] N. Wiliani, R. Hesananda, N. S. Rahmawati, and E. H. Prianggara, "Application of machine learning for bitcoin exchange rate prediction against us dollar," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 7, no. 2, pp. 67–74, Feb. 2022, doi: 10.33480/jitk.v7i2.2880.

[25] A. L. R. Jauhari, R. L. Ariany, F. Fardillah, and A. Ayu, "Profile of students' statistical reasoning capabilities in introductory social statistics courses," in *Journal of Physics: Conference Series*, Feb. 2021, vol. 1764, no. 1. doi: 10.1088/1742-6596/1764/1/012118.

[26] B. Abdul Wahid, "2568 Accredited "Rank 4," 2021. [Online]. Available: https://iocscience.org/ejournal/index.php/mantik

[27] N. S. Wibowo, M. Aziziah, I. G. Wiryawan, and E. Rosdiana, "Implementasi metode regresi linier pada rancang bangun sistem informasi monitoring nutrisi tanaman hidroponik kangkung," *JTIM: Jurnal Teknologi Informasi dan Multimedia*, vol. 4, no. 1, pp. 13–24, May 2022, doi: 10.35746/jtim.v4i1.186.

[28] M. S. Rahman, "Analysis regression and path model: The influence both instagram and tiktok in improving students' vocabulary," *Sketch Journal*, vol. 1, no. 1, 2021.

[29] P. R. Sihombing, "Aplikasi pemodelan logit, probit dan cloglog pada regresi binomial (Studi kasus: pemodelan penyakit jantung)," *Jurnal Multidisiplin Madani*, vol. 2, no. 6, pp. 2599–2610, 2022, doi: 10.55927/mudima.v2i6.430.