

Application of Data Mining to Determine Promotion Strategy Using Algorithm Clustering at SMK Yadika 1

Jerry Watulangkouw^{1*)}

¹Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
email: jerrywatulangkouw@gmail.com

Abstract – The Promotion Strategy is very important to achieve the desired target, in determining the School Promotion Strategy for the results of new student admissions and in recommending the right promotion that can be used to overcome the problems faced by SMK Yadika which experienced a decrease in the number of new students from the 2017/2018 class entering 267 new student, then experiencing difficulties in determining promotion strategies, and promotion decisions taken by the school are sometimes not right on target, even though the position of SMK Yadika has a very strategic environment or place that can produce and get a lot of students. This study aims to apply the K-Means algorithm in the Promotion Strategy grouping which produces seven clusters based on the K Optimal Davies Bouldin Index so that it can be used to determine the right promotion strategy and develop an information system prototype to assist schools in compiling and deciding the right promotion. The results of this research, schools can carry out promotions based on the origin of the student's school, promotions based on the field of study of interest, promotions based on the study program expertise, promotions based on competency skills, and promotions based on the district where the student lives or domicile. With the results of clustering using the K-Means methodology, Cluster 1 (17.71%), cluster 2 (32.67%), cluster 3 (10.43%), cluster 4 (5.7%), cluster 5 (4.55%), cluster 6 (3.34%), and cluster 7 (25.78%).

Keywords – Data Mining, Promotion Strategy, Clustering, Davies Bouldin Index, CRIPS-DM

I. INTRODUCTION

Determining the right Promotion Strategy at this time can outperform the competition between schools, so promotional strategy efforts for an educational institution are needed to be able to get new students and students. Schools as educational service providers need to improve themselves and learn to have the initiative to increase customer satisfaction, in this case prospective students and students. Therefore, a promotion strategy, especially in the field of education, is needed to win a competition between schools so as to increase the interest of prospective new students or students to see the strategic position of the school that can produce a lot of students and students, and also to increase the acceleration of quality improvement and school management professionalism. Competition between vocational education institutions actually provides benefits for customers in this case prospective students and students, this is because customers have many choices in deciding which schools are suitable for their prospective students and students.

In Previous Research Asril et al (2015) with research on graduate data analysis with data mining to support the Promotion strategy of Lancang Kuningan University, the problem obtained The number of private universities and high schools that have been established requires Lancang Kuning University to carry out a series of various active promotions, so that no less competitive with other universities, in this study using the K-Means Clustering Algorithm Method with the attributes of Address, Name of

Region, and GPA while the results obtained are four clusters are formed. Rony (2016) with research on the Application of Data Mining to Use the Clustering Algorithm to Determine New Student Promotion Strategies at the LP3I Jakarta Polytechnic the problems obtained With the very large amount of data the LP3I Polytechnic has difficulty getting identification of prospective students who register at the LP3I Polytechnic Jakarta, In This study uses the K-Means Clustering Algorithm Method, with the attribute of residence, while the results obtained form 4 clusters

Promotional strategies currently running in schools tend to pick up or come to nearby junior high schools by distributing brochures, presenting school profiles, putting up banners, organizing events or competitions in junior high schools, and finally promoting through social media. Then every year Yadika 1 Vocational School has a diverse number of students, such as in the 2017/2018 semester academic year Odd there are 267 new students, in 2018/2019 Odd the number of new students is 263 students, while in 2019 semester / 2020 Odd The number of new students is 231 students. By looking at the decreasing number of students per semester, a strategy is needed to promote Yadika 1 Vocational School.

II. RESEARCH METHODOLOGY

"Data mining is the analysis of data to find clear relationships and conclude that they were not previously known in a way that is currently understood and useful for the owner of the data".

From some of the opinions above, it can be concluded that Data Mining is a technique used to explore and find a previously unknown relationship from a large and large number of data so that information is needed and can be used later.

The terms data mining and knowledge discovery in databases (KDD) are often used interchangeably to describe the process of extracting hidden information in a large database. And the stages in this data mining are like the KDD process which can be broadly explained as follows han jiawei in Sumangkut et al [1]:

1. Data Selection

Selection of new data sets of operational data needs to be done before the stage of extracting information in data mining begins. The selected data that will be used for the data mining process is stored in a file, separate from the operational database.

2. Pre-processing/Cleaning

Before the data mining process can be carried out, it is necessary to carry out a cleaning process on the data that is the focus of KDD. The cleaning process includes, among others, removing duplicate data, checking for inconsistent data, and correcting errors in data, such as typographical errors.

3. Transformation

Coding is a transformation process on the data that has been selected, so that the data is suitable for the data mining process.

4. Data Mining

Data mining is the process of looking for interesting patterns or information in selected data using certain techniques or methods.

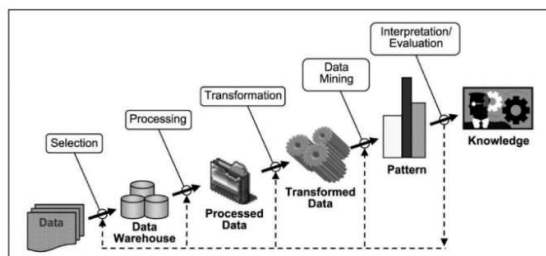
5. Interpretation/Evaluation

The pattern of information generated from the data mining process needs to be displayed in a form that is easily understood by interested parties.

6. Knowledge

The last stage of the data mining process is how to formulate decisions or actions from the results of the analysis obtained

Furthermore, Budiman [2] describes the Knowledge Discovery in Database Stages as shown in Figure 2.



According to Badrul [3] data mining is divided into several groups based on the tasks that can be done, which are as follows:

1. Description

Sometimes researchers and analysts simply want to try to find ways to describe the patterns and trends contained in the data. For example, polling officers may not be able to find information or facts that those who are not professional enough will have little support in the presidential election. Descriptions of patterns and trends often provide possible explanations for a pattern or trend.

2. Estimation

Estimation is almost the same as classification, except that the estimation target variable is more numerical than categorical. The model is built using a complete record that provides the value of the target variable as the predicted value. Furthermore, in the next review, the estimated value of the target variable is made based on the value of the predictive variable.

3. Prediction

Prediction is almost the same as classification and estimation, except that in predicting the value of the results will be in the future. Some of the methods and techniques used in classification and estimation can also be used (for appropriate circumstances) for prediction.

4. Classification

In classification, there are categorical variable targets. For example, income classification can be separated into three categories, namely high income, medium income, and low income.

5. Clustering

Clustering is grouping records, observing, or paying attention and forming classes of objects that have similarities. Cluster is a collection of records that have similarities with one another and have dissimilarities with records in other clusters.

6. Association

The task of association in data mining is to find attributes that appear at one time. In the business world it is more commonly called shopping cart analysis.

Several previous studies that have the same object of study regarding educational assistance are summarized as a study review in this paper. The research in question is:

1. This research was conducted by Asril, Wiza and Yunefri (2015) to determine the promotion strategy of Lancang Kuningan University. The full title is Graduate Data Analyst with data mining to support the promotion strategy of Lancang Kuningan University. The problem faced in this research is the number of private universities and high schools that have been established requiring Lancang Kuning University to

- carry out a series of various active promotions, so as not to lose to compete with other universities. The right promotion strategy uses the K-means Clustering Algorithm [4]
2. Research conducted by Priambudi (2015) to determine product sales strategy. The full title of this research is PT Mayora's Product Sales Strategy using the Apriori method and Data Mining Implementation. The problem in this research is how to make an application that can assist the development of schools in determining marketing strategies by utilizing transaction data. The process of implementing the right sales strategy is carried out through the Apriori Method and the implementation of data mining [5].
 3. Research conducted by Rony (2016) to determine new student promotion strategies. The full title of the research is Application of Data Mining to Use Clustering Algorithm to Determine New Student Promotion Strategy at LP3I Polytechnic Jakarta. The process of applying the right promotion strategy is done through the K-Means Clustering algorithm, starting with a random selection which is the number of clusters that you want to form. Then set the K values randomly, temporarily the value becomes the center of the cluster or commonly called the centroid [6].
 4. This research was conducted by Suprawoto (2016) to support the selection of marketing strategies. Title Classification of Student Data Using the K-Means Method to support the selection of marketing strategies. The problem faced is the amount of data that has accumulated from year to year needs to be analyzed to be able to open up opportunities to produce useful information in making alternative decisions for higher education management. The process of implementing the right promotion strategy is done through the K-Means Clustering algorithm [7].
 5. This research was conducted by Wirta and Erlin (2016) to choose a promotion strategy for new student admissions. The full title is Implementation of the K-Means Cluster Analysis Method to choose a new student admission promotion strategy. The process of applying the right promotion strategy is done through the K-Means Clustering algorithm starting with a random selection which is the number of clusters that you want to form [8].
 6. This research was conducted by Mochammad C. A. (2018) to determine the promotion strategy at the Baitussalam Intensive Vocational School Tanjunganom Nganjuk. The full title of this research is Data Mining using the K-Means algorithm to determine the promotion strategy at Baitussalam Intensive Vocational School Tanjunganom Nganjuk. The process of applying the right promotion strategy is done through the K-Means Clustering algorithm, starting with a random selection which is the number of clusters that you want to form. Then assign K values randomly [9].
 7. This research was conducted by Achyani (2018) for the optimization of direct marketing predictions. The full title is There are classification and regression problems with linear or nonlinear kernels which can be an ability of classification learning algorithms. There are classification and regression problems with linear or nonlinear kernels which can be an ability of the classification learning algorithm. The process of applying the Particle Swarm Optimization method for direct marketing prediction optimization [10].
 8. This research was conducted by Kusumo, Sedyono and Marwata (2019) to support higher education promotion strategies. The full title is Apriori Algorithm Analysis to Support Higher Education Promotion Strategy. The problem faced is that universities are currently required to have a competitive advantage by utilizing all available resources (Azimah & Sucahyo, 2007). The high level of competition between educational institutions has resulted in each institution having to be able to manage its institution professionally. The process of implementing the right promotional strategy information support is carried out through the Apriori Algorithm Method [11].
 9. This research was conducted by Jaini (2019) for grouping sales and product marketing strategies. The problems faced in conducting promotions are based on products that are experiencing a trend in society. In addition, new products that are approaching the expiration date will be promoted. It affects the effectiveness of marketing. The process of implementing the right sales support and product marketing strategy is carried out through the Fuzzy C-Means and K-Medoids Algorithm Methods [12].
 10. This research was conducted by Takdirillah (2020) on transaction data to support sales strategy information. Full Title Implementation of Data Mining Using Apriori Algorithm Against Transaction Data to Support Sales Strategy Information. Problems regarding stockpiling that can harm retail store entrepreneurs are quite common. The process of implementing the right sales strategy information support is carried out through the Apriori Algorithm Method [13].

III. RESEARCH METHODOLOGY

A. Research Methods

In this study, the methodology used is CRISP-DM to analyze and process data. The research stages are divided into several stages as shown in Figure 1.

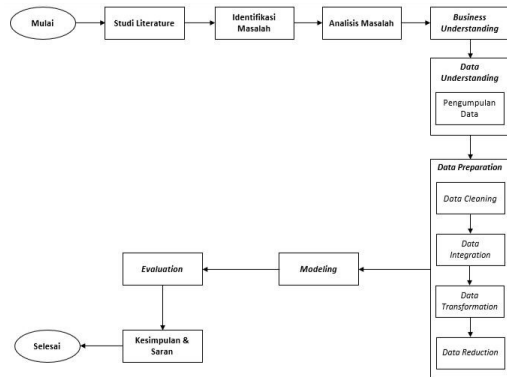


Figure 1. Research Stages

1. **Study of literature**
The author tries to dig up more information about the application of the K-Means method and related research.
2. **Identification of problems**
The author identifies problems related to the topic being researched.
3. **Problem Analysis**
Analyze problems related to research. At this stage, a sub-stage is carried out where the CRISP-DM methodology is applied. The CRISP-DM process can be explained as follows:
 - a. *Business Understanding*
At this stage it focuses on understanding the goals or objectives to be achieved in this research and the research schedule strategy.
 - b. *Data Understanding*
At this stage it focuses on understanding the data to be able to recognize what kind of data will be used for research purposes. At this stage, secondary data was collected from SMK Yadika 1 for the 2017-2019 period.
 - c. *Data Preparation*
At this stage, the data cleaning and transformation process is carried out. In this study using data cleaning and data transformation.
 - d. *Modeling*
At this stage, the data clustering process is carried out by the model and then generates a number of rules. In this study using the K-Means method.
 - e. *Evaluation*
At this stage, an evaluation of the model that has been carried out is carried out, whether the K-Means method can produce the best choice recommendation strategy cluster to be applied to the Promotion strategy.
4. **Conclusion and suggestions**

In this process, conclusions and suggestions are made on the research that has been done.

B. Technique Conclusion and suggestions

The data collection method in this study is a secondary data collection method sourced from the Yadika 1 Vocational School. Class, Year of Report, Field of Study

of Expertise, Program of Study of Expertise, Competence of expertise, Origin of School. As shown in table 1.

Table 1. Data for Yadika Vocational High School Students 1

No	Nama	Alamat	Agama	Umur	Gender	Agama	Umur	Gender	Agama	Umur	Gender	Agama	Umur	Gender	Agama	Umur	Gender
1	Adi Nugroho	Jember, Jl. Bina Bangsa	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L
2	Adi Nugroho	Jember, Jl. Bina Bangsa	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L
3	Adi Nugroho	Jember, Jl. Bina Bangsa	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L
4	Adi Nugroho	Jember, Jl. Bina Bangsa	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L
5	Adi Nugroho	Jember, Jl. Bina Bangsa	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L	Islam	17	L

1. Data Preprocessing

The data used must go through preprocessing data, because the source of the data obtained from the database is still dirty, meaning that it consists of some incomplete data, missing or empty data, lack of certain attributes or appropriate attributes, noise data, and inconsistent data. Preprocessing consists of 4 types, namely data cleaning, data integration, data transformation, and data reduction. The number of data from Yadika 1 Vocational School students from 2017-2020 was obtained as many as 1026. In this study, data preprocessing to be carried out was data cleaning to clean incomplete data (noisy), data integration to integrate data with supporting data such as master items. as item category information, data transformation to transform data into simplified criteria and are needed in research without modifying data and data reduction to eliminate incomplete data, eliminate noisy data and correct inconsistent data.

2. Analysis, Design and Testing Techniques

a. Analysis Techniques

In the analysis technique, the writer applies the K-Means method to cluster items into seven clusters. The K-Means method calculates clusters by adding a k counter with one or k=k+1 in each iteration until it meets the validity limit of the quality cluster quality.

The K-Means method is by forming clusters on student data by measuring and calculating variables as shown in table 3.1. The results formed will be evaluated using the Davies-Bouldin Index measurement. Likewise, defined by another author, “Davies-Bouldin Index is a cluster validation method based on the results of the clustering. The DBI measurement approach is to maximize the inter-cluster distance and minimize the intra-cluster distance. The smaller the DB Index value indicates the most optimal cluster scheme. The greater the purity value (closer to 1), the better the quality of the cluster” (Widiarina, 2015)..

The clustering is divided into 7 according to the Optimal K obtained in the Davies Bouldin Index (DBI) which is validated with the help of third-party tools, namely Python.

The K-Means method is calculated based on the initialization term, namely the minimum distance between centers cluster, initialization is used to measure the

separation between clusters, which can be seen in equation 3.1.

$$inter = \min \{ \|m_k - m_{kk}\| \} \quad \forall k = 1, 2, \dots, K - 1 \text{ dan } k = k + 1, \dots, K \dots \dots \dots (1)$$

Then the term intra is used to measure the cohesiveness of a group. The standard deviation is used to check the proximity of the data points of each cluster which can be seen in equation 3.2.

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - X_m)^2} \quad (2)$$

The cluster that will be analyzed produces a pattern of adjacent item sets from the results of student data analysis which will be used as item recommendations for the promotion strategy of Yadika 1 Vocational High School, West Jakarta City.

b. Design Technique

After conducting the analysis, then proceed with system design based on the problem analysis that has been carried out, namely:

- 1) The first stage is analyzing business understanding and preprocessing data using the CRISP DM method, clustering student data using the K-Means method, calculating the K-Means method testing by calculating the Centroid value and Davies Bouldin Index to get product results for strategy recommendations promotion.
- 2) The second stage is to design system interactions with actors using Use Case Diagrams and Use Case Diagrams.
- 3) The third stage is to design a database using streamlit as a library and also to upload the dataset and to download the results.
- 4) The fourth stage is designing the system interface (user interface).
- 5) The fifth stage is designing applications using the Python programming language.

c. Testing technique

After doing the design, then proceed with the system testing technique based on the research design that has been done, namely :

- 1) The technique of testing the results of clustering is by looking at the results of the Davies Bouldin Index, if the purity results are close to 1, it indicates the better the cluster formed..
- 2) The testing technique in the development of this information system is using black boc testing is a test that is carried out only observing the results of execution through test data and checking the functionality of the software..

C. Research Steps

Based on the research method, sample selection method, data collection method, analysis technique, design and testing of food data, research steps were formed which can be seen in Figure 2..

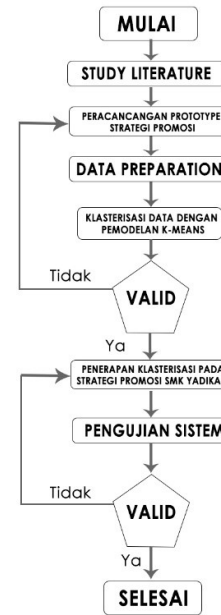


Figure 2. Research Stages

Description of Research Steps:

1. Start
2. Study Literature
This stage is a series of activities related to the method of collecting library data, reading and recording and managing research materials
3. This stage is the process of designing a display system design. Promotional strategies are made as examples to develop products as an illustration for direct userslangsung
4. Data Preparation
This stage includes all activities to build the final dataset (data to be processed at the modeling stage) from raw data. This stage can be repeated several times. This stage also includes the selection of tables, records, and data attributes, including the process of cleaning and transforming data to be used as input in the modeling stage.
5. Data Clustering with K-Means Modeling
At this stage, the selection and application of various modeling techniques will be carried out and several parameters will be adjusted to obtain optimal values.

In particular, there are several different techniques that can be applied to the same data mining problem. On the other hand, there are modeling techniques that require special data formats. So at this stage it is still possible to return to the previous stage.

6. Valid

If at this stage it is appropriate, it will proceed to the Clustering Application stage. If at this stage it is not appropriate or fails, it will return to the prototype design stage.

7. Implementation of Clustering on Promotion Strategy for Yadika 1 Vocational High School 1

At this stage of implementing clustering, it is one of the tools in data mining that aims to group objects into clusters in the promotion strategy at SMK Yadika 1.

8. System Testing

In the testing phase of this system based on the research design that has been carried out, namely the technique of testing the results of clustering by looking at the results of the Davies Bouldin Index and the testing technique in developing this information system using black box testing.

9. Valid

If at this stage it is appropriate, then everything is finished and running smoothly. If at this stage it is not appropriate or fails, it will return to the stage of implementing clustering in the promotion strategy of Yadika Vocational High School 1.

10. Finish

IV. RESULTS AND DISCUSSION

A. CRISP-DM Method

1. Business Understanding

There are management difficulties in determining the Promotion Strategy of the Yadika 1 Vocational High School in West Jakarta. So in this study, the application of data mining with the K-Means method will be carried out to cluster the promotion strategy into seven clusters according to the Optimal K obtained when searching through the Davies Bouldin Index. Of the seven clusters, Tigas will be a promotional recommendation for promotion strategies at SMK Yadika 1, West Jakarta City.

2. Data Understanding

In this study, the attributes used in this study refer to the criteria approach that has the most influence on the promotion strategy. Based on the results of observations that have been made, the authors obtained student data at SMK Yadika 1. The data obtained in this study was obtained through the Administrative section of SMK Yadika, the data obtained were student data from 2017 to 2019 The number of students in this study amounted to 1034 students with each department. From 761 student data that has been obtained, then the data is in Cleaning (Data Cleaning) then obtained a total of students who can be used in this study amounted to 508 students, Attributes before being selected amounted to 21 attributes.

3. Data Preparation

The data obtained for this study were 508 student data from 1036 student data obtained at SMK Yadika 1

West Jakarta. To get quality data, the author uses the initial data technique, data selection, data labeling, and conversion of data labeling results. This stage performs some preparation of data processing. Preparation of the data process, namely: Data Cleaning.

4. Data Integration and Data Reduction

This data selection is a data selection process by focusing on data that can be used to determine the promotion strategy of the Yadika 1 Vocational School. After getting the selection results, the attributes used are 11 attributes, including name, gender, city of birth, religion, sub-district, batch, year of the last report card, field of study of expertise, program of study of expertise, competence of expertise, and the origin of the student's school. The results of the attributes after being selected can be seen in table 2

Table 2. Attributes After Selection

No	Nama Atribut	Tipe Data	Keterangan
1.	Name	String	Nama Siswa
2.	Gender	String	Jenis kelamin siswa
3.	City of Birth	string	Kota Lahir
4.	Religion	string	Agama siswa
5.	Districts	string	Kecamatan siswa
6.	Force	int	Angkatan masuk sekolah
7.	Report Three Year	int	Tahun 3 Raport siswa
8.	Field of study Expertise	String	Bidang Studi Keahlian siswa
9.	Skill Study program	String	Program Studi Keahlian
10.	Skill Competency	String	Kompetensi keahlian
11.	Student's School Origin	String	Asal SMP Siswa

The next stage is data integration. Data integration is the process of converting or merging data into a format suitable for processing in data mining. Often the data that will be used in the data mining process has a format that cannot be directly used. Therefore, the format needs to be changed. In this study, nominal data is initialized in the form of numbers so that it can be processed using the K-means Clustering algorithm.

At this stage, data labeling is carried out on the data that has been selected. The data can be seen in Table 4.5 :

Table 3. Selected data

No	Attribute Name	Data Type
1.	Name	String
2.	Gender	String
3.	City of Birth	String
4.	Religion	String
5.	Districts	String
6.	Force	Int
7.	Report Three Year	Int
8.	Field of study Expertise	String
9.	Skill Study program	String
10.	Skill Competency	String
11.	Student's School Origin	String

The results of the data labeling are then transferred to notepad++ with .arff format. The form of the data is shown in Figure 3 :

The following are the steps for clustering using the k-means algorithm, namely::

- 1. Step 1: Determine the desired number of clusters (cluster = 7).
- 2. Step 2: choose the initial centroid randomly. In this step, 7 pieces of data will be randomly selected as centroids.
- 3. Step 3: calculate the distance with the centroid (iteration 1)

In this step, the nearest centroid of each data will be determined, and the data will be applied as a member of the group closest to the centroid.

The data used are as follows:

- Data : (1,18,4,16,2,2,1,1,1,113)
- Centroid M1 : (1,18,4,10,2,2,1,1,1,113)
- Centroid M2 : (2,60,4,7,2,2,1,1,1,112)
- Centroid M3 : (1,43,1,16,2,2,1,1,1,56)
- Centroid M4 : (2,18,2,10,2,2,1,1,1,121)
- Centroid M5 : (2,18,4,16,2,2,1,1,1,41)
- Centroid M6 : (2,18,1,10,2,2,1,2,2,75)
- Centroid M7 : (2,18,2,10,2,2,1,2,2,72)
- Data : (1,18,4,16,2,2,1,1,1,113)

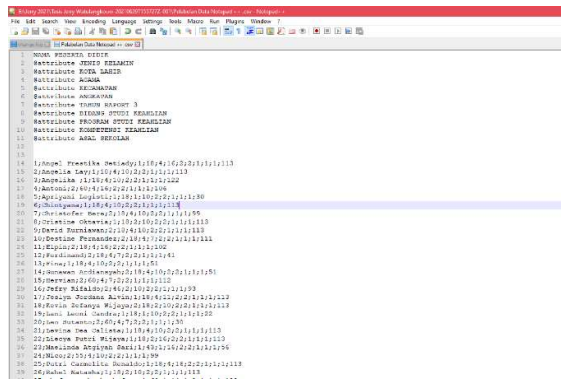


Figure 3. Conversion of Data Labeling Results

5. Data Transformation

The data used in this study is student data of SMK Yadika 1 West Jakarta City for the 2017-2020 academic year which is stored in a Microsoft Excel database by the School Administration, then data transformation is carried out by initializing data from nominal data into numbers. after the data is already in initial form

After all student data is transformed into numbers, then the data is converted into binary format in the initial database used, namely Microsoft Excel. In this study, the data that has been converted into numbers has a total of 10 attributes. The following is the initial data of students in the form of images that can be seen in Table 3 :

Table 3. Initialization of Student Data

NO	NAMA	ALAMAT	NO HP	JENIS KELAMIN	AGAMA	MATA PELAJARAN	NILAI MATEMATIKA	NILAI BAHASA INDONESIA	NILAI BAHASA INGGRIS	NILAI IPA	NILAI IPS
1	Agung Pratomo	1	18	4	16	2	2	1	1	113	
2	Agung Pratomo	1	18	4	16	2	2	1	1	113	
3	Agung Pratomo	1	18	4	16	2	2	1	1	113	
4	Agung Pratomo	1	18	4	16	2	2	1	1	113	
5	Agung Pratomo	1	18	4	16	2	2	1	1	113	
6	Agung Pratomo	1	18	4	16	2	2	1	1	113	
7	Agung Pratomo	1	18	4	16	2	2	1	1	113	
8	Agung Pratomo	1	18	4	16	2	2	1	1	113	
9	Agung Pratomo	1	18	4	16	2	2	1	1	113	
10	Agung Pratomo	1	18	4	16	2	2	1	1	113	
11	Agung Pratomo	1	18	4	16	2	2	1	1	113	
12	Agung Pratomo	1	18	4	16	2	2	1	1	113	
13	Agung Pratomo	1	18	4	16	2	2	1	1	113	
14	Agung Pratomo	1	18	4	16	2	2	1	1	113	
15	Agung Pratomo	1	18	4	16	2	2	1	1	113	
16	Agung Pratomo	1	18	4	16	2	2	1	1	113	
17	Agung Pratomo	1	18	4	16	2	2	1	1	113	
18	Agung Pratomo	1	18	4	16	2	2	1	1	113	
19	Agung Pratomo	1	18	4	16	2	2	1	1	113	
20	Agung Pratomo	1	18	4	16	2	2	1	1	113	
21	Agung Pratomo	1	18	4	16	2	2	1	1	113	
22	Agung Pratomo	1	18	4	16	2	2	1	1	113	
23	Agung Pratomo	1	18	4	16	2	2	1	1	113	
24	Agung Pratomo	1	18	4	16	2	2	1	1	113	
25	Agung Pratomo	1	18	4	16	2	2	1	1	113	
26	Agung Pratomo	1	18	4	16	2	2	1	1	113	
27	Agung Pratomo	1	18	4	16	2	2	1	1	113	
28	Agung Pratomo	1	18	4	16	2	2	1	1	113	
29	Agung Pratomo	1	18	4	16	2	2	1	1	113	
30	Agung Pratomo	1	18	4	16	2	2	1	1	113	

B. K-Means Method

Researchers perform calculations using an equation to calculate the distance between data on K-Means using the Euclidiance Distance (D) formula. The equations used are as follows :

$$D(x_2, x_1) = \sqrt{\sum_{j=1}^p (x_{2j} - x_{1j})^2} \dots \dots \dots (3)$$

- Description :
P = Data dimensions
x1 = Point Position 1
x2 = Point Position 2

The data used are 508 student data consisting of 11 attributes that are used for examples of calculations using K-Means Clustering manually.

$$DM1 = \sqrt{\frac{(1-1)^2 + (18-18)^2 + (4-4)^2 + (16-10)^2 + (2-2)^2 + (2-2)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (113-113)^2}{36}}$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 1 is 36 .

$$DM2 = \sqrt{\frac{(1-2)^2 + (18-60)^2 + (4-4)^2 + (16-7)^2 + (2-2)^2 + (2-2)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (113-112)^2}{1847}}$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 2 is 1847 .

$$DM3 = \sqrt{\frac{(1-1)^2 + (18-43)^2 + (4-1)^2 + (16-16)^2 + (2-2)^2 + (2-2)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (113-56)^2}{3883}}$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 3 is 3883 .

$$DM4 = \sqrt{\frac{(1-2)^2 + (18-18)^2 + (4-2)^2 + (16-10)^2 + (2-2)^2 + (2-2)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (113-121)^2}{105}}$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 4 is 105 .

$$DM5 = \sqrt{\begin{matrix} (1-2)^2 + (18-18)^2 + (4-4)^2 + \\ (16-16)^2 + (2-2)^2 + (2-2)^2 + \\ (1-1)^2(1-1)^2 + (1-1)^2(113-41)^2 \end{matrix}} = 5185$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 5 is 5185

$$DM6 = \sqrt{\begin{matrix} (1-2)^2 + (18-18)^2 + (4-1)^2 + \\ (16-10)^2 + (2-2)^2 + (2-2)^2 + \\ (1-1)^2(1-2)^2 + (1-2)^2(113-75)^2 \end{matrix}} = 1492$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 6 is 1492 .

$$DM7 = \sqrt{\begin{matrix} (1-2)^2 + (18-18)^2 + (4-2)^2 + \\ (16-10)^2 + (2-2)^2 + (2-2)^2 + \\ (1-1)^2(1-2)^2 + (1-2)^2(113-72)^2 \end{matrix}} = 1724$$

From the results of the above calculations, it is found that the distance between student 1 and the center of cluster 7 is 1724 .

In this case $d(m_i, m_j)$ represents the Euclidean distance from m to m_j . Calculating WCV That is by choosing the smallest distance between the data and the centroid in each cluster. The following is the closest distance (iteration 1) in the form of a table which can be seen in table 4:

Table 4. Nearest Distance (Iteration 1)

Jarak Terdekat
21,84
15,00
777,14
1048,89
131,19
1472,20
26,22
12770,47
12772,18
12338,11
10419,81
1698,11
2603,75
2604,18
12553,49
8714,78
12771,11
12770,90
487,82

909,49
12771,75
12782,61
3292,78
9803,11
21,84

WCV = 17707.41

So that the ratio = $BCV/WCV = 3583.43 / 17707.41 = 0,20236895$

Because this step is the first iteration then proceed to the next step.

Based on the calculations that have been carried out, the results for each cluster are obtained, for Cluster 1 as many as 111 with a percentage of 21.85%, Cluster 2 as many as 32 with a percentage of 6.3%, Cluster 3 as many as 30 with a percentage of 5.9%, Cluster 4 as many as 83 with a percentage of 16.33%, Cluster 5 as many as 99 with a percentage of 19.48%, Cluster 6 as many as 100 with a percentage of 19.68% and Cluster 7 as many as 53 with a percentage of 10.43%. The following is a graph of the results of manual calculations obtained as shown in Figure 4:

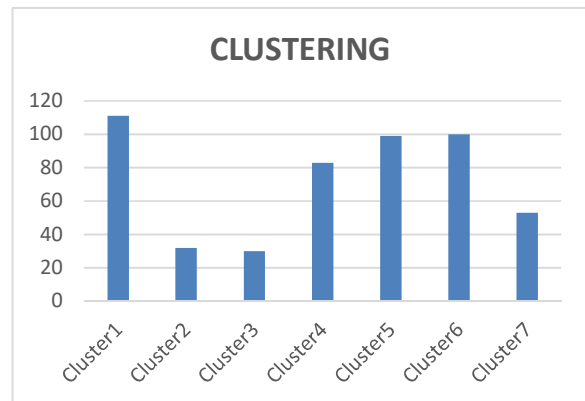


Figure 4. Manual Calculation

C. Results of Data Visualization of Each Attribute Using WEKA

1. Results of Visualization of Religious Attributes

The following is a form of attribute visualization that is used as a reference for promotional strategies using WEKA tools, namely:

a. Gender Attribute Visualization

Visualization of the Address Attribute. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The minimum statistic has a value of 0, the maximum statistic has a value of 2, the statistical mean (average) has a value of 1.283, the statistical standard deviation has a value of 0.455, Gender Attribute Visualization can be seen in Figure 5.

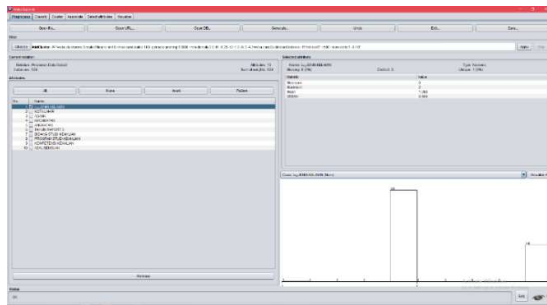


Figure 5 Visualization of Gender Attributes

b. Visualisasi b. City of Birth Attribute Visualization

Visualization of the City of Birth Attributes. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The minimum statistic has a value of 0, the maximum statistic has a value of 72, the statistical mean (average) has a value of 23,337, the statistical standard deviation has a value of 13,848, Visualization of the City of Birth Attributes can be seen in Figure 6.

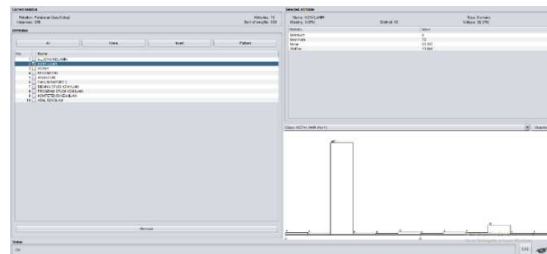


Figure 6. Visualization of Birth City Attributes

c. Religious Attribute Visualization

Visualization of Major Attributes. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The minimum statistic has a value of 1, the maximum statistic has a value of 5, the statistical mean (average) has a value of 1.758, the statistical standard deviation has a value of 1.058, the Visualization of Religious Attributes can be seen in Figure 7..

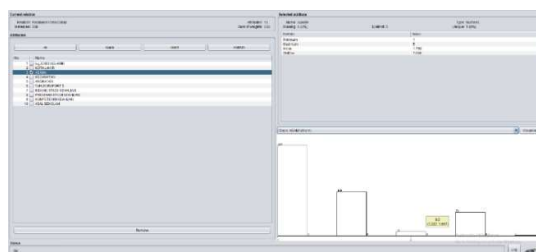


Figure 7. Visualization of Religious Attributes

d. District Attribute Visualization

Visualization of General Subject Value Attributes. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The

minimum statistic has a value of 0, the maximum statistic has a value of 16, the statistical mean (average) has a value of 11,352, the standard deviation statistic has a value of 3.334, Visualization of District Attributes can be seen in Figure 8.

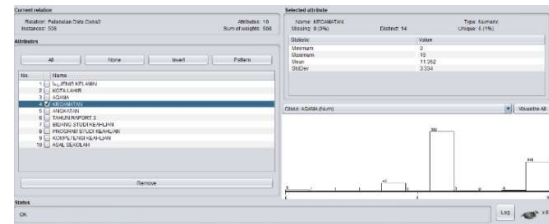


Figure 8. Visualization of District Attributes

e. VForce Attribute Visual

Visualization of School Origin Attributes. It is known that from 508 data in the column selected attribute there is no missing data as much as 0 or 0%. The minimum statistic has a value of 0, the maximum statistic has a value of 3, the statistical mean (average) has a value of 2.413, the statistical standard deviation has a value of 0.571, Visualization of Force Attributes can be seen in Figure 9.

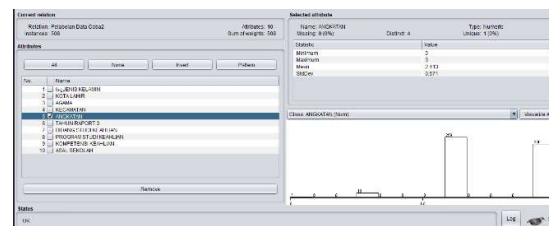


Figure 9 Visualization of Force Attributes

f. Last Report Year Attribute Visualization

Visualization of Religious Attributes. It is known that from 508 data there are 0 or 0% missing data. The minimum statistic has a value of 0, the maximum statistic has a value of 3, the statistical mean (average) has a value of 2.419, the statistical standard deviation has a value of 0.561, Visualization of the Attributes of the Year Reports can be seen in Figure 10.

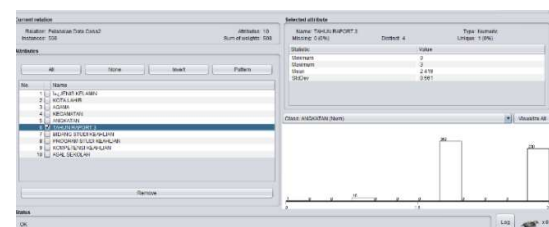


Figure 10 Visualization of the Last Report Year Attributes

g. Visualization of Field of Study Attributes of Expertise Visualisasi dari Atribut Bidang Studi Keahlian. Diketahui bahwa dari 508 data terdapat missing data sebanyak 10 atau 1%. Pada statistik minimum terdapat nilai 1, statistik maximum terdapat nilai 2, statistik mean (rata-rata) terdapat nilai 1,495, statistik standard deviasi terdapat nilai 0,5, Visualisasi Atribut Bidang Studi keahlian dapat dilihat

di Gambar 11.



Figure 11 Visualization of the Attributes of the Field of Expertise

h. Visualisasi h. Visualization of Skills Study Program Attributes

visualization of Kelurahan Attributes. It is known that from 508 data there are 0 or 0% missing data. The minimum statistic has a value of 1, the maximum statistic has a value of 3, the statistical mean (average) has a value of 1.591, the statistical standard deviation has a value of 0.645, Visualization of the Attributes of the Skills Study Program can be seen in Figure 11.

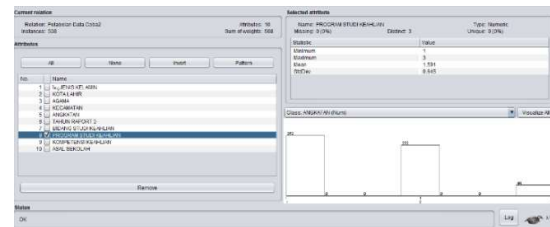


Figure 11. Visualization of Skills Study Program Attributes

i. Visualization of Skill Competency Attributes

It is known that from 508 data there are 0 or 0% missing data. The minimum statistic has a value of 1, the maximum statistic has a value of 3, the statistical mean (average) has a value of 1.591, the statistical standard deviation has a value of 0.645. Visualization of competency attributes can be seen in Figure 12.

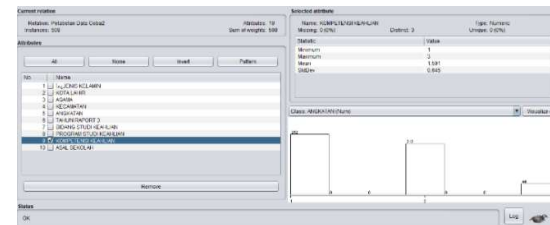


Figure 12 Visualization of Skill Competency Attributes

2. Visualization of School Origin Attributes

Visualization of the Residence Type Attribute. It is known that from 508 data there are 0 or 0% missing data. The minimum statistic has a value of 1, the maximum statistic has a value of 3, the statistical mean (average) has a value of 1.247, the statistical standard deviation has a value of 0.645, Visualization of School Origin Attributes can be seen in Figure 13..

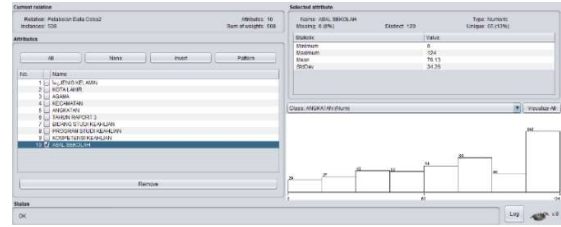


Figure 13 Visualization of School Origin Attributes

2. Cluster Analysis Results With Weka Tools

The results of the cluster analysis, where there are 7 predetermined clusters, the calculation is continued until all data is counted and produces groups into clusters with minimal distances. The iteration was stopped because there were the same cluster center numbers in the 8th iteration. The results of the cluster formed after the 8th iteration did not change, so the iteration was stopped. Clusters were chosen randomly after that the closest distance to the clusters was found in Cluster 4, Cluster 2, Cluster 3, Cluster 6, Cluster 5 and Cluster 7. From the results of cluster 1 it can be seen that the number of students was 57 students (11%). cluster 2 shows that the number of students is 75 students (15%), Then from the results of cluster 3 it can be seen that the number of students is 70 students (14%), then from the results of cluster 4 it can be seen that the number of students is 182 (36%), then from the results cluster 5 shows that the number of students is 44 (9%), then from the results of cluster 6 it can be seen that the number of students is 58 (11%), then from the results of cluster 7 it can be seen that the number of students is 22 (4%). The results of the cluster analysis with the Weka tools can be seen in Figure 14::

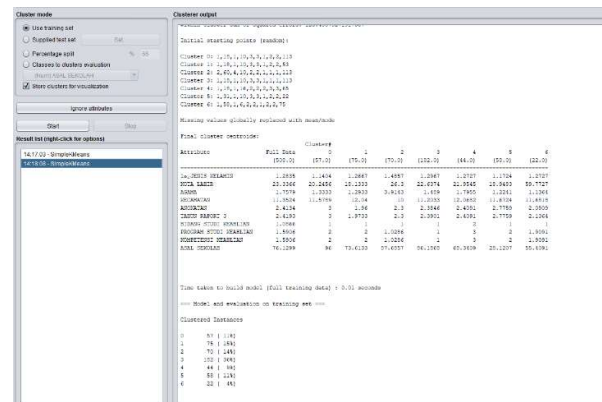


Figure 14 Cluster Analysis

3. Clustering Results Using Weka

From the data from the results of the clustering that has been carried out, it can be determined several promotional strategies that can be carried out by SMK Yadika 1 in conducting promotions based on the Marketing Department which are divided into 7 groups / clusters. The results of Clustering using Weka can be seen in table 5.

Table 5. Clustering Results Using Weka

Cluste r 1	Cluste r 2	Cluste r 3	Cluste r 4	Cluste r 5	Cluste r 6	Clus ter 7
1	1	2	1	1	1	1
18	18	60	18	18	31	58
1	1	4	1	1	1	1
10	10	10	10	16	10	6
3	3	2	3	2	3	2
3	3	2	3	2	3	2
1	1	1	1	2	1	1
2	2	1	1	3	2	2
2	2	1	1	3	2	2
113	53	113	113	65	75	75

4. Cluster Analysis Results Using Python

a. Selection of K values using the Davies Bouldin Index (DBI) method

After clustering, the next step is to analyze the results of the clustering using Python to get the "Davies Bouldin index" value. From the analysis carried out, for the lowest DBI with a DBI value of 0.552 and the optimal K is 7, and this value is far below 1.0 or it can be concluded that the grouping process is quite good according to Figure 15 below.

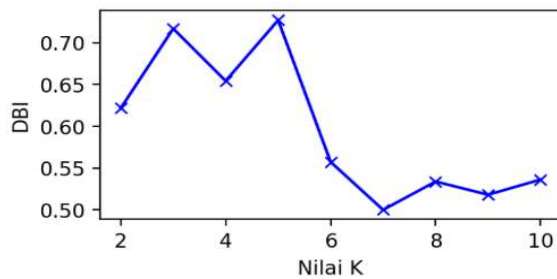


Figure 15 Davies Bouldin Index and K Optimal Values

D. K-means Modeling

After seeing the results of the Davies Bouldin Index (DBI) Method, for K-Means Modeling using Python, we can see in Figure 16 below :

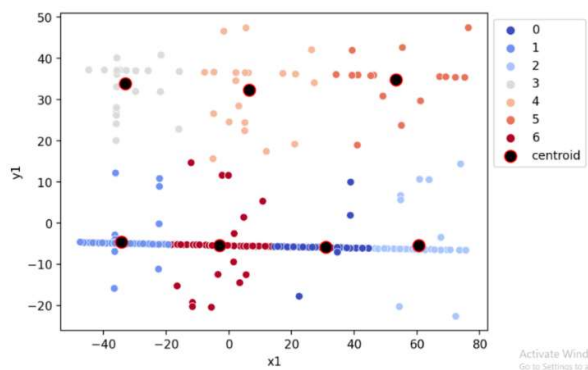
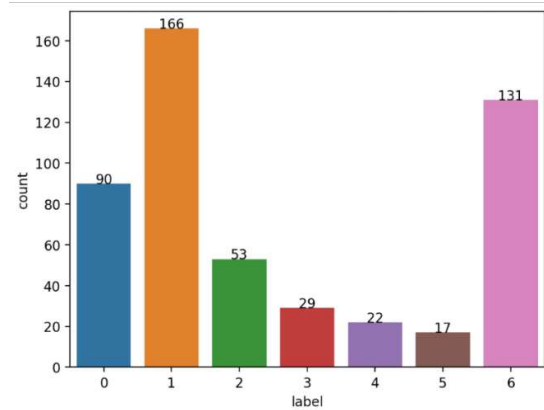


Figure 16 K-Means Modeling

1. Clustering Results

The results of the cluster analysis, where there are 7 clusters that have been determined according to the optimal K, the calculation results in groups into clusters

with a minimum distance. Because there are the same cluster center numbers in the 5th iteration. that the number of students is 90 students (17.71%), Then from the results of cluster 2 it can be seen that the number of students is 166 students (32.67%), Then from the results of cluster 3 it can be seen that the number of students is 53 students (10.43%), then from the results of cluster 4 it can be seen that the number of students is 29 (5.7%), then from the results of cluster 5 it can be seen that the number of students is 22 (4.33%), then from the results of cluster 6 it can be seen that the number of students is 17 (3.34%), then from the results of cluster 7 it can be seen that the number of students as many as 131 (25.78%). with the data population according to Figure 17.



```
[7]:
```

K	dbi
0	2 0.635089
1	3 0.741507
2	4 0.684332
3	5 0.769454
4	6 0.599900
5	7 0.552802
6	8 0.615288
7	9 0.632581
8	10 0.653840

Figure 17 Clustering Results

2. Business Use Case

The following design is a business use case for the system built on the K-Means Calculation application design to determine the best promotion strategy according to Figure 19 below :

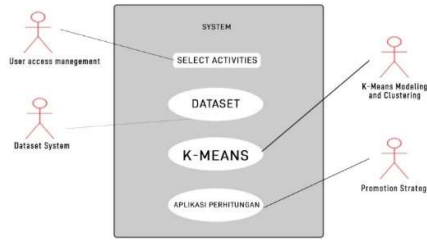
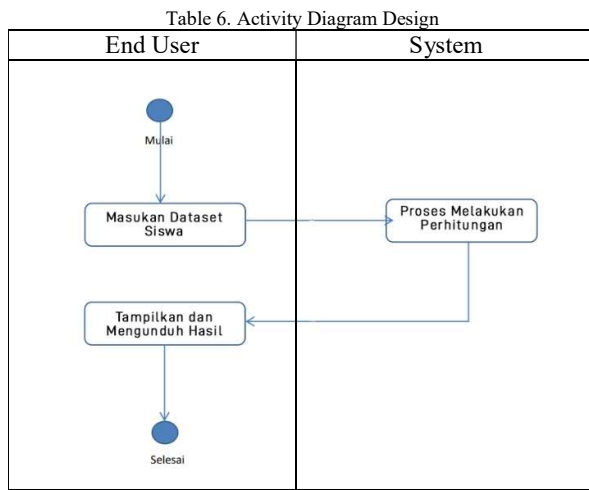


Figure 19. Business Use Case

3. Activity Diagram

The following is an activity diagram design that was built, the stages of the process of using the K-Means calculation application to determine the promotion strategy in accordance with the system design according to Table 6 below :



4. Prototype Implementation

At the implementation stage of this prototype, the prototype was built using the Python programming language with streamlit as the interface and uploading the dataset and downloading the results, the following is the initial display of the prototype calculation application according to Figure 20 below :

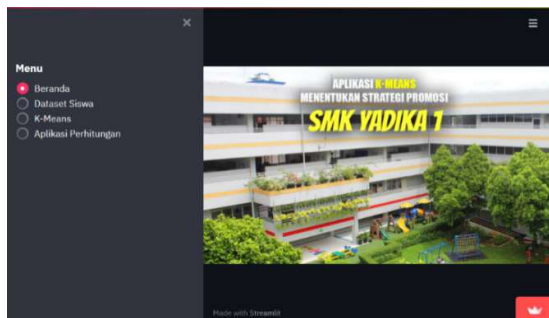


Figure 20 Initial Screen of Applications

On the first page of the prototype there are several menus including the homepage, student datasets, K-Means modeling, and Calculation Applications as the K-Means

calculation process to determine the best promotion strategy. On the student dataset page to see the attributes that will be used in the K-Means search process according to Figure 21 below.

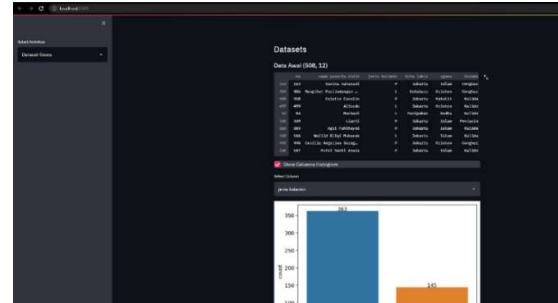


Figure 21 Student Dataset Pages

On the K-Means page there are the results of selecting the K value using the Davies Bouldin Index (DBI) Index, the K-Means Model Simulation with the K value can be changed and the clustering results according to Figure 22 below

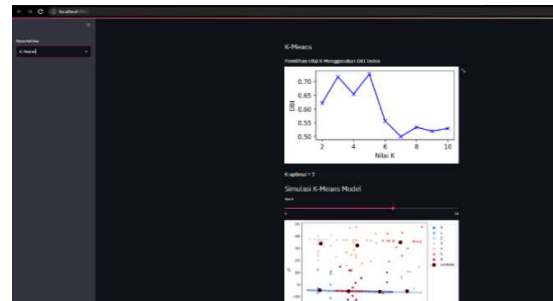



Figure 22 K-Means . Prototype Pages

5. Black Box Testing

a. Black Box Testing on the Calculation Application Page

In the Black Box test on the calculation application page to see whether the expected results are appropriate or not, it can be seen in table 7.

Tabel 7 Hasil Aplikasi Perhitungan K-Means

No	Testing Scenario	Expected results	Conclusion
1.	Upload a Dataset whose data is incorrect and does not match the format that has been set in the calculation application, then let it process the inputted data, Testing:	The system will display errors or process data that does not match the format set in the calculation application. Test result : 	Already appropriate
2.	Upload a Dataset whose data is	The system will show the process that took place until	Already

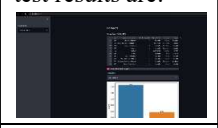



	incorrect True according to the format that has been set in the calculation application, then let it process the inputted data, Testing:	the result and can be downloaded. Test result :	appropriate
--	--	---	-------------










b. *Black Box Testing on Dataset Pages*

Pada pengujian *Black Box* pada halaman dataset perhitungan untuk melihat hasil yang diharapkan apakah sesuai atau tidak bisa dilihat pada tabel 8.

Table 8 Student Dataset Results

No	Testing Scenario	Expected results	Conclusion
1.	The system will display the dataset menu if you click the dataset menu button.	The system displays the dataset menu, the test results are:	Already appropriate
			
2.	In the Dataset, the system will display a menu for each Gender attribute if you click the button according to the attribute that will be displayed.	The system displays the Gender attribute on the dataset menu, the test results are:	Already appropriate
			
3.	In the dataset, the system will display a menu for each attribute City of birth if you click the button according to the attribute that will be displayed.	The system displays the City of birth attribute on the dataset menu. The test results are:	Already appropriate
			
4.	In the dataset, the system will display the Religion attribute menu if you click the button according to the attribute to be displayed.	The system displays the Religion attribute on the dataset menu, the test results are:	Already appropriate
			
5.	In the dataset, the system will display the sub-	The system displays the sub-	Already appropriate


	district attribute menu if you click the button according to the attribute that will be displayed.	on the dataset menu, the test results are:	
			
6.	In the Dataset, the system will display the Force attribute menu if you click the button according to the attribute to be displayed.	The system displays the Force attribute on the dataset menu, the test results are:	Already appropriate
			
7.	In the dataset, the system will display the last Report Year attribute menu for the last report card if you click the button according to the attribute to be displayed.	The system displays the last Report Year attribute in the dataset menu, the test results are:	Already appropriate
			
8.	In the dataset, the system will display the attributes of the field of expertise if you click the button according to the attribute that will be displayed.	The system displays the attributes of the field of expertise on the dataset menu, the test results are:	Already appropriate
			
9.	In the dataset, the system will display an attribute menu for the study program of expertise if you click the button according to the attributes that will be displayed.	The system displays the attributes of the skill study program on the dataset menu, the test results are:	Sudah sesuai
			
10.	In the dataset, the system will display the skill competency attributes on the dataset menu, the test results are:	The system displays the skill competency attributes on the dataset menu, the test results are:	Already appropriate
			

	the attribute to be displayed.		
11	In the dataset, the system will display the menu for each school attribute in detail if you click the button according to the attribute to be displayed.	The system displays the school's attributes on the dataset menu. The test results are: 	Already appropriate

c. Black Box Testing on K-Means Pages

In the Black Box test on the calculation application page to see whether the expected results are appropriate or not, it can be seen in table 9.

Table 9 Results of K-Means

No	Testing Scenario	Expected results	Conclusion
1.	The system will display the K-Means menu, if you click the dataset menu button, the results of the K-Means Model will be displayed.	The system displays the K-Means menu, the test results are: 	Already appropriate
2.	The K-Means Menu will display a K-Means Model Simulation whose K value can be changed if you click the simulation button with a K value from numbers 2 to 10 then it will be displayed according to your choice.	The system displays the K-Means Model Simulation by selecting the value of K 10 on the K-Means menu. The test results are:	Already appropriate
3.	The K-Means menu will display Cluster Options from Cluster 1 to Cluster 10, according to the Optimal K we want.	The system displays a choice of clusters that match our choice, the test results are:	Already appropriate

V. CONCLUSION

A. Conclusion

Based on the results of observations and research that has been done at SMK Yadika 1, it can be concluded as follows::

1. Applying the K-Means Method to determine the best promotion strategy in accordance with SMK Yadika 1.
 - a. The application of the K-Means Clustering Algorithm resulted in 7 clusters where Cluster 1 90 data (17.71%), cluster 2 166 data (32.67%), cluster 3 53 data (10.43%), cluster 4 29 data (5.7%), cluster 5 22 data (4.55%), cluster 6 17 data (3.34%), and cluster 7 131 data (25.78%) and the Davies Boulding Index (DBI) with a value of 0.552 and an optimal K of 7.
 - b. Based on the calculations that have been made, the Promotion Strategy that can be carried out by the school so that the promotion is carried out more effectively and efficiently are:
 - 1) Promotion based on Student's School Origin.
 - 2) Promotion based on the most sought after Field of Study of Expertise.
 - 3) Promotion based on Skills Study Program
 - 4) Promotion based on Skill Competence
 - 5) Promotion Student
2. Develop a data mining prototype to determine the best strategy according to the needs of SMK Yadika 1.

B. SUGGESTION

From the conclusions mentioned above, the authors provide suggestions for further development of the application of the K-means Clustering Method to determine the best promotion strategy, namely by developing a system for language that is easier to understand so as to assist schools in receiving the information generated and this research can be developed with adding the amount of new data in the attributes of the data mining prototype.

REFERENCES

- [1] Sumangkut, K., Lumenta, A. S. M. and Tulenan, V. 2016. Analysis of Daily Mart supermarket shopping patterns to determine the layout of goods using the FP-Growth Algorithm. *Journal of Informatics Engineering*, 8(1).
- [2] Budiman, I. 2015. Application of Classification Data Mining Functions for Prediction of Timely Student Study Period in Higher Education Academic Information Systems. *Journal of Jupiter*, 7(1), pp. 39–50.
- [3] Badrul, M. 2016. Association Algorithm With Apriori Algorithm for Sales Data Analysis. *Journal of Pilar Nusa Mandiri*, 12(2), pp. 121–129.
- [4] Asril, E., Wiza, F. and Yunefri, Y. 2015. Analysis of Graduate Data with Data Mining to Support the Promotion Strategy of Lancang Kuning University. *Journal of Information & Communication Technology Digital Zone*, 6(2), pp. 24-32.
- [5] Priambudi, D. S. 2015. Pt.Mayora's Product Sales Strategy Using Apriori Methods And Data Mining Implementation. Thesis Article, pp. 1–9.
- [6] Rony, S. 2016. Data Mining Application Using K-Means Clustering Algorithm to Determine New Student Promotion Strategy (Case Study: Polytechnic Lp3i Jakarta). *Journal of Lanterns ICT*, 3(1), pp. 76–92.

- [7] Suprawoto, T. 2016. Classification of Student Data Using the K-Means Method to Support the Selection of Marketing Strategies. *Journal of Informatics and Computers*, 1(1), pp. 12–18.
- [8] Wirta, A. and Erlin. 2016. Implementation of the K-Means Cluster Analysis Method for Selecting New Student Admission Promotion Strategies. *National Seminar on computer science*, pp. 9–15.
- [9] Mochammad C. A. 2018. Data Mining Uses the K-Means Algorithm to determine promotion strategies in intentional vocational schools. *Thesis Article*, pp. 1–12.
- [10] Achyani, Y. E. 2018. Application of Particle Swarm Optimization Method in Optimizing Direct Marketing Predictions. *Journal of Informatics*, 5(1), pp. 1–11.
- [11] Kusumo, H., Sedyono, E. and Marwata, M. 2019. Analysis of Apriori Algorithms to Support Higher Education Promotion Strategies. *Walisongo Journal of Information Technology*, 1(1), p. 49.
- [12] Jaini, A. 2019. Application of the Fuzzy C-Means Algorithm and K-Medoids for Grouping Sales and Product Marketing Strategies. *State Islamic University of Sultan Syarif Kasim Riau*.
- [13] Takdirillah, R. 2020. Application of Data Mining Using Apriori Algorithm Against Transaction Data to Support Sales Strategy Information. *Journal of Informatics Education*, 4(1), pp. 37–46.