STUDENT SEGMENTATION TO DETERMINE PROMOTION STRATEGIES FOR NEW STUDENT ADMISSIONS USING THE K-MEANS ALGORITHM

Leny Tritanto Ningrum[1*] and Istiqoomatun Nisaa[2]
[1]*Information System Study Program, Binaniaga Indonesia University Bogor*
[2]*Industrial Management Study Program, Telecomunication Academy Bogor*

Abstract

College competition every academic year always increases, this is clearly seen in student recruitment where the right strategy is needed to attract students to apply. Because of this competition, the number of students who enroll in universities each registration period decreases, which causes financial losses for the university management. The purpose of this study is to group student data into 4 segments to determine the characteristics of students who register based on the sources of information they get. Information sources are promotional strategies applied by universities including brochures, posters, advertisements, social media, whatsapp, and / or broadcasar SMS. The data used is student data for the academic year 2020, 2021, and 2022. The method used in this study is the K-Means algorithm with the attributes used, namely subdistrict, school origin, study program, class, year of entry, parental occupation, and source of information. The result of this study is the formation of student data groups into 4 segments that refer to more dominant, standard, low, and less information sources so as to provide advice to university management to implement the most appropriate promotion strategy to increase the number of students who register.

Keywords: Data Mining, Enrollment, K-Means, Segmentation, Student Data

Introduction

Currently, competition to get students at universities is increasingly visible. This has led to more and more college choices that are considered by high school graduate students to continue their education to a higher level (Kristanto et al., 2020). Universities need to implement the right promotion strategy to attract prospective applicants either from the community or high school students so that the number of students does not continue to decline (Sopandi, 2011). The strategy used is usually in the form of distributing brochures, banners, advertisements, scholarships, and tuition discounts (Anggreini, 2019). To determine the right promotion strategy, universities can utilize stored student data to explore knowledge and understand patterns of student data characteristics by segmenting data groupings in order to get the right combination of promotional strategies and increase the interest of prospective students who register(Silalahi, 2020).

Based on the description in the introduction, it can be identified that the problems faced by universities, especially private universities, are:

1. The decreasing number of students applying
2. Not exactly the promotion strategy applied
3. The amount of data stored in the database that has never been analyzed to extract knowledge from the data is not impossible to provide guidance on the implementation of the right strategy

The objectives to be achieved in this study are: Menganalisa data mahasiswa agar dapat ditemukan pengetahuan yang bermanfaat

1. Student data segmentation
2. Determining the right promotion strategy
3. Increase the number of students applying

Theoretical Basis

Student data segmentation is carried out to determine the pattern of characteristics of students who register in order to strive for promotional strategies. The research method applies a data mining model using the K-Means algorithm where the existing student data will be grouped so that it can form student segmentation based on the characteristics of the data.

a. Data Mining

Data mining is applied to analyze data from various different perspectives by referring conclusions into useful information to increase profits, cost efficiency, promotional marketing patterns, or others (Yunita, 2018). One method of data mining that can be done is a data clustering model that groups data into several clusters (Sinaga & Yang, 2020).

b. CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) is a standard process for implementing data mining that was developed with the aim of conducting industry analysis as a strategy for solving a problem from business or research (Suhanda, Kurniati, & Norma, 2020). The steps for implementing CRISP-DM consist of 6 stages, namely Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment, as shown in figure 1.

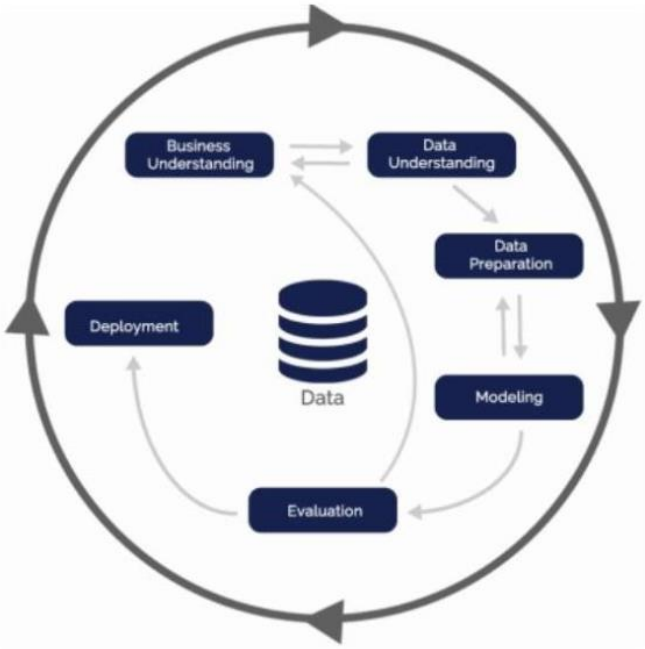

Figure 1. CRISP-DM Step

1. Business Understanding Phase
    a. Determination of project objectives and needs in detail within the scope of the business or research unit as a whole.
    b. Translate goals and constraints into formulas of data mining problems.
    c. Prepare the initial strategy to achieve the goal.
2. Data Understanding Phase
    a. Collecting data, if the data comes from more than one database then the data integration process or Data Integration is carried out
    b. Develop data inquiry analysis to further identify data and seek prior knowledge.
    c. Evaluating data quality, checking data and cleaning invalid data or Data Cleaning process.
    d. If desired, select a small number of data groups that may contain patterns from the problem.

3. Data Preparation Phase

    a. Prepare the initial data, the data set that will be used for the entire next phase or the Data Selection process.
    b. Select cases and variables to be analyzed, according to the analysis to be carried out.
    c. Make changes to variables if needed.
    d. Prepare the initial data so that it is ready for modeling tools or Data Transformation.

4. Modeling Phase
    a. Select and apply appropriate modeling techniques.
    b. Calibrate model rules to optimize results.
    c. Can use some of the same techniques for the same problem.
    d. Can return to the data processing phase if needed to transforming data into a form of specific needs

5. Evaluation Phase
    1. Evaluate the modeling stage to ensure that the model results are in accordance with the research

6. Deployment Phase
    1. Using the generated and presented model or knowledge presentation process.

c. Algoritma K-Means

The K-Means algorithm is one of the clustering algorithms included in the unsupervised learning category that is able to analyze data based on data distance in euclidean distanace (Abriyanto & Damastuti, 2019). The steps of grouping data with the K-Means algorithm are: (Nur, Zarlis, & Nasution, 2017) :

a. Determining centroid value
b. Select K from the data set as centroid
c. Allocate data to the nearest centroid using the distance matrix
d. Recalculate centroid values based on each cluster
e. Repeat steps 3 and 4 until the cluster allocation does not change

d. Study Review

In each study, it is necessary to review its relation to previous studies to analyze the problem and the application of the methods used. These studies are used as a reference for solving problems and methods so that research sustainability can be carried out.

This research will analyze student data to be segmented so that the characteristics of students who register can be explored more deeply and can be used as suggestions and input to determine the promotion strategy that best suits students.

Some previous studies related to this research include research conducted by (Sugiarti, Nahulae, Panggabean, & Sianturi, 2018) about decision support system to determine campus promotion strategy policies with the Weight Aggregated Sum Product Assessment method (WASPAS) with the aim of determining promotion policies on campus. Other research conducted by (Sandag, Yahuda Putra, Wurangian, &; Believer Tulangow, 2019) about strategy analysis to increase the target of new students using the K-Means algorithm which is motivated by problems due to the accumulation of student data that has never been analyzed to find out the knowledge and possible information that can be obtained from the data. Further research conducted by (Handayanto, Latifa, Saputro, & Waliansyah, 2019) about the analysis and application of Support Vector Machine (SVM) algroitms in data mining to determine promotional strategies and described that data processing is carried out to obtain data patterns to explore knowledge.

e. Research Framework

The promotion strategy is a way for universities to attract prospective students to register as new students. Various ways are done for campus promotion, but in reality the target achieved has not been as desired and even lower than

the predicted estimate. For this reason, a way is needed to assist the management in implementing any promotional strategy in accordance with the times and technology in order to achieve the desired target. Based on this problem, it is necessary to make it in an appropriate research framework, while the concept of research carried out in this process is as seen in figure 2.

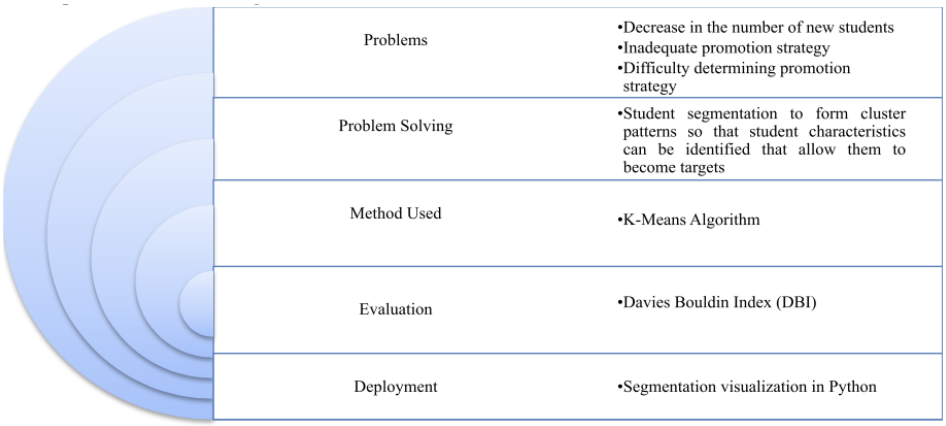

Figure 2. Research Framework

Methodology

This research was conducted to determine the right promotion strategy for new student admissions using the K-Means algorithm method. Based on the process flow, this research refers to the standard concept of the data mining process, namely CRISP-DM which consists of 6 steps, including: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment is as seen in figure 3.

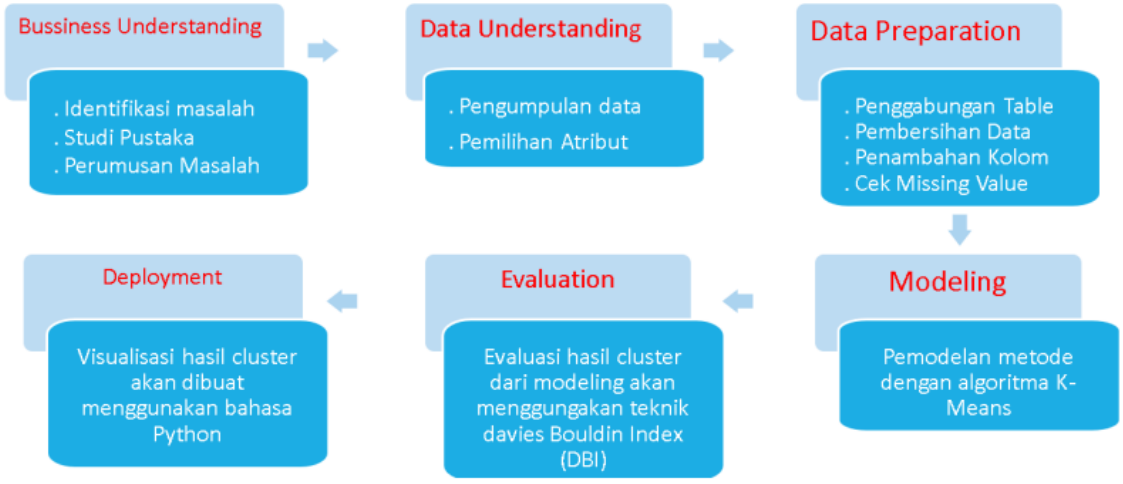

Figure 3. Research step using CRISP-DM

a. Business Understanding

The problem complained by the university manager is related to the number of new students who register has decreased. So that this research is expected to form a student segment by finding cluster patterns from student data

that registers. Based on the student data cluster pattern, the characteristics of the students who register will be seen which will then be visualized to display the promotion strategy chosen based on the cluster formed.

1. Identify the Problem
   In this phase, an analysis of the object of research is carried out and problems that occur in the university environment, especially in terms of new student admissions.
2. Literature Study
   At this stage, various literature studies related to problems related to problem identification are studied. Literature studies are obtained from various journal references with the same problem to be continued in the problem formulation process.
3. Problem Statement
   At this stage, the formulation of the existing problem is explained so that it can be adjusted to the research objectives

b. Data Understanding

This stage is a data collection process carried out to obtain primary data through documentation techniques by looking directly at the sources of documents related to research. Data collection is carried out by extracting student data sourced from the Academic Information System. The documentation data obtained is in the form of student data in several registration periods. Student data obtained is student data in the 2020-2021 and 2021-2022 Academic Years. Based on references from several previous research sources, in this study the attributes to be used include sources of cost, school origin, region, class, parental occupation, year of entry, and sources of information. At this stage there are 2 processes, namely:

1. Data Collecting
   The process of collecting data and understanding the content of the data is completed at this stage so that the desired information can be known.
2. Feature Selection
   After understanding the content of the existing data, the next stage is the selection of the attributes contained in the data which will then be adjusted to the research topic. Overall, the student table has 41 attributes, but based on references from a review of studies related to similar research that has been done before, this study will only use 7 attributes, namely: region, school origin, source of cost, class, year of entry, parent occupation, and source of enrollment information.

c. Data Preparation

1. Merging tables, namely based on existing data, the data needed for research is separated into 2 tables, namely data from 2020-2021, and 2021-2022. For data analysis purposes, the three data are combined into a unified table that is intact and can be used as a student dataset.
2. The cleaning process, which is after having the appropriate data, the next step is to clean the data because there are several records whose values do not match. Cleaning is carried out to produce valid data as needed so as to facilitate the data processing process by analyzing data quality by changing, correcting or deleting data that is not in accordance with research needs. The cleaning process, which is after having the appropriate data, the next step is to clean the data because there are several records whose values do not match. Cleaning is carried out to produce valid data as needed so as to facilitate the data processing process by analyzing data quality by changing, correcting or deleting data that is not in accordance with research needs.

c. Modeling

The initial stage of the modeling stage with the K-Means algorithm is that a randomly selected record will be used as a centroid (data center) which will then calculate the distance between the centroid and other data so that the closest data distance is obtained and forms a data group (cluster).

d. Evaluation

In this process, testing of cluster results from the modeling stage is carried out to determine the best cluster pattern to be used. The testing technique will be carried out using the Davies Bouldin Index (DBI) method. The cluster will be used as a basis for determining what promotional strategy will be applied.

e. Deployment

The results of the evaluation stage will be visualized at the deployment stage using Python language so that it can be clearly seen the cluster formed and can provide an overview to the promotion team to determine the right combination of promotion strategies.

Results and Discussions

a. Business Understanding

Based on the background of the problem, interviews and observations made on the object of research, in this phase it can be identified that the problems that occur are:

1. The decline in the number of new student admissions tends to decrease from year to year.
2. The promotion strategy applied has not been on target
3. Management has difficulty in determining promotion strategies

Therefore, in this study, clustering of student data will be carried out using the K-Means algorithm to determine the right promotion strategy so that it can help the management to make decisions.

b. Data Understanding

At this stage, what is done is the process of understanding the data. The data obtained is in the form of student data for the class of 2020-2021 and 2021-2022 with a total of 824 records from all study programs. Overall, the number of attributes contained in the student data is 41, but only 7 attributes will be used. In detail, these attributes can be seen in table 1.

Tablel 1. Student data attributes

No	Nama Atribut	Tipe Data	Keterangan
1.	Wilayah	String	Wilayah tempat tinggal mahasiswa berdasarkan area berupa Kabupaten, Kota, atau luar wilayah Bogor
2.	Asal Sekolah	String	Asal Sekolah Mahasiswa
3.	Pekerjaan Orang Tua	String	Pekerjaan Orang Tua
4.	Sumber biaya	String	Sumber dana pembiayaan kuliah mahasiswa
5.	Kelas	String	Kelas Perkuliahan Mahasiswa
6.	Tahun Masuk	String	Tahun masuk mahasiswa
7.	Info Pendaftaran	String	Sumber Informasi Mahasiswa Saat mendaftar

c. Data Preparation

After going through the data understanding process, the next process is to prepare data preparation which will then be used as data analysis in modeling, where it takes some data process preparation, namely by combining data. The process of combining data carried out is student data obtained from the information system section consisting of 2 tables, namely student data for the academic year 2020-2021 and 2021-2022. After merging data, the next phase is data cleaning so that the data used is of higher quality and ready for analysis. This data will be prepared according to

the needs of the analysis using predetermined attributes, namely regional attributes, school origin, cost source, class, parents' occupation, year of entry, and source of enrollment information.

d. Modeling

In this study, the method used was the K-Means algorithm as a method of data analysis. The process of analyzing data with the K-Means algorithm requires several stages, namely:

1. Determine the number of k centroids According to the results of the cluster test, it is determined that the number of k to be used is as many as 4.
2. Select the data as centroid. Determining the centroid is obtained using the formula (C = Number of Data / C + 1)
3. Allocate all data to the closest centroid with a predefined distance metric.
4. Recalculate centroid C based on the data that follows each cluster.
5. Repeat steps 3 and 4 until the cluster condition does not change.

After going through several data preparation processes, 825 data were obtained to be used as data analysis with 7 attributes as data to be processed in K-Means.

Before the modeling stage begins, the value of k is first determined. Based on the results of the exploration test of the number of k values with the DBI technique on rapid miners, the following results were obtained:

Figure 4. The test result determines the number of clusters

Based on the k value test diagram, it can be seen that the DBI score at the time of k = 4 occurs initial intersection and is in accordance with the cluster k value test technique. After determining the next value, a modeling process is carried out with the number k = 4. In the modeling process, the results of student data clustering ended in the 14th iteration because there were no changes to the cluster grouping on the data. The cluster group results can be seen in table 2.

Table 2. Hasil pengelompokan cluster

Data ke-	wilayah	sekolah	biaya	kelas	pekerjaan	Thnmasuk	info	C1	C2	C3	C4	Jarak Data Terdekat antar Cluster	Kelompok Klaster
1	2	2	2	2	13	1	1	6	5	2	4	2	Cluster 3
2	1	1	2	1	13	1	3	5	5	2	3	2	Cluster 3
3	2	2	1	4	13	1	5	2	4	3	1	1	Cluster 4
4	2	1	2	1	13	1	3	5	5	2	3	2	Cluster 3
5	1	1	1	3	13	1	4	3	4	2	1	1	Cluster 4
6	1	1	3	4	13	1	3	4	4	2	3	2	Cluster 3
7	3	1	2	3	13	1	5	3	5	3	2	2	Cluster 4
8	3	1	2	3	13	1	5	3	5	3	2	2	Cluster 4
9	1	4	2	1	13	1	3	5	5	3	4	3	Cluster 3
10	2	2	2	1	9	1	3	6	2	5	5	2	Cluster 2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
816	2	1	2	2	13	2	3	1	4	2	1	2	Cluster 2
817	1	1	1	4	10	2	7	6	6	6	5	3	Cluster 1
818	1	3	3	4	13	2	1	4	3	4	3	3	Cluster 3
819	2	2	1	4	13	2	1	3	3	4	3	2	Cluster 3
820	1	2	2	3	12	2	3	2	3	3	2	2	Cluster 2
821	1	1	1	4	13	2	5	3	4	4	3	2	Cluster 4
822	2	1	1	4	13	2	2	3	3	4	3	2	Cluster 2
823	2	1	1	4	13	2	1	3	4	4	3	2	Cluster 2
824	1	3	2	2	13	2	1	3	3	3	3	3	Cluster 2
825	1	3	1	4	10	2	3	4	3	5	3	2	Cluster 3

e. Evaluation

After the modeling stage is complete and gets the cluster results, the calculation results from modeling are obtained the final centroid value from the student dataset. After knowing the final centroid value of each attribute in each cluster, the next stage is the testing stage of the resulting cluster to determine the quality of the cluster by calculating the data distance of each record. Centroid attributes of each data are obtained from centroid data on each cluster based on cluster grouping. The data distance is obtained from the difference between the attribute value and the centroid attribute in each data. An overview of the results of this stage can be in table 3.

Table 3. Data distance between records

Data ke-	cen_ wilayah	cen_ sekolah	cen_ biaya	cen_ kelas	cen_ pekerjaan	cen_ thn_masuk	cen_ info	Jarak Data
1	1,56	1,67	1,66	3,21	12,95	0,73	2,52	2,07
2	1,56	1,67	1,66	3,21	12,95	0,73	2,52	2,47
3	1,62	1,65	1,63	3,16	12,92	0,46	4,76	1,31
4	1,56	1,67	1,66	3,21	12,95	0,73	2,52	2,44
5	1,62	1,65	1,63	3,16	12,92	0,46	4,76	1,45
6	1,56	1,67	1,66	3,21	12,95	0,73	2,52	1,87
7	1,62	1,65	1,63	3,16	12,92	0,46	4,76	1,69
8	1,62	1,65	1,63	3,16	12,92	0,46	4,76	1,69
9	1,56	1,67	1,66	3,21	12,95	0,73	2,52	3,33
10	1,52	1,58	1,78	3,17	9,09	1,60	3,55	2,41
...	...	...	...	...	...	...	...	...
816	1,56	1,67	1,66	3,21	12,95	0,73	2,52	2,01
817	1,54	1,54	1,63	3,21	12,68	0,45	6,89	3,35
818	1,56	1,67	1,66	3,21	12,95	0,73	2,52	2,90
819	1,56	1,67	1,66	3,21	12,95	0,73	2,52	2,30
820	1,56	1,67	1,66	3,21	12,95	0,73	2,52	1,82
821	1,62	1,65	1,63	3,16	12,92	0,46	4,76	2,09
822	1,56	1,67	1,66	3,21	12,95	0,73	2,52	1,89
823	1,56	1,67	1,66	3,21	12,95	0,73	2,52	2,37
824	1,56	1,67	1,66	3,21	12,95	0,73	2,52	2,76
825	1,52	1,58	1,78	3,17	9,09	1,60	3,55	2,21

After obtaining the distance of data in each cluster, then the testing stage uses the Davies Bouldin Index (DBI) technique, where this technique will measure the level of homogeneity between members in a cluster called the Sum of Square Within (SSW) and the distance between clusters is called the Sum of Square Between (SSB). After obtaining SSW and SSB values, the ratio between clusters is then determined to calculate the DBI value.

SSW is generated from the average value of data distance in each cluster listed in table 4.18. In this process, SSW is calculated from the 1st SSW (SSW1) to the 4th SSW (SSW4) because it adjusts to the number of clusters at the modeling stage. SSW1 is calculated based on the distance of grouped data in cluster 1 as many as 116, SSW2 is calculated based on the distance of grouped data in cluster 2 as many as 223, SSW3 is calculated based on the distance of grouped data in cluster 3 as many as 294, and SSW4 is calculated based on the distance of grouped data in cluster 4 as many as 192.

SSW1 = Number of data distances on cluster 1/ Number of data clusters 1
SSW1 = 211,21 / 116
SSW1 = 1,84
SSW2 = Number of data distances on cluster 2 / Number of data clusters 2
SSW2 = 583,16 / 223
SSW2 = 2,63
SSW3 = Number of data distances on cluster 3/ Number of data clusters 3
SSW3 = 548,98 / 294
SSW3 = 1,87
SSW4 = Number of data distances on cluster 4/ Number of data clusters 4
SSW4 = 354,22 / 192
SSW4 = 1,85

The DBI value is obtained by dividing the number of n clusters multiplied by the largest ratio value. Below is the process of calculating the DBI value for the cluster result test.

DBI = (1/n cluster)* R14
DBI = (1/4) * 1,71
DBI = 0,43

The calculation result of DBI = 0.43 shows a positive value below 0.5 and close to 0, thus it can be stated that the cluster produced at the modeling stage meets the criteria for a good cluster and can proceed to the next stage.

f. Deployment

The results of the cluster performance test show that the cluster that has been produced has good quality, for that the next process is to determine the characteristics of each cluster member and can be seen in table 4.

Tabel 4. Hasil karakteristik data mahasiswa pada cluster

Atribut	Value	Jumlah	Strategi Promosi
Wilayah	Kota Bogor	61	Utamakan promosi pada wilayah Kota dan Kabupaten Bogor.
	Kabupaten Bogor	47	
	Luar Bogor	8	
Asal Sekolah	SMK	77	Fokuskan promosi dan tingkatkan kerjasama dengan SMK di wilayah Kota Bogor
	SMA	25	
	MA	4	
	UMUM	10	
Sumber Biaya	Orang Tua	52	Informasikan adanya program beasiswa
	Sendiri	55	
	Beasiswa	9	
Kelas Perkuliahan	Karyawan	8	Sosialisasikan adanya kelas Karyawan dan Lanjutan.
	Lanjutan	11	
	Reguler Sore	46	
	Reguler Pagi	51	
Pekerjaan Orang Tua	IRT	8	Promosikan adanya program beasiswa kuliah di UNBIN.
	Wiraswasta	13	
	Lainnya	95	
Tahun Masuk	2020-2021	90	Upayakan pendekatan dan berikan reward pada mahasiswa angkatan 2020-2021 yang berhasil mendaftarkan mahasiswa baru.
	2021-2022	26	
Sumber Informasi	Guru	13	Upayakan temukan terobosan dan trik promosi baru agar dapat menambah minat pendaftar.
	Lainnya	103	

Conclusion

Based on problem identification, problem formulation, research objectives and hypotheses it can be concluded that:

1. The application of the K-Means algorithm can determine the most optimal cluster with the results of 4 clusters because based on the test results from clusters that have been produced with the DBI technique produces a positive value of 0.4. Thus, the cluster results can be said to be good because the DBI value is positive and below 0.5.
2. Development of data mining prototypes is needed to determine promotional strategies

Reference

Abriyanto, A., & Damastuti, N. (2019). Segmentasi Mahasiswa Dengan Unsupervised Algoritma Guna Membangun Strategi Marketing Penerimaan Mahasiswa. *Insand Comtech: Information Science and Computer Technology Journal*, *4*(2), 10–18.

Anggreini, N. L. (2019). Teknik Clustering Dengan Algoritma K-Medoids Untuk Menangani Strategi Promosi Di Politeknik Tedc Bandung. *Jurnal Teknologi Informasi Dan Pendidikan*, *12*(2), 1–7. https://doi.org/10.24036/tip.v12i2.215

Handayanto, A., Latifa, K., Saputro, N. D., & Waliansyah, R. R. (2019). Analisis dan Penerapan Algoritma Support Vector Machine (SVM) dalam Data Mining untuk Menunjang Strategi Promosi. *JUITA : Jurnal Informatika*, *7*(2), 71. https://doi.org/10.30595/juita.v7i2.4378

Kristanto, T., Hadiansyah, W. M., Nasrullah, M., Amalia, A., Anggraini, E. Y., & Firmansyah, A. (2020). Strategi Pemasaran Digital Dalam Peningkatan Penerimaan Mahasiswa Baru Menggunakan Analisis SWOT. *Multinetics*, *6*(2), 128–133. https://doi.org/10.32722/multinetics.v6i2.3415

Nur, F., Zarlis, M., & Nasution, B. B. (2017). Penerapan Algoritma K-Means Pada Siswa Baru Sekolahmenengah Kejuruan Untuk Clustering Jurusan. *InfoTekJar (Jurnal Nasional Informatika Dan Teknologi Jaringan)*, *1*(2), 100–105. https://doi.org/10.30743/infotekjar.v1i2.70

Silalahi, N. (2020). *TIN : Terapan Informatika Nusantara Penentuan Strategi Promosi Universitas Budi Darma Menggunakan Algoritma K-Means Clustering TIN : Terapan Informatika Nusantara*. *1*(1), 40–46.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Sopandi, E. (2011). Tinjauan atas kegiatan Promosi PTS di Jawa Barat Tahun 2010. *Manajerial*, *9*(April), 87–101.

Sugiarti, S., Nahulae, D. K., Panggabean, T. E., & Sianturi, M. (2018). Sistem Pendukung Keputusan Penentuan Kebijakan Strategi Promosi Kampus Dengan Metode Weighted Aggregated Sum Product Assesment (WASPAS). *JURIKOM (Jurnal Riset Komputer)*, *5*(2), 103–108. Retrieved from http://ejurnal.stmik-budidarma.ac.id/index.php/jurikom%7CPage%7C103

Suhanda, Y., Kurniati, I., & Norma, S. (2020). Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik. *Jurnal Teknologi Informatika Dan Komputer MH Thamrin*, *6*(2), 12–20.

Yunita, F. (2018). Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru. *Sistemasi*, *7*(3), 238. https://doi.org/10.32520/stmsi.v7i3.388