NEW STUDENT CLUSTERIZATION BASED ON NEW STUDENT ADMISSION USING DATA MINING METHOD

Anita Diana[1*)], Atik Ariesta[2], Arief Wibowo[3], Diva Ajeng Brillian Risaychi[4]

[1,4] Sistem Informasi, [2] Manajemen Informatika, [3] Teknik Informatika
Universitas Budi Luhur
Jakarta, Indonesia
https://www.budiluhur.ac.id/
*[1]anita.diana@budiluhur.ac.id, [2]atik.ariesta@budiluhur.ac.id, [3]arief.wibowo@budiluhur.ac.id,
[4]divarisaychi@gmail.com

(*) Corresponding Author

Abstract*—The process of admitting new students to the Faculty of Information Technology (FTI) at Universitas Budi Luhur produces a large amount of student data in the form of student profile data and other data. This happens, causing a buildup of new student data, thus affecting the search for information on that data. This study aims to classify regular undergraduate admissions data at the Faculty of Information Technology (FTI) Universitas Budi Luhur by utilizing the data mining process using the clustering technique. The algorithm used for Clustering is the K-Means algorithm. K-Means is a non-hierarchical clustering data method that can group student data into several clusters based on the similarity of the data so that student data with the same characteristics are grouped in one cluster, and those with different characteristics are grouped in another cluster. An implementation using RapidMiner is used to help find accurate values. This research described the clusters from data on regular undergraduate admissions at the Faculty of Information Technology (FTI) at Universitas Budi Luhur. This will help recommend decision-making to determine the marketing promotion strategy for each study program at Universitas Budi Luhur. Based on the K-Means algorithm cluster results, the majors or study programs of interest in each school from which new students come can also be seen.*

Keywords: *K-Means Algorithm, Data Mining, New Student Clustering.*

Abstrak—*Proses penerimaan mahasiswa baru Fakultas Teknologi Informasi (FTI) Universitas Budi Luhur menghasilkan data mahasiswa yang sangat banyak berupa data profil mahasiswa dan data lainnya. Hal tersebut terjadi menimbulkan penumpukan data mahasiswa baru, sehingga mempengaruhi pencarian informasi terhadap data tersebut. Penelitian ini bertujuan untuk melakukan pengelompokan terhadap data penerimaan mahasiswa baru Strata-1 reguler di Fakultas Teknologi Informasi (FTI) Universitas Budi Luhur dengan memanfaatkan proses data mining dengan menggunakan teknik Clustering. Algoritma yang digunakan untuk pembentukan Clustering adalah algoritma K-Means. K-Means merupakan salah satu metode data non-hierarchical clustering yang dapat mengelompokkan data mahasiswa ke dalam beberapa cluster berdasarkan kemiripan dari data tersebut, sehingga data mahasiswa yang memiliki karakteristik yang sama dikelompokkan dalam satu cluster dan yang memiliki karakteristik yang berbeda dikelompokkan dalam cluster yang lain. Implementasi menggunakan RapidMiner digunakan untuk membantu menemukan nilai yang akurat. Penelitian ini menghasilkan deskripsi klaster apa saja yang terbentuk dari dapat data penerimaan mahasiswa baru Strata-1 reguler di Fakultas Teknologi Informasi (FTI) Universitas Budi Luhur. Hal ini akan membantu merekomendasikan pengambilan keputusan untuk menentukan strategi promosi pemasaran masing-masing program studi yang ada di Universitas Budi Luhur. Berdasarkan hasil cluster algoritma K-Means ini pula, dapat dilihat jurusan atau program studi yang diminati di masing-masing sekolah asal mahasiswa baru.*

Kata Kunci*: Algoritma K-Means, Data Mining, Klasterisasi Mahasiswa baru.*

INTRODUCTION

The application of information technology in education can bring about change, especially in producing abundant data about students and the resulting learning process. At tertiary educational institutions, data can be obtained based on historical data so that the data will increase continuously, for example, student data, especially new student data. Admitting new students to a tertiary institution produces abundant data in the form of profiles of these new students. This will repeatedly happen in a college. Continuous accumulation of student data will slow the search for information on that data. Based on the abundance of student data, confidential information can be found by processing the data so that it is helpful for the university.

Student data processing must be done to find important information in the form of new knowledge (knowledge discovery). Data mining is looking for patterns or interesting information in selected data using specific techniques, methods, or algorithms. The proper method or algorithm selection depends on the goals and process of Knowledge Discovery in the Database (KDD). One method in data mining used in this study is Clustering, which identifies objects with similarities.

Cluster formation is one technique used in extracting a data's trend pattern. Data mining is usually synonymous with extracting data that is quite large and grouped into neatly arranged data. The Clustering that is applied uses the K-Means Clustering algorithm. The K-Means Clustering algorithm can classify data in the same group and different data in different groups. The K-Means Clustering algorithm itself is a non-hierarchical data clustering method that groups data in the form of one or more clusters or groups. Data with the same characteristics are grouped in one cluster/group, and data with different characteristics are grouped with other clusters or groups so that the data in one cluster or group has a small degree of variation. So that it will be seen that the data group of regular undergraduate freshmen based on their profile at the Faculty of Information Technology (FTI) at Universitas Budi Luhur is not structured to be structured.

The problem that will be examined and analyzed in this study is the absence of a regular Strata-1 new student clustering model based on new student admissions using the data mining method at the Faculty of Information Technology, Universitas Budi Luhur. The limitation of the problem in this research is that the data being analyzed is data on regular fresh-man undergraduates at the Faculty of Information Technology (FTI) Universitas Budi Luhur for 2019, 2020, and 2021 classes.

According to their profile, this research aims to apply the K-means clustering algorithm to regular undergraduate admissions data at the Faculty of Information Technology (FTI) Universitas Budi Luhur. The results of grouping data using a clustering algorithm are expected to assist decision-making in determining the marketing promotion strategy for each study program at Universitas Budi Luhur.

In a previous book (Santrock 2007), new students are a transition from mid-adolescence (middle) to late adolescence or a new status worn by late adolescents in their first year of college. Tertiary institutions routinely carry out New Student Admissions (PMB) at every new school opening. In practice, PMB has many selection paths for registration, depending on the policies of each tertiary institution (Nasir 2018)

Previous publications stated that The K-Means algorithm is a method commonly used to solve clustering problems such as pattern recognition, partitioning, and taxonomic grouping problems in plants. The K-Means algorithm relies on randomly selecting the initial cluster center. It can affect the clustering results because the initial center of the cluster changes in each simulation. (Khairati et al. 2019)

In previous publications, it was written that Clustering is a method used in data mining by finding and grouping data that has similar characteristics between the data and other data collected. One of the practical and fast clustering methods used is the k-means method which aims to create clusters of objects in k partitions based on attributes. (Anindya Khrisna Wardhani 2016)

In other publications, it is written that Clustering is grouping several data sets, observations, or other examples of a category that are distinguished from each other because of the similarity of objects. K-Means is an algorithm that can group data based on the shortest distance between data centers in a cluster. Based on the experimental results, K-Means show better evaluation results than K-Medoids when processing small data sets. (Ramadhani dan Januarita AK. 2017)

In other publications, it is written that Clustering is one of the techniques used in data mining. Defining clusters in data mining is a technique for grouping data into specific clusters, allowing data in clusters to have similarities and differences that differ from data in other clusters. (Sucipto 2019)

In other publications, it is stated that K-Means clustering is used to generate data grouping of new students so that they can find out the regional pattern of new students, which can be used to determine the right promotion strategy. Researchers transform school data into SMA, SMK, MA, and others. The student's address is also transformed to produce the area of origin. Data conversion is also carried out, such as gender, school origin, student address, and study program. (Rahmalinda dan Jananto 2022)

In other publications, it is written that this study aims to classify student data using the data mining method with the clustering method. The algorithm used is the K-means clustering algorithm. The K-Means algorithm is an iterative clustering algorithm that divides the data set into k clusters defined initially. The Fast Miner 5.3 implementation is used to find the exact value. (Damanik dan Sigiro 2021)

Other publications stated that Another research used K-Means to group new students using school origin, chosen study program/significance, and UAN score. The variables used are then transformed as preparation for Clustering. The research produced 3 clusters that can be used as a basis for determining each study program's promotion strategy. (Yunita 2018)

In other studies, it is stated that K-Means is also used for data classification of new student admissions. Data processing and classification process using rapid Miner. Before Clustering, the data is transformed first. The transformed data includes the school's name, area, and gender. In the initial clustering process, what is done is to determine the center point (centroid) first. The study's clustering process continued after the centroid was obtained, producing 3 clusters. (Udariansyah dan Ibrahim 2022)

In other research, it is stated that The large number of new students at the university means more data enters the database server—the more data that comes in, the more data accumulates. Data mining is collecting data relating to new students' acceptance. Data mining techniques can process a mass of information into meaningful information. The technique used to classify prospective new student data is a clustering technique. The results of this study become a reference for the university in implementing strategies for prospective new students (Asroni, Fitri, dan Prasetyo 2018)

In other research, it is stated that The final results of this study show that the K-Means algorithm can group provinces into three clusters with specifications based on the level of national fuel consumption. The grouping of Indonesian provinces according to fuel economy can be solved using the K-Means algorithm. The K-Means algorithm can collect large amounts of data, but it is not efficient enough to classify accurately because the determination of the center (midpoint) in the early stages of the K-Means algorithm significantly affects the cluster results as a test result. Done with different centroids also produce different clustering results. (Mahartika dan Wibowo 2019)

In other research, it is stated that This research is descriptive quantitative. Data were obtained by clustering data with data mining techniques using the K-Means algorithm. The grouping of BPJS patient data from the data mining process aims to generate new information about the grouping of BPJS patient data in the Sidoarjo region. It can be a reference for hospitals in disease prevention issues in areas where patients suffer from serious illnesses. (Ali dan Masyfufah 2021)

In other publications, it is stated that the collection of first-year students in the desired course choices is carried out to determine the extent to which students have increased and decreased in these individual courses. Its purpose is to serve as a decision aid for administrators when determining strategies for increasing student numbers in the future. The K-Means algorithm is a data mining algorithm that can be used for data clustering. By grouping the data, it can be seen that the number of students in each program has increased or decreased. These results can be used as management evaluation material for adding new students to any program, so programs with disproportionate faculty-student ratios are suitable. (Nopriandi dan Haswan 2022)

From the results of other research that has been done, the authors are interested in researching clustering new students at the Faculty of Information Technology (FTI) Universitas Budi Luhur, which is expected to help recommend decision-making to determine marketing promotion strategies for each study program at Universitas Budi Luhur.

MATERIALS AND METHODS

The primary method of analyzing the K-Means algorithm is as follows (Wardhani 2016)

a. Determine the number of clusters (k) and randomly set the cluster center.
b. Calculate the distance of each data to the cluster center.
c. Group data into clusters with the shortest distance.
d. Compute the new cluster center.
e. Repeat steps A to D until no more data moves to another cluster.

The clustering process begins by identifying the data to be clustered, Xij (i=1,…,n; j=1,…,m), where n is the amount of data to be clustered, and m is the number of variables. At the beginning of the iteration, the center of each cluster is determined randomly, Ckj (k=1,…,k; j=1,…,m), and then the distance between each data and each cluster center is calculated. To calculate the distance of the Ith data (xi) at the kth cluster center (ck), given the name (dik), the Euclidean Distance formula can be used as shown in equation (1), as follows:

$$D_{ij} = \sqrt{\sum_{k=1}^{m} X_{ij} - C_{jk}^{2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where:

Dij : object distance between data values and cluster center values
m : number of data dimensions
Xij : data value from the k-th dimension
Cjk : cluster center value of the kth dimension

To calculate the new centroid, you can use equation (2) as follows:

$$C = \frac{\sum m}{n} \text{ ......} (2)$$

Where:
C: data centroids
m: a data member that belongs to a certain centroid
n: the number of data that is a member of a certain centroid

Additional calculation steps, namely the ratio between BCV (Between Cluster Variation) and WCV (Within Cluster Variation) as follows:

1) Calculate BCV or equation formula (1):

C1 = (3; 2,59; 1), C2 = (1; 3,46; 2), C3 = (1; 3,02; 3)

d(C1,C2) = $\sqrt{(3-1)^2 + (2{,}59 - 3{,}46)^2 + (1-2)^2}$ = 2,40

d(C1,C3) = $\sqrt{(3-1)^2 + (2{,}59 - 3{,}02)^2 + (1-3)^2}$ = 2,86

d(C2,C3) = $\sqrt{(1-1)^2 + (3{,}46 - 3{,}02)^2 + (2-3)^2}$ = 1,09

BCV = d(C1,C2) + d(C1,C3) + d(C2,C3) = 2,40 + 2,86 + 1,09 = 6,35

Then the value of BCV (Between Cluster Variation) is 6.35

2) Calculating WCV or equation (2):

WCV = $\sum(\text{the minimum distance to the center of each cluster})^2$

To calculate the ratio can use the formula:

$$Ratio = \frac{BCV}{WCV} \text{ ......} (3)$$

The Davies Bouldin (DB) validity index calculates the average value of each point in the data set. The calculation of the value of each point is the sum of the compactness values divided by the distance between the two cluster center points as separation (Khairati et al. 2019). This approach is to maximize inter-cluster distances and minimize intra-cluster distances, which can be calculated using Equation 4 below:

$$S_i = \frac{1}{|c_i|} \sum_{x \in ci} \{|x - z_i|\} \text{ ......} (4)$$

Where c_i is the number of points that enter cluster i, x is data, and z_i is the centroid of cluster i. While the distance between clusters is defined in Equation 5 below:

$$d_{ij} = |z_i - z_j| \text{ ......} (5)$$

Where z_i centroid of cluster i and z_j centroid of cluster j. Calculation of the distance d_ij can use Euclidean. Next, we will define R_(i,qt) for the c_i cluster in Equation 6 below:

$$R_{i,qt} = \max_{j,j \neq i} \{\frac{S_{i,q} + S_{j,q}}{d_{ij,t}}\} \text{ ......} (6)$$

Furthermore, the Davies Bouldin Index is defined in Equation 7 below:

$$DB = \frac{1}{k} \sum_{i=1}^{k} Ri, qt \text{ ......} (7)$$

From this equation, k is the number of clusters used. The smaller the DBI value obtained (non-negative >= 0), the better the cluster obtained from the K-means grouping used (Ramadhani dan Januarita AK. 2017).

In the context of this research, the method used is experimental, referring to previous studies on data mining (Wati 2016). This experimental research phase consists of five steps: data collection, initial data processing, proposed method, experiment and testing, and evaluation and validation of results. In this study, the type of research conducted was in the form of experiments on primary data, namely data on regular first-year students of Strata-1 students at the Faculty of Information Technology (FTI) Universitas Budi Luhur Class of 2019, 2020, 2021.

This study also uses descriptive analysis techniques, which are carried out to analyze the data used for the results of data collection with literature studies, interviews, and observations to get the demographics of research data. The analysis technique carried out in this study uses the clustering data mining method with the K-Means algorithm. This method is used to process the data collected for solving research problems—evaluation of clustering results based on the Davies Bouldin Index value.

RESULTS AND DISCUSSION

The steps in this research are as follows:

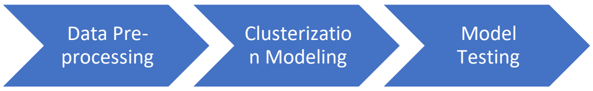


Figure 1. Research steps

The system design for the K-Means algorithm can be seen in Figure 1. In the figure, it is explained that this system's initial process is to enter data pre-processing. The data is processed into several groups or clusters using the K-Means algorithm. After the cluster rules have been formed, information is calculated using the Euclidean distance formula to calculate the distance between data to determine group members from each cluster. K-Means algorithm performance test compared to comparison algorithms such as K-Medoids, and model testing using the Davies Bouldin Index (DBI) scale.

The steps to perform data mining following KDD rules are as follows:

a. Data Selection

The research data is a selection from several database tables, including student data and student registration data at Universitas Budi Luhur. The data includes name, name, address, gender, school origin, date of birth, city of origin, study program, high school major, parents' data, and others.

b. Pre-processing / Cleaning

This step is taken to clean up duplicate data and inconsistent data. Deleting is carried out on incomplete data, contains inconsistencies, and is invalid.

c. Data Transformation

Transformations are performed to adjust the data to be processed, such as transforming the home address and the school's origin location into kilometers. The school's origin is transformed into SMA, SMK, and MA. Study program data is transformed into TI, SI, SK, KA, MI, etc.

d. Data Mining

Data mining is done to find information or patterns for Clustering using the K-Means algorithm.

e. Interpretation / Evaluation

The KDD's final step is to explain the meaning of each formed cluster. This step is done by dividing the centroid from the last iteration by the number of members of each cluster so that a description of each cluster can be concluded. In addition, a cluster evaluation was carried out using the Davies Bouldin Index method to calculate the average value of each point in the data set.

1. Data Pre-processing

The research data were obtained with assistance from the Directorate of Information Technology (DTI) at Universitas Budi Luhur for regular undergraduate student data at the Faculty of Information Technology, Universitas Budi Luhur batch 2019, 2020, and 2021, as many as 1,144 data and as many as 13 columns consisting of NIM, Name, date of birth, gender, address, city of origin, school of origin, high school major, father's occupation, mother's occupation, social studies semester 1, the study program of choice, year of entry.

At this stage, cleaning or elimination of inconsistent data is carried out. The deletion is carried out on data with incomplete attributes and is not by the conditions of the study. Incomplete data conditions in the pre-processing phase consisted of 13 data without date of birth and address values. From several incomplete data on these attributes, some data has incompleteness in several attributes at once. Then there are ten student employee data, so the year of entry will be 2022. Then there are 60 student data with the Department of Informatics Management; this data is also deleted because it does not match the research data, namely regular undergraduate students of the Faculty of Information Technology. The total number of data that was eliminated or deleted was 83 data.

Then at this stage also, data adjustments are made to inconsistent data. The data adjustment step is carried out on incomplete attribute data. Incomplete data conditions were found in the pre-processing phase, consisting of data without School Origin, High School Major, Age, City of Origin, and distance traveled values. There are 260 blank data for data from school, replaced with the highest school origin. For data values for SLTA Department, there are 668 blank data, filled with the highest central value. For the City of Origin data values that are blank, as many as 78 data are adjusted for the address.

Meanwhile, age data is obtained from the date of birth owned, and mileage data is obtained by measuring the distance from high school to Universitas Budi Luhur (in km). The total data that has been adjusted is 1,006 data.

2. Data Demografi

The research data analyzed consisted of several attributes resulting from data selection, including master data for regular undergraduate students at the Faculty of Information Technology, Universitas Budi Luhur, in batches of 2019, 2020, and 2021. The form of primary data which became material for research analysis (student data) was 1,061 data, where from 1,144 original data, then 83 data were eliminated at the pre-data processing stage. From the modeling data, it is known that there are 1,061 demographic data for regular Undergraduate students in the Faculty of Information Technology, Universitas Budi Luhur, that there are 391 students in the 2019 class, 353 students in the 2020 class, and 317 students in the 2021 class.

The following research problem-solving phase was solved using a CRISP-DM-based methodology with phases starting from Business Understanding to Deployment.

a. Business Understanding

Based on the data collected, business understanding has been identified, which is the object of this research. The clustering business process based on data from regular Undergraduate students at the Faculty of Information Technology, Universitas Budi Luhur, class of 2019, 2020, and 2021 can be seen in Figure 2.

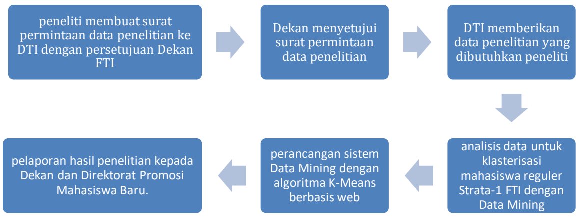
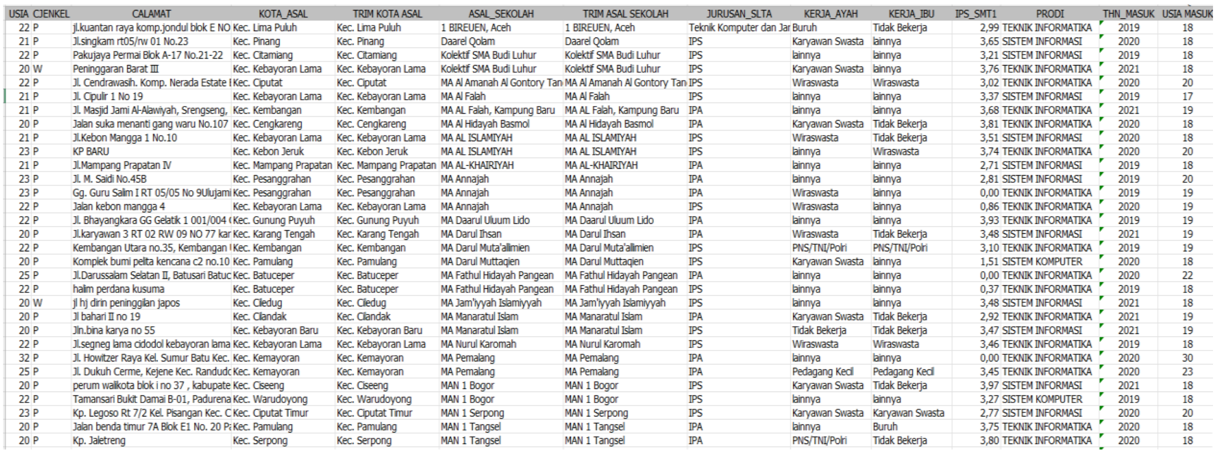

peneliti membuat surat permintaan data penelitian ke DTI dengan persetujuan Dekan FTI
Dekan menyetujui surat permintaan data penelitian
DTI memberikan data penelitian yang dibutuhkan peneliti
analisis data untuk klasterisasi mahasiswa reguler Strata-1 FTI dengan Data Mining
perancangan sistem Data Mining dengan algoritma K-Means berbasis web
pelaporan hasil penelitian kepada Dekan dan Direktorat Promosi Mahasiswa Baru.

Figure 2. Business Process Clusterization of regular New Student Strata-1 Faculty of Information Technology Universitas Budi Luhur.

In these 1,061 data, there is a data adjustment on the attribute data that is not complete. There are 260 blank data for data from school, replaced with the highest school origin. For data values for SLTA Department, there are 668 blank data, filled with the highest central value. For the City of Origin data values that are blank, as many as 78 data are adjusted for the address. Additional age data is obtained from the date of birth owned, and additional school zoning data is obtained by looking at the distance from the high school to Univ—Budi Luhur (in km). The total data that has been adjusted is 1,006 data. The addition of age data is shown in Figure 3.



USIA	CJENKEL	CALAMAT	KOTA_ASAL	TRIM KOTA ASAL	ASAL_SEKOLAH	TRIM ASAL SEKOLAH	JURUSAN_SLTA	KERJA_AYAH	KERJA_IBU	IPS_SMT1	PRODI	THN_MASUK	USIA MASUK
22	P	jl.kuantan raya komp.jondul blok E NO	Kec. Lima Puluh	Kec. Lima Puluh	1 BIREUEN, Aceh	1 BIREUEN, Aceh	Teknik Komputer dan Jar	Buruh	Tidak Bekerja	2,99	TEKNIK INFORMATIKA	2019	18
21	P	Jl.singkam rt05/rw 01 No.23	Kec. Pinang	Kec. Pinang	Daarel Qolam	Daarel Qolam	IPS	Karyawan Swasta	lainnya	3,65	SISTEM INFORMASI	2020	18
22	P	Pakujaya Permai Blok A-17 No.21-22	Kec. Citamiang	Kec. Citamiang	Kolektif SMA Budi Luhur	Kolektif SMA Budi Luhur	IPS	lainnya	lainnya	3,21	SISTEM INFORMASI	2019	18
20	W	Peninggaran Barat III	Kec. Kebayoran Lama	Kec. Kebayoran Lama	Kolektif SMA Budi Luhur	Kolektif SMA Budi Luhur	IPS	Karyawan Swasta	lainnya	3,76	TEKNIK INFORMATIKA	2021	18
22	P	Jl. Cendrawasih. Komp. Nerada Estate	Kec. Ciputat	Kec. Ciputat	MA Al Amanah Al Gontory Tan	MA Al Amanah Al Gontory Tan	IPS	Wiraswasta	Wiraswasta	3,02	TEKNIK INFORMATIKA	2020	20
21	P	Jl. Cipulir 1 No 19	Kec. Kebayoran Lama	Kec. Kebayoran Lama	MA Al Falah	MA Al Falah	IPS	lainnya	lainnya	3,37	SISTEM INFORMASI	2019	17
21	P	Jl. Masjid Jami Al-Alawiyah, Srengseng,	Kec. Kembangan	Kec. Kembangan	MA AL Falah, Kampung Baru	MA AL Falah, Kampung Baru	IPA	lainnya	lainnya	3,68	TEKNIK INFORMATIKA	2021	19
20	P	Jalan suka menanti gang waru No.107	Kec. Cengkareng	Kec. Cengkareng	MA Al Hidayah Basmol	MA Al Hidayah Basmol	IPA	Karyawan Swasta	Tidak Bekerja	3,81	TEKNIK INFORMATIKA	2020	18
21	P	Jl.Kebon Mangga 1 No.10	Kec. Kebayoran Lama	Kec. Kebayoran Lama	MA AL ISLAMIYAH	MA AL ISLAMIYAH	IPS	Wiraswasta	Tidak Bekerja	3,51	SISTEM INFORMASI	2020	18
23	P	KP BARU	Kec. Kebon Jeruk	Kec. Kebon Jeruk	MA AL ISLAMIYAH	MA AL ISLAMIYAH	IPS	lainnya	Wiraswasta	3,74	TEKNIK INFORMATIKA	2020	20
21	P	Jl.Mampang Prapatan IV	Kec. Mampang Prapatan	Kec. Mampang Prapatan	MA AL-KHAIRIYAH	MA AL-KHAIRIYAH	IPA	lainnya	lainnya	2,71	SISTEM INFORMASI	2019	18
23	P	Jl. M. Saidi No.45B	Kec. Pesanggrahan	Kec. Pesanggrahan	MA Annajah	MA Annajah	IPA	lainnya	lainnya	2,81	SISTEM INFORMASI	2019	20
23	P	Gg. Guru Salim I RT 05/05 No 9Ulujami	Kec. Pesanggrahan	Kec. Pesanggrahan	MA Annajah	MA Annajah	IPA	Wiraswasta	lainnya	0,00	TEKNIK INFORMATIKA	2019	19
22	P	Jalan kebon mangga 4	Kec. Kebayoran Lama	Kec. Kebayoran Lama	MA Annajah	MA Annajah	IPS	Wiraswasta	lainnya	0,86	TEKNIK INFORMATIKA	2020	19
22	P	Jl. Bhayangkara GG Gelatik 1 001/004 (	Kec. Gunung Puyuh	Kec. Gunung Puyuh	MA Daarul Uluum Lido	MA Daarul Uluum Lido	IPA	lainnya	lainnya	3,93	TEKNIK INFORMATIKA	2019	19
20	P	Jl.karyawan 3 RT 02 RW 09 NO 77 kar	Kec. Karang Tengah	Kec. Karang Tengah	MA Darul Ihsan	MA Darul Ihsan	IPA	Wiraswasta	Tidak Bekerja	3,48	SISTEM INFORMASI	2021	19
22	P	Kembangan Utara no.35, Kembangan	Kec. Kembangan	Kec. Kembangan	MA Darul Muta'allimien	MA Darul Muta'allimien	IPS	PNS/TNI/Polri	PNS/TNI/Polri	3,10	TEKNIK INFORMATIKA	2019	19
20	P	Komplek bumi pelita kencana c2 no.10	Kec. Pamulang	Kec. Pamulang	MA Darul Muttaqien	MA Darul Muttaqien	IPS	Karyawan Swasta	lainnya	1,51	SISTEM KOMPUTER	2020	18
25	P	Jl.Darussalam Selatan II, Batusari Batuc	Kec. Batuceper	Kec. Batuceper	MA Fathul Hidayah Pangean	MA Fathul Hidayah Pangean	IPA	lainnya	lainnya	0,00	TEKNIK INFORMATIKA	2020	22
22	P	halim perdana kusuma	Kec. Batuceper	Kec. Batuceper	MA Fathul Hidayah Pangean	MA Fathul Hidayah Pangean	IPS	lainnya	lainnya	0,37	TEKNIK INFORMATIKA	2019	18
20	W	jl hj dirin peninggilan japos	Kec. Ciledug	Kec. Ciledug	MA Jam'iyyah Islamiyyah	MA Jam'iyyah Islamiyyah	IPS	lainnya	lainnya	3,48	SISTEM INFORMASI	2021	18
20	P	Jl bahari II no 19	Kec. Cilandak	Kec. Cilandak	MA Manaratul Islam	MA Manaratul Islam	IPA	Karyawan Swasta	Tidak Bekerja	2,92	TEKNIK INFORMATIKA	2021	19
20	P	Jln.bina karya no 55	Kec. Kebayoran Baru	Kec. Kebayoran Baru	MA Manaratul Islam	MA Manaratul Islam	IPS	Tidak Bekerja	Tidak Bekerja	3,47	SISTEM INFORMASI	2021	19
22	P	Jl.segneg lama cidodol kebayoran lama	Kec. Kebayoran Lama	Kec. Kebayoran Lama	MA Nurul Karomah	MA Nurul Karomah	IPS	Wiraswasta	Wiraswasta	3,46	TEKNIK INFORMATIKA	2019	18
32	P	Jl. Howitzer Raya Kel. Sumur Batu Kec.	Kec. Kemayoran	Kec. Kemayoran	MA Pemalang	MA Pemalang	IPA	lainnya	lainnya	0,00	TEKNIK INFORMATIKA	2020	30
25	P	Jl. Dukuh Cerme, Kejene Kec. Randudc	Kec. Kemayoran	Kec. Kemayoran	MA Pemalang	MA Pemalang	IPA	Pedagang Kecil	Pedagang Kecil	3,45	TEKNIK INFORMATIKA	2020	23
20	P	perum walikota blok i no 37 , kabupate	Kec. Ciseeng	Kec. Ciseeng	MAN 1 Bogor	MAN 1 Bogor	IPS	Karyawan Swasta	Tidak Bekerja	3,97	SISTEM INFORMASI	2021	18
22	P	Tamansari Bukit Damai B-01, Padurena	Kec. Warudoyong	Kec. Warudoyong	MAN 1 Bogor	MAN 1 Bogor	IPS	lainnya	lainnya	3,27	SISTEM KOMPUTER	2019	18
23	P	Kp. Legoso Rt 7/2 Kel. Pisangan Kec. C	Kec. Ciputat Timur	Kec. Ciputat Timur	MAN 1 Serpong	MAN 1 Serpong	IPS	Karyawan Swasta	Karyawan Swasta	2,77	SISTEM INFORMASI	2020	20
20	P	Jalan benda timur 7A Blok E1 No. 20 P	Kec. Pamulang	Kec. Pamulang	MAN 1 Tangsel	MAN 1 Tangsel	IPA	lainnya	Buruh	3,75	TEKNIK INFORMATIKA	2020	18
20	P	Kp. Jaletreng	Kec. Serpong	Kec. Serpong	MAN 1 Tangsel	MAN 1 Tangsel	IPA	PNS/TNI/Polri	Tidak Bekerja	3,80	TEKNIK INFORMATIKA	2020	18

Figure 3. Student data with the addition of age data

The second stage is a statistical analysis of student data resulting from pre-processing data declared complete. The descriptive statistical analysis results showed that the students who became the object of study were dominated by students who chose the Informatics Engineering study program, as many as 629 students who chose the Information Systems study program as many as 404 students. In comparison, who chose the Computer Systems study program as many as 28 students. Complete data of the student's chosen study program is shown in Table 1.

Table 1. Table of student's Choice of Study Program

No.	Study Program	Amount
1	Teknik Informatika	629
2	Sistem Informasi	404
3	Sistem Komputer	28
	TOTAL	1.061

The results of the descriptive statistical analysis also show that the highest number of students who are the object of study in the 2019 entry year, namely 391 students, the number of students in the 2020 entry year is 353 students, while the number of students in the 2021 entry year is 317 students. Complete data on the number of students by year of entry is shown in Table 2.

Table 2. Table of data Number of students by year of entry

No.	Year Of Entry	Amount
1	2019	391
2	2020	353
3	2021	317
	TOTAL	1.061

The descriptive statistical analysis results also show that the number of students who are the object of study, the highest number of senior high school majors are from science majors, namely 434 students, and the number of students from the Computer and Network Engineering department is 348 students. In comparison, the number of students from the Department IPS is as many as 109.

b. Data Understanding

At this stage, data is understood from database tables needed for clustering modeling in this study. The data consists of 1 (one) table, namely student data. Table specifications that are understood to be analyzed are shown in Table 3.

Table 3. Research Data Table

Tabel: mahasiswa

No.	Field	Description
1.	NIM	Student ID Number
2.	Name	Student name
3.	Date of birth	Student's date of birth
4.	Gender	Student Gender
5.	Address	Student home address
6.	Hometown	The city of origin is the student's home district
7.	Which school are you from	Name of the student's school of origin
8.	High School Department	Student majors during high school
9.	Father's occupation	Student's father's job
10.	Mother's job	Occupation of the mother of the student
11.	Semester 1 IPS	Grade Point Average Semester 1 student
12.	Preferred study program	The name of the student's chosen study program when registering
13.	Entry Year	Student entry year
14.	entry age	Age of student when enrolled
15.	mileage	distance from high school to Univ. Budi Luhur (in km)

Data selection is made according to the rules to form a relational database suitable for obtaining primary research data.

c. Data Preparation

By clustering data mining, a transformation process is needed on the data, namely to obtain data from high school, origin from high school majors, father's and mother's occupations, chosen study program when registering, and School Zoning. Transformations are performed to adjust the data so that it can be processed. This transformation uses values with rules as shown in, among others, Table 4, table 5, table 6, table 7, and Table 8.

Table 4. Data Transformation from High School Majors

No	High School Majors	Transformation
1	SAINTEK	1
2	SOSHUM	2

Table 5. Transformation of Father's and Mother's Occupational Data

No	Father's and Mother's Occupation	Transformation
1	Does not work	0
2	Private sector employee	1
3	PNS/TNI/Polri	2
4	Self-employed	3
5	Small Traders/small traders	4
6	Other	5
7	Already dead	6
8	Laborer	7
9	retired	8
10	Farmers/breeders	9

Table 6. Elective Study Program Data Transformation

No.	Study Program	Transformation
1	Sistem Informasi	1
2	Teknik Informatika	2
3	Sistem Komputer	3

Table 7. School Origin Data Transformation

No.	Program Studi	Transformation
1	SMK	1
2	SMA	2
3	Sederajat	3

Table 8. School Zoning Data Transformation

No	School Zoning	Transformation
1	0 KM – 5 KM	1
2	6 KM – 10 KM	2
3	11 KM – 15 KM	3
4	16 KM – 20 KM	4
5	>21KM	5

3. Pemodelan Klasterisasi dan Evaluasi

The clustering modeling stage in this study was carried out using the Rapidminer Studio 9.0.0 software. The modeling design of the Rapidminer Study is shown in Figure 4.

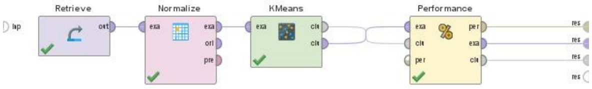


Figure 4. Clustering Modeling Design in Rapidminer Studio

From the modeling design, as shown in Figure 4, it is known that the modeling process begins by inputting data on RapidMiner with the Retrieve feature, data for new students. The K-Means method is used for modeling, and the Davies-Bouldin Index (DBI) is used to benchmark the validity of the clusters formed. The modeling process uses experiments that divide the data from 2 (two) to 10 (ten) clusters. In the evaluation feature in RapidMiner Studio, normalizing and maximizing settings are made to ensure that the evaluation process is carried out on the best model. The modeling results are shown in Table 9.

Table 9. Clusterization modeling results

No.	Cluster	DBI value
1.	2	2,433
2.	3	2,057
3.	4	1,898
4.	5	1,743
5.	**6**	**1,597**
6.	7	1,604
7.	8	1,695
8.	9	1,732
9.	10	1,625

In Table 9, it can be seen that with various experiments conducted, starting from k=2 to k=10, it turns out that the most optimal (lowest) clustering model is formed in modeling with a total of six clusters (k=6), that is, with a DBI value of 1,597. The optimal distribution of cluster members is shown in Table 10.

Table 10. Number of Members in Each Cluster for k=6

Cluster	Number of Members
I	35
II	40
III	114
IV	395
V	116
VI	331
Total	**1061**

Table 10 shows that by modeling with 6 (six) clusters, the most significant number of cluster members is in cluster 4, namely 395. For cluster 1, as many as 35 members. For cluster 2, as many as 40 members. For cluster 3, as many as 114 members. As for cluster 4, there are 395 members. For cluster 5, as many as 116 members, and for cluster 6, as many as 331. A description of each cluster that is formed is shown in Table 11

Table 11. Profile of the Cluster formed at k=6

Cluster	Average		
	School Zoning	GPA	Entry Age
I	49,69	3,20	20,11
II	165,44	3,27	18,53
III	138,00	3,25	19,04
IV	94,16	3,31	18,58
V	89,38	3,24	18,78
VI	60,31	3,18	18,76

Based on Table 11, it is known that the specifications for cluster I are new students with most high school locations located in the DKI Jakarta buffer zone, with an age of 20-21 years when registering. In terms of academic quality, this cluster has an average semester 1 GPA above 3.20.

Cluster II is new students with high school locations mostly outside Jabodetabek, aged 18-19 years old when registering. In terms of academic quality, this cluster has an average semester 1 GPA above 3.27.

Cluster III is new students, with most high school locations outside Jabodetabek, aged 19 years old when registering. In terms of academic quality, this cluster has an average semester 1 GPA above 3.25.

Cluster IV is new students, with most high school locations in the Jakarta area, aged 18-19 years when registering. Regarding academic quality, this group has the highest average semester 1 GPA above 3.31.

Cluster V are new students, with most high school locations located in the Jakarta area, aged 18-19 years old when registering. In terms of academic quality, this cluster has an average semester 1 GPA above 3.24.

Cluster VI are new students, with most high school locations located in the Jakarta area, aged 18-19 years old when registering. In terms of academic quality, this cluster has the lowest semester 1 GPA of 3.18.

Regarding model evaluation, most of the clusters formed had the best DBI values in the 1.6 to 1.8 models with 4, 5, 6, 7, 8, 9, and 10 clusters. The relatively poor DBI values are in modeling with 1 and 2 clusters between 2.4 and 2.1. Under these conditions, the formation of clusters that are relatively ideal for new regular undergraduate students at the Faculty of Information Technology, Universitas Budi Luhur, is recommended at position 6 or 7 clusters, based on evaluation values using the Davies Bouldin Index (DBI), namely cluster 6 with DBI = 1.597, and cluster 7 = 1.604.

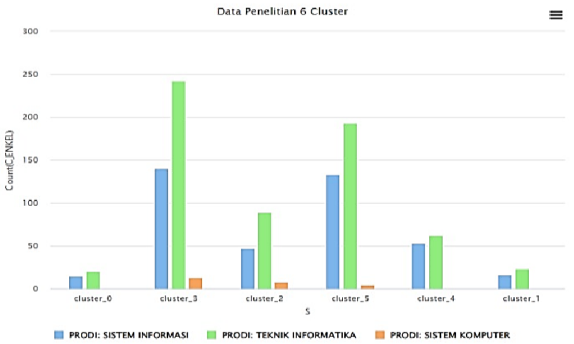


Figure 5. Number of Members of Each Cluster for the student's chosen study program

Based on Figure 5, it can be seen that the most selected study program is Informatics Engineering. The second most position is Information Systems. Moreover, the least chosen by new students is the Computer System.

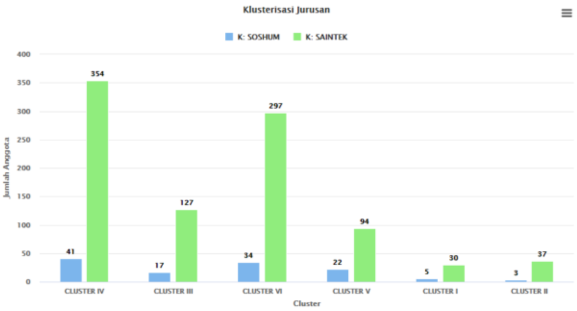


Figure 6. Number of Clusters from High School Majors for 6 clusters

Figure 6 shows that the highest school majors are in cluster 4, with 354 SAINTEK and 41 SOSHUM. The second most position is in cluster 6, with 297 SAINTEK and 34 SOSHUM. Next, it is in cluster 3, 127 SAINTEK and 127 SAINTEK. There are 17 SOSHUMs. Furthermore, in cluster 2, there are 37 SAINTEK and 3 SOSHUM. Moreover, the fewest new students are in clusters 1, 30, SAINTEK, and 5 SOSHUM.

ACKNOWLEDGEMENTS

The researcher would like to thank Universitas Budi Luhur, especially the DRPPM of Universitas Budi Luhur, for funding support for this research with a research contract number A/UBL/DRPM/61/12/22.

CONCLUSION

Several conclusions can be conveyed from all stages of the research, including student and academic data, which can be elaborated as data for modeling the Clustering of regular new undergraduate students at the Faculty of Information Technology, Universitas Budi Luhur, with a total of 1,061 data. From the research results, with various experiments conducted, from k=2 to k=10, it turns out that the most optimal (lowest) clustering model is formed in modeling with a total of six clusters (k=6), with a DBI value of 1.597. Evaluation of the Davies Bouldin Index (DBI) used in Rapidminer Studio for clustering modeling in this study produces an excellent DBI value (fit the model) in cluster 6 with DBI = 1.597. Clustering modeling performed using the K-Means algorithm runs smoothly on the RapidMiner Studio v.9 software. Cluster modeling, which is relatively ideal for new regular undergraduate students of the Faculty of Information Technology, Universitas Budi Luhur, is recommended at position 6. With modeling with 6 (six) clusters, the most significant number of cluster members is in cluster 4, namely 395 members. For cluster 1, as many as 35 members. For cluster 2, as many as 40 members. For cluster 3, as many as 114 members. As for cluster 4, there are 395 members. For cluster 5, as many as 116 members, and for cluster 6, as many as 331. From the research results, it was found that the most selected study program was Informatics Engineering. The second most position is Information Systems. Moreover, the least chosen by new students is the Computer System. Moreover, the highest school majors are from cluster 4, with 354 SAINTEK and 41 SOSHUM.

This research has not yet reached the system development stage. For further research, system development can be carried out. This research can also be developed, and the results improved in the future by adding other selection data, for example, original information about Budi Luhur University, and so on, if Clustering will be focused on segmenting new students.

REFERENCE

Ali, Amir, dan Lilis Masyfufah. 2021. “Klasterisasi Pasien BPJS Dengan Metode K-Means Clustering Guna Menunjang Program Jaminan Kesehatan Nasional Di Rumah Sakit Anwar Medika Balong Bendo Sidoarjo.” *Jurnal Wiyata* 8(1):8–22. doi: http://dx.doi.org/10.56710/wiyata.v8i1.427.

Anindya Khrisna Wardhani. 2016. “Implementasi Algoritma K-Means Untuk Pengelompokkan Penyakit Pasien Pada Puskesmas Kajen Pekalongan.” *Jurnal Transformatika* 14(1):30–37. doi: 10.26623/transformatika.v14i1.387.

Asroni, Asroni, Hidayatul Fitri, dan Eko Prasetyo. 2018. “Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokkan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus:

Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik).” *Semesta Teknika* 21(1):60–64. doi: 10.18196/st.211211.

Damanik, Nurafni, dan Mula Sigiro. 2021. “Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Pada Penerimaan Mahasiswa Baru Sebagai Metode Promosi.” *Jurnal Teknik Informatika Komputer Universal (Jutisal)* 4(2):158.

Khairati, A. F., A. A. Adlina, G. F. Hertono, dan B. D. Handari. 2019. “Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA.” Hal. 161–70 in *Prosiding Seminar Nasional Matematika (PRISMA)*. Vol. 2.

Mahartika, Indah Rizky, dan Arief Wibowo. 2019. “Data Mining Klasterisasi dengan Algoritme K-Means untuk Pengelompokkan Provinsi Berdasarkan Konsumsi Bahan Bakar Minyak Nasional.” Hal. 87–91 in *Seminar Nasional Sistem Informasi dan Teknologi (SISFOTEK)*.

Nasir, Mohamad. 2018. *Peraturan Menteri Riset, Teknologi, Dan Pendidikan Tinggi Republik Indonesia Nomor 60 Tahun 2018 Tentang Penerimaan Mahasiswa Baru Program Sarjana Pada Perguruan Tinggi Negeri*.

Nopriandi, Helpi, dan Febri Haswan. 2022. “Analisis Klasterisasi Mahasiswa Baru dalam Memilih Program Studi dengan Menggunakan Algoritma K-Means.” *Journal of Information System Research (JOSH)* 3(4):666–71. doi: 10.47065/josh.v3i4.1986.

Rahmalinda, Nanda Ayu, dan Arief Jananto. 2022. “Penerapan Metode K-Means Clustering Dalam Menentukan Strategi Promosi Berdasarkan Data Penerimaan Mahasiswa Baru.” *Jurnal Tekno Kompak* 16(2):163–75.

Ramadhani, Rima Dias, dan Dwi Januarita AK. 2017. “Evaluasi K-Means dan K-Medoids pada Dataset Kecil.” Hal. 20–24 in *Seminar Nasional Informatika dan Aplikasinya (SNIA)*.

Santrock, John W. 2007. *Psikologi Pendidikan Edisi Kedua*. Kencana Prenada Media Group.

Sucipto, Adi. 2019. “Klasterisasi Calon Mahasiswa Baru Menggunakan Algoritma K-Means.” *Jurnal Science Tech* 5(2):50–56. doi: https://doi.org/10.30738/jst.v5i2.5829.

Udariansyah, Devi, dan Deny Rahmat Ibrahim. 2022. “Klasifikasi Data Penerimaan Mahasiswa Baru Pada Universitas Bina Darma Menggunakan Algoritma K-Means Clustering.” *Jurnal Pendidikan dan Konseling* 4(4):2692–2701.

Wardhani, Anindya Khrisna. 2016. “Implementasi Algoritma K-Means Untuk Pengelompokkan Penyakit Pasien Pada Puskesmas Kajen Pekalongan.” *Jurnal Transformatika* 14:30–37.

Wati, Risa. 2016. “Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan Menggunakan Naive Bayes.” *Jurnal Evolusi* 4(BSI):25–31.

Yunita, Fitri. 2018. “Penerapan Data Mining Menggunkan Algoritma K-Means Clustring Pada Penerimaan Mahasiswa Baru (Studi Kasus: Universitas Islam Indragiri).” *Sistemasi* 7(3):238–49. doi: 10.32520/stmsi.v7i3.388.