**Acceleration, Quantum, Information Technology and Algorithm Journal (AQILA)**
Vol. 1, No. 1, Juni 2024, pp 6-15
ISSN: 3062-8555, DOI: 10.62123/aqila.v1i1.23

# K-Means Algorithm for Clustering Students Based on Areas of Expertise (A Case Study)

Yuyun Yusnida Lase[1*], Christian Roi Tua Sinaga[1], Muhammad Rivan Nugroho[1], Muhammad Rasyid Ridha[1]

[1]Politeknik Negeri Medan, Jalan Almamater No 1 Kampus USU, Medan, Indonesia

*Corresponding Email: yuyunlase@polmed.ac.id

**DOI :** 10.62123/aqila.v1i1.23

**Received** : May 12, 2024
**Revised** : May 25, 2024
**Accepted** : May 30, 2024

***Keywords:***
Students Expertise
K-Means
Clustering
Prediction

## ABSTRACT

Every student of the Software Engineering Technology study program, the Computer and Informatics Department of the Politeknik Negeri Medan in the lecture process is required to take all the courses in the curriculum. From the courses inthe curriculum, there are several subject groups that shape student expertise, such as Software Engineering, System Analyst, Database Administrator, and IT Entrepreneur. It is hoped that this expertise will later be used as a reference by students in carrying out their thesis at the end of lectures. The purpose of this study is to group students based on their respective expertise based on data processing of student course scores related to each skill. The data used is data on student scores for batches of 2020 and 2021 with a range of courses from semester 1 to semester 3. The data is tested by implementing the K-Means algorithm. The results of the tests that have been carried out show the grouping of students based on their expertise, with 7 times the number of iterations. Then, data testing was carried out with the RapidMiner application to get the results of the distribution of cluster members obtained, including 12 students occupying Software Engineer skills, 21 students with System Analyst skills, 5 students with Database Administrator skills, and 31 students with IT Entrepreneur skills, along with the distribution chart. Thus, the K-Means algorithm is quite good at grouping students based on their expertise.

## 1. INTRODUCTION

The Software Engineering Technology study program, the Department of Computers and Informatics, State Polytechnic of Medan, is an applied undergraduate education level Diploma 4 which was established on 05 May 2020 based on the Decree of the Ministry of Education and Culture No.43704/A5/HK/2020, oneof the objectives of which is to produce human resources who master the field software engineering technology so that it can compete at national and international levels. In an effort to realize this goal, the Software Engineering Technology Study Program implements a curriculum based on the Decree of the Minister of Education and Culture No.232/U/2000 where the learning curriculum is composed of Personality Development Courses, Science and Skills Courses, Work Skills Courses, Work Behavior Courses, and Community Life Courses. Every student is required to carry out several processes in lectures to complete the education level and get an applied bachelor's degree, thesis or final assignment is the culmination of thelecture process. The Thesis or Final Project is one of the requirements that must be met by students to complete their education at tertiary institutions [1]. Thesis can be considered as an indicator of the percentage of understanding and achievement of a student's knowledge, a student studies according to a certain area of expertise [2], so that students can be expected to become professional and quality human resources in accordance with the field of expertise based on the abilities mastered by thestudents [3], Field of expertise students are grouped based on student expertise on the values of certain courses that students have obtained [4]. Therefore, grouping student areas of expertise based on the abilities mastered is very necessary [5]. To group the student data, a data processing technique that can be used is Data Mining. Data mining is a data mining process as well as a process of filtering data against large data sets to obtain information from the data [6]. One method of data mining is clustering. Clustering [7] means separating data into several groups based on the similarity between the data objects [8]. The K-Means algorithm is a clustering algorithm that is commonly used to identify groups and grouping a casedata set. K-means includes an iterative grouping procedure that partitions and groups a large number of data objects.The grouping processed by K-means is based on data that has similarities so that the results of the grouping can be analyzed. In related research that has been carried out [9], with the title Implementation of the K- Means Clustering Algorithm for Selection of Achieving Students Based on Activeness in the Learning Process discusses grouping active and achieving students by implementing the K-Means algorithm.

Based on the results of research on 59 students, students were grouped into 3 clusters with a distribution of 8 students in the inactive student cluster, 21 less active student clusters, and 30 active student clusters, where the results ofthis research can be used as a reference for evaluation by the parties. school. The K-Means algorithm was also implemented in research [10] entitled K- Means Algorithm for Grouping Student Thesis Topics, resulting in grouping of students according to their expertise, where grouping with the highest cluster shows that the student's ability in each group of areas of expertise, which indicates Students have good skills in that field, so they can choose essay topics that suit their group of areas of expertise. Research [5] entitled Data Mining Grouping Student Expertise Using the K-Means Algorithm (Case Study: Cic University Cirebon) discusses grouping students using the k-means algorithm, after carrying out 3 iterations and testing with the Tanagra application, the results obtained produce the same number of clusters and number of members with an accuracy of 80%. Research conducted by Dina [11] entitled Application of the K-Means Clustering Algorithm to Determine the Capabilities of IT Employees with an accuracy of 40%, resulted in 5 clusters of employees with excellent, good, fair, poor and very poor abilities. These results were obtained because the centroid was chosen randomly, resulting in different iterations and number of members.

In research [12] entitled K-Means clustering for the classification of educational qualification standards and work experience of vocational school teachers in Indonesia which implements the K-Means algorithm, the results obtained were that cluster 1 consisted of 8 provinces, cluster 2 consisted of 19 provinces, and cluster 3 consists of 7 provinces. Cluster 2 was found to be the best cluster where the province met the requirements to become a model for quality human resource management in Indonesia. Based on the background and relevant research above, this research carried out the implementation of the K-Means algorithm in grouping students based on their areas of expertise with the aim of finding out groups of students based on their expertise as a reference or support for students and campuses in selectingthesis or Final assignment topics.

## 2. LITERATURE REVIEW

### 2.1 Clustering

Clustering is defined as placing a set of combinations of objects that have similarities, in such a way that these objects are related to one another compared to other groups of sets. Data objects are grouped into clusters so that other objects that are in one cluster will have similarity and tend to have a high resemblanceto one another [13][14][15][16]. Grouping of data or objects that can be done can minimize the variation of data in a group/cluster.

### 2.2 K-Means algorithm

K-Means algorithm is an algorithm in Data Mining where an unsupervised modeling process is carriedout and the process of grouping data into partitions is carried out, where the data in the pzzartition group have the same characteristics and are different from data in other group partitions [17zzz]. The K-means algorithm requires k input parameters and divides a set of n objects into k clusters so that the degree of similarity between members in a cluster is high and the clusters are gradually very low [18]. The similarity or similarity of the member data objects to the cluster can be measured by the closeness of the object to the mean value in the cluster which is commonly called the cluster centroid. To group a data object into clusters, you can follow the following steps in the K-Means algorithm process [19]:

1. Determine how many clusters you want to form, so the value of k is the number of clusters.
2. Determine how many centroid values of each cluster are formed. The centroid value is determinedrandomly from the data set and the number of centroids is equal to the number of clusters.
3. Calculate the distance of each data with each centroid with the Euclidean distance formula:

$$d(X,Y) = \sum_{j=i}^{x} \left\{ p_j(X) - p_j(Y) \right\}^2 \tag{1}$$

Information:

d = data distance to centroid
p = Data object value
X = Data on record
Y = Centroidvalue

4. Based on the results of calculating the distance between the data and the centroid, the data isgrouped based on the smallest distance or minimum distance.
5. From grouping the data, then find a new centroid based on the membership of each cluster bycalculating the cluster average. To determine it can use the formula:

$$C\ (i) = \frac{x1 + x2 + \cdots + xn}{\Sigma x} \tag{2}$$

Information:

x1...xn = value of data records 1 to n
Σx = number of data records.

6. After that, repeat Step 3 again for several iterations. The iteration stops or is not carried out again after the iteration group has no changes in the previous iteration.

## 3. RESEARCH METHODS

In this study, the stages in the process of grouping students based on areas of expertise with datamining are explained. The research process flowchart can be seen in Figure 1.

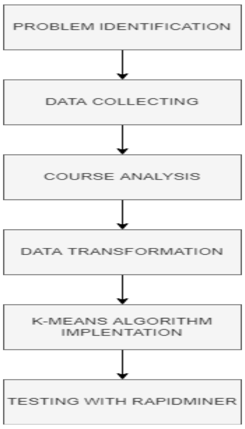


**Figure 1.** Research Flowchart

### 3.1 Problem analysis

Based on the problems that have been found through observation and literature study, an analysis is carried out, this stage is carried out to understand the existing problems. By analyzing the problem, the problem can be well understood, and at this stage the writer determines what datato collect and conducts literature study on existing methods related to this research.

### 3.2 Data collection

Data collection is done by:

1. Observation
   Observation is directly observing the object of research by observing data on course grades and student interest in areas of expertise related to student expertise.
2. Study of literature
   Literature studies are obtained by looking for references from various literary sources, such as books and relevant publications, journals, and articles to clarify problems and solutions for research.

### 3.3 Course analysis

The data used in this study is in the form of data on student scores in the Software EngineeringTechnology Study Program, Politeknik Negeri Medan class of 2020 and 2021 starting from semesters 1 to 3 based on their subject area of expertise. The subjects used as assessments in this study can be seen in Table 1.

**Table 1.** Course Group

| No. | Subject | Group |
|---|---|---|
| 1 | Logic and Algorithms | SoftwareEngineer |
| 2 | Logic and AlgorithmPractice | |
| 3 | OOP I | |
| 4 | OOP Practice I | |
| 5 | Information Technology | Systems Analyst |
| 6 | Information Technology Practices | |
| 7 | Data and NetworkCommunications | |
| 8 | Operating system | |
| 9 | Database System | Database Administrators |
| 10 | Database SystemPractice | |
| 11 | Data Structure | |
| 12 | Data Structure Practice | |
| 13 | Management information System | |

| | | |
|---|---|---|
| **14** | Graphic Design / WebDesign | IT Entrepreneur |
| **15** | OOP II (Project based) | |
| **16** | OOP II Practice (Project-based) | |

### 3.4 Data transformation

At this stage the process of analyzing student grades data that has been obtained at the data collection stage, the data is grouped into tables that are easy to understand [20], then the data is cleaned with the aim of reducing and removing inconsistent data, duplication, and integration or merging data from several data sources, the purpose of this data transformation is to change the data and make the data more accurate, clear, and suitable for research needs. As for the data resulting from the transformation can be seen in Table 2.

**Table 2.** Student Test Data

| No. | Student | Software Engineer | SystemsAnalyst | Database Administrators | IT Entrepreneur |
|---|---|---|---|---|---|
| **1** | Student1 | 66 | 67 | 60 | 52 |
| **2** | Student2 | 83 | 86 | 87 | 86 |
| **3** | Student3 | 76 | 75 | 70 | 79 |
| **4** | Student4 | 79 | 75 | 81 | 79 |
| **5** | Student5 | 71 | 73 | 72 | 69 |
| **6** | Student6 | 74 | 79 | 84 | 83 |
| **7** | Student7 | 73 | 76 | 74 | 75 |
| **8** | Student8 | 74 | 78 | 81 | 84 |
| **9** | Student9 | 79 | 76 | 82 | 81 |
| **10** | Student10 | 80 | 78 | 81 | 80 |
| **…** | … | … | … | … | … |
| **65** | Student65 | 71 | 69 | 75 | 81 |
| **66** | Student66 | 83 | 81 | 82 | 87 |
| **67** | Student67 | 80 | 79 | 79 | 80 |
| **68** | Student68 | 77 | 79 | 79 | 81 |
| **69** | Student69 | 78 | 75 | 82 | 83 |

### 3.5 Application of the k-means algorithm

Based on the data that has been obtained and collected and has been transformed, then the calculation process is carried out with the K-Means algorithm. The result of this process is in the form of data grouping. In the next stage, the data will be tested using RapidMiner.

### 3.6 Testing with RapidMiner

Tests were carried out on the calculated data to ensure the results of applying the K-Means algorithm.

## 4. DISCUSSION AND RESULT

### 4.1 Implementation of the k-means algorithm

K-Means algorithm for data has the goal of grouping data according to the clusters formed [21] The following is the K-Means implementation process [22]:

1. Determine the number of clusters or data groups. The number of clusters formed corresponds to the predetermined number of expertise groups, which is 4 clusters. Therefore, the value of k = 4, namely: Software Engineering (SE), System Analyst (SA), Database Administrator (DA), IT Preneur (IT).
2. Determine the starting centroid point. Centroid points are taken randomly on student data, each cluster has 4 centroid values according to the number of k. Initial centroid values can be seen in Table 3.

**Table 3.** Initial *Centroid* Values

| | **Initial Centroid of Each Cluster** | | | |
|---|---|---|---|---|
| | SE | SA | DA | IT |
| **C1** | 80.51 | 80,2 | 80,2 | 82 |
| **C2** | 69 | 69 | 72.5 | 74.5 |
| **C3** | 71 | 69 | 61 | 56 |
| **C4** | 75 | 77 | 72 | 76 |

3. Calculate the distance from each data to the centroid value. On the calculation of the distance (Euclidean) formula (1) is used. Following are the results of calculating cluster distances with Microsoft Excel so that they are obtained as shown in Table 4.

**Table 4.** Calculation Of Iteration Distance 1

| **Data *i*** | **N1** | **N2** | **N3** | **N4** | **C1** | **C2** | **C3** | **C4** |
|---|---|---|---|---|---|---|---|---|
| **Student1** | 65.5 | 67.0 | 60.0 | 51.5 | 41.9 | 26,6 | 21.5 | 31 |
| **Student2** | 82.5 | 85.8 | 86.5 | 86.3 | 9,5 | 28 | 34,9 | 20,8 |
| **Student3** | 75.8 | 74.8 | 70.0 | 78.5 | 13,1 | 9,8 | 12,7 | 4 |
| **Student4** | 79.3 | 74.8 | 81.3 | 78.8 | 6,7 | 14,9 | 23 | 10,7 |
| **Student5** | 70.5 | 72.5 | 72.0 | 68.5 | 20,4 | 7,1 | 11,4 | 10,4 |
| **Student6** | 73,8 | 78.8 | 83.8 | 83.3 | 7,8 | 17,7 | 26,6 | 13,8 |
| **Student7** | 72.5 | 76,3 | 74,3 | 75,3 | 12,8 | 8,1 | 14,8 | 3,9 |
| **Student8** | 74,3 | 78.3 | 80.8 | 83.5 | 6,7 | 16 | 24,2 | 11,4 |
| **Student9** | 78.5 | 75.5 | 82.3 | 81.0 | 5,7 | 16,1 | 24.5 | 11,9 |
| **Student10** | 80.0 | 77,8 | 81.0 | 79.8 | 3,6 | 16,9 | 24,2 | 10,7 |
| **…** | … | … | … | … | … | … | … | … |
| **Student65** | 70.5 | 68.5 | 75,3 | 81.3 | 16,3 | 7,3 | 16,6 | 11,7 |
| **Student66** | 83.0 | 81.0 | 81.5 | 86.5 | 5,2 | 23.5 | 29,7 | 16,3 |
| **Student67** | 80.0 | 79.0 | 78.8 | 79.5 | 3,4 | 16,6 | 22,8 | 9 |
| **Student68** | 76.5 | 79.3 | 78.5 | 81.0 | 4,6 | 15,2 | 22,1 | 8,3 |
| **Student69** | 77,8 | 74.5 | 81.8 | 82.5 | 6,6 | 15,7 | 24,3 | 12,2 |

4. Grouping data according to the minimum distance to *the centroid*. The data group for each is inTable 5.

**Table 5.** Members Of The Iteration Cluster 1

| **Data *i*** | **N1** | **N2** | **N3** | **N4** | **C1** | **C2** | **C3** | **C4** |
|---|---|---|---|---|---|---|---|---|
| **Student1** | 65.5 | 67.0 | 60.0 | 51.5 | | | * | |
| **Student2** | 82.5 | 85.8 | 86.5 | 86.3 | * | | | |
| **Student3** | 75.8 | 74.8 | 70.0 | 78.5 | | | | * |
| **Student4** | 79.3 | 74.8 | 81.3 | 78.8 | * | | | |
| **Student5** | 70.5 | 72.5 | 72.0 | 68.5 | | | | * |
| **Student6** | 73,8 | 78.8 | 83.8 | 83.3 | * | | | |
| **Student7** | 72.5 | 76,3 | 74,3 | 75,3 | | | | * |
| **Student8** | 74,3 | 78.3 | 80.8 | 83.5 | * | | | |
| **Student9** | 78.5 | 75.5 | 82.3 | 81.0 | * | | | |
| **Student10** | 80.0 | 77,8 | 81.0 | 79.8 | * | | | |
| **…** | … | … | … | … | | | | |
| **Student65** | 70.5 | 68.5 | 75,3 | 81.3 | | * | | |
| **Student66** | 83.0 | 81.0 | 81.5 | 86.5 | * | | | |
| **Student67** | 80.0 | 79.0 | 78.8 | 79.5 | * | | | |
| **Student68** | 76.5 | 79.3 | 78.5 | 81.0 | * | | | |
| **Student69** | 77,8 | 74.5 | 81.8 | 82.5 | * | | | |

5. Next is to determine the new centroid value to carry out the next iteration process, the centroid value can be obtained from the data that has been grouped in the previous iteration based on the cluster. In Table 6 the centroid values for the 2nd iteration using Microsoft Excel.

**Table 6.** Centroid 2nd Iteration

| **Centroid Iteration - 2** | | | | |
|---|---|---|---|---|
| | **SE** | **SA** | **DA** | **IT** |
| **C1** | 80.55 | 80.52 | 80,36 | 82,28 |
| **C2** | 69,62 | 69.50 | 72.50 | 73,87 |
| **C3** | 71.00 | 69.75 | 61.50 | 56.50 |
| **C4** | 75,73 | 77,21 | 72,21 | 76,73 |

6. Determine the 2nd iteration cluster. The 2nd iteration cluster grouping can be seen in Table 7.

**Table 7.** Members Of The 2nd Iteration Cluster

| Data *i* | N1 | N2 | N3 | N4 | C1 | C2 | C3 | C4 |
|---|---|---|---|---|---|---|---|---|
| **Student1** | 65.5 | 67.0 | 60.0 | 51.5 | | | * | |
| **Student2** | 82.5 | 85.8 | 86.5 | 86.3 | * | | | |
| **Student3** | 75.8 | 74.8 | 70.0 | 78.5 | | | | * |
| **Student4** | 79.3 | 74.8 | 81.3 | 78.8 | * | | | |
| **Student5** | 70.5 | 72.5 | 72.0 | 68.5 | | * | | |
| **Student6** | 73,8 | 78.8 | 83.8 | 83.3 | * | | | |
| **Student7** | 72.5 | 76,3 | 74,3 | 75,3 | | | | * |
| **Student8** | 74,3 | 78.3 | 80.8 | 83.5 | * | | | |
| **Student9** | 78.5 | 75.5 | 82.3 | 81.0 | * | | | |
| **Student10** | 80.0 | 77,8 | 81.0 | 79.8 | * | | | |
| **…** | … | … | … | … | | | | |
| **Student65** | 70.5 | 68.5 | 75,3 | 81.3 | | * | | |
| **Student66** | 83.0 | 81.0 | 81.5 | 86.5 | * | | | |
| **Student67** | 80.0 | 79.0 | 78.8 | 79.5 | * | | | |
| **Student68** | 76.5 | 79.3 | 78.5 | 81.0 | * | | | |
| **Student69** | 77,8 | 74.5 | 81.8 | 82.5 | * | | | |

7. The iteration continues until all centroid values do not change or the changes that occur are no longer significant. From the results of the research above, it was carried out 7 times to get a centroid value that did not change anymore. So, the final cluster results from the K-Means calculation can be seen in Table 8.

**Table 8.** Final Result Cluster

| Data *i* | N1 | N2 | N3 | N4 | C1 | C2 | C3 | C4 |
|---|---|---|---|---|---|---|---|---|
| **Student1** | 65.5 | 67.0 | 60.0 | 51.5 | | | * | |
| **Student2** | 82.5 | 85.8 | 86.5 | 86.3 | * | | | |
| **Student3** | 75.8 | 74.8 | 70.0 | 78.5 | | * | | |
| **Student4** | 79.3 | 74.8 | 81.3 | 78.8 | | | | * |
| **Student5** | 70.5 | 72.5 | 72.0 | 68.5 | | * | | |
| **Student6** | 73,8 | 78.8 | 83.8 | 83.3 | | | | * |
| **Student7** | 72.5 | 76,3 | 74,3 | 75,3 | | * | | |
| **Student8** | 74,3 | 78.3 | 80.8 | 83.5 | | | | * |
| **Student9** | 78.5 | 75.5 | 82.3 | 81.0 | | | | * |
| **Student10** | 80.0 | 77,8 | 81.0 | 79.8 | | | | * |
| **…** | … | … | … | … | | | | |
| **Student65** | 70.5 | 68.5 | 75,3 | 81.3 | * | | | |
| **Student66** | 83.0 | 81.0 | 81.5 | 86.5 | | | | * |
| **Student67** | 80.0 | 79.0 | 78.8 | 79.5 | | | | * |
| **Student68** | 76.5 | 79.3 | 78.5 | 81.0 | | | | * |
| **Student69** | 77,8 | 74.5 | 81.8 | 82.5 | | | | * |

### 4.2 Implementation of the k-means clustering algorithm with RapidMiner

RapidMiner is a GUI-based software that is used to perform data analysis using existing methods in data mining such as Naive Bayes, KNN, K-Means, and others [23]. The input data used is in the form of csv data which contains data on student scores that have passedthe data transformation stage. Figure 2. shows the *K-Means* implementation scheme in RapidMiner.

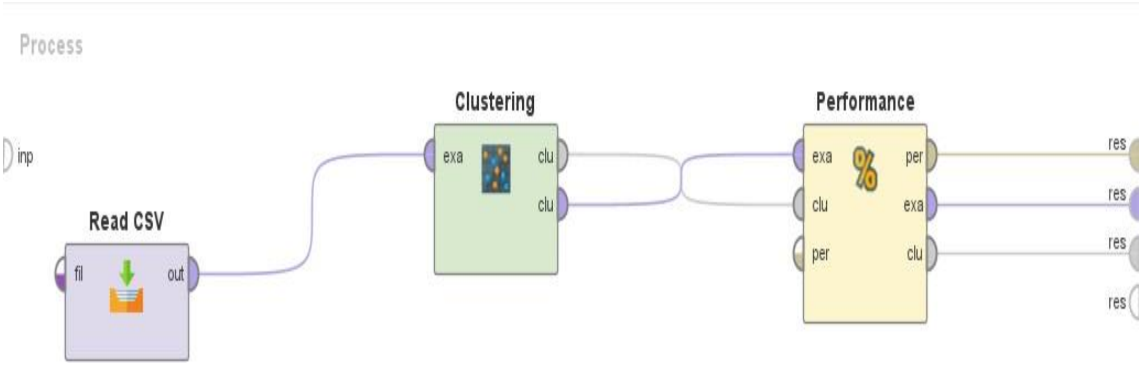
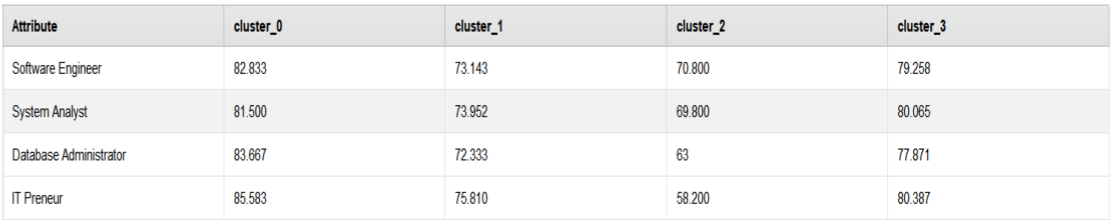


**Figure 2.** K-Means Process Schematic on RapidMiner

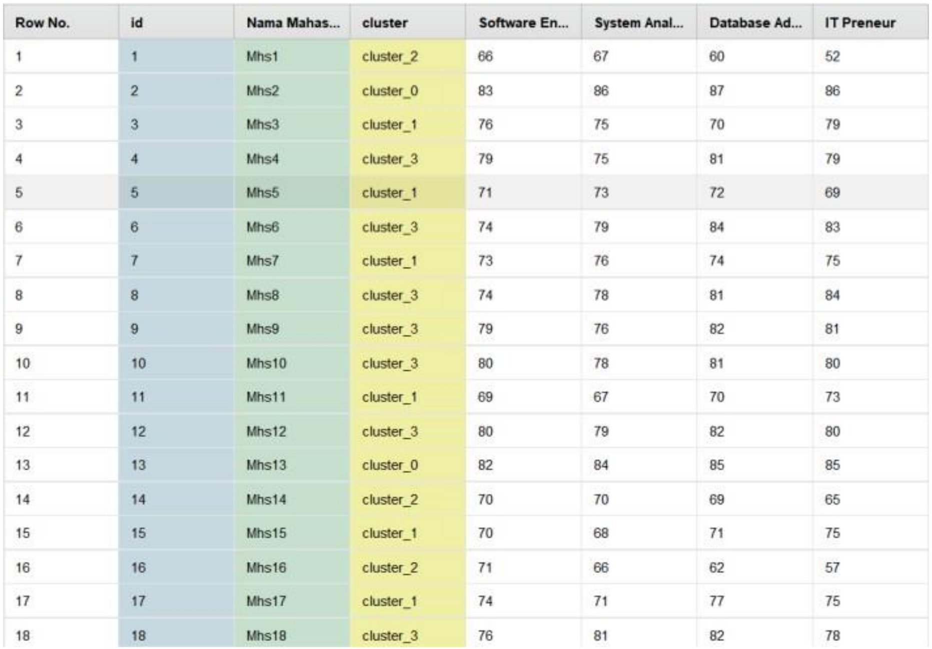
After designing the process scheme, set the number of k according to the desired process (k = 4). Next,run the analysis process. Based on the process results from RapidMiner, the centroid value is in Figure 3. As well as the cluster grouping results in Figure 4.



| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 |
|---|---|---|---|---|
| Software Engineer | 82.833 | 73.143 | 70.800 | 79.258 |
| System Analyst | 81.500 | 73.952 | 69.800 | 80.065 |
| Database Administrator | 83.667 | 72.333 | 63 | 77.871 |
| IT Preneur | 85.583 | 75.810 | 58.200 | 80.387 |

**Figure 3.** Centroid results on RapidMiner

The results of the clustering conducted based on the values of centroids in each cluster can beobserved in Figure 4, which identifies the clusters against the entire dataset of student grades. The complete results of the analysis using RapidMiner are shown in Figure 5, Figure 6, and Figure 7.

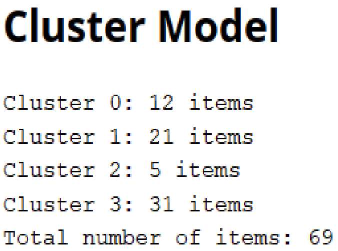


| Row No. | id | Nama Mahas... | cluster | Software En... | System Anal... | Database Ad... | IT Preneur |
|---|---|---|---|---|---|---|---|
| 1 | 1 | Mhs1 | cluster_2 | 66 | 67 | 60 | 52 |
| 2 | 2 | Mhs2 | cluster_0 | 83 | 86 | 87 | 86 |
| 3 | 3 | Mhs3 | cluster_1 | 76 | 75 | 70 | 79 |
| 4 | 4 | Mhs4 | cluster_3 | 79 | 75 | 81 | 79 |
| 5 | 5 | Mhs5 | cluster_1 | 71 | 73 | 72 | 69 |
| 6 | 6 | Mhs6 | cluster_3 | 74 | 79 | 84 | 83 |
| 7 | 7 | Mhs7 | cluster_1 | 73 | 76 | 74 | 75 |
| 8 | 8 | Mhs8 | cluster_3 | 74 | 78 | 81 | 84 |
| 9 | 9 | Mhs9 | cluster_3 | 79 | 76 | 82 | 81 |
| 10 | 10 | Mhs10 | cluster_3 | 80 | 78 | 81 | 80 |
| 11 | 11 | Mhs11 | cluster_1 | 69 | 67 | 70 | 73 |
| 12 | 12 | Mhs12 | cluster_3 | 80 | 79 | 82 | 80 |
| 13 | 13 | Mhs13 | cluster_0 | 82 | 84 | 85 | 85 |
| 14 | 14 | Mhs14 | cluster_2 | 70 | 70 | 69 | 65 |
| 15 | 15 | Mhs15 | cluster_1 | 70 | 68 | 71 | 75 |
| 16 | 16 | Mhs16 | cluster_2 | 71 | 66 | 62 | 57 |
| 17 | 17 | Mhs17 | cluster_1 | 74 | 71 | 77 | 75 |
| 18 | 18 | Mhs18 | cluster_3 | 76 | 81 | 82 | 78 |

**Figure 4.** Cluster Grouping Results

### 4.3 Result evaluation

After implementing the K-Means algorithm manually and using Rapid Miner, the same results wereobtained in both processes. The results of the division of *clusters* and the members of each *cluster* can be seen in Figure 5 where*cluster 3* (IT Preneur) has the largest distribution of members, and *cluster 2* (Database Administrator) has thesmallest distribution of members.

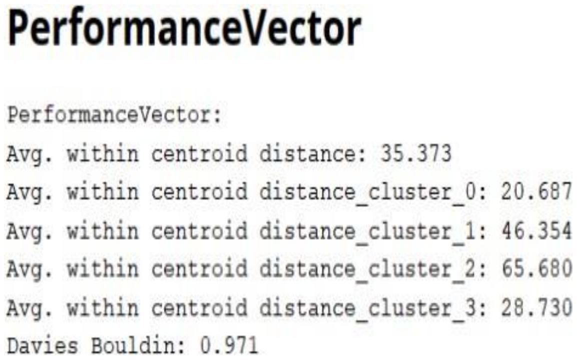


**Cluster Model**

```
Cluster 0: 12 items
Cluster 1: 21 items
Cluster 2: 5 items
Cluster 3: 31 items
Total number of items: 69
```

**Figure 5.** Cluster Distribution Results on RapidMiner

The Davies-Bouldin metric index is a measurement scale used to evaluate the quality of clustering generated by data clustering algorithms, one of which is the K-Means algorithm, thus aiding in assessing how well the clustering performs on a given dataset [5]. The smaller the Davies-Bouldin value, the better the obtained clustering result. In the Performance Vector results (as seen in Figure 6), a Davies-Bouldin index of 0.971 was obtained, indicating that each member within the cluster exhibits relatively strong similarity, as it approaches the value of



PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 35.373
Avg. within centroid distance_cluster_0: 20.687
Avg. within centroid distance_cluster_1: 46.354
Avg. within centroid distance_cluster_2: 65.680
Avg. within centroid distance_cluster_3: 28.730
Davies Bouldin: 0.971
```

**Figure 6**. Performance Vector results on RapidMiner

A graphic plot of the distribution of data sets in each cluster can be seen in Figure 7. The clusters represent students' areas of expertise, including:

1. Cluster_0 represents Software Engineering
2. Cluster_1 represents System Anlayst
3. Cluster_2 represents Database Administrator
4. Cluster_3 represents IT Entrepreneur

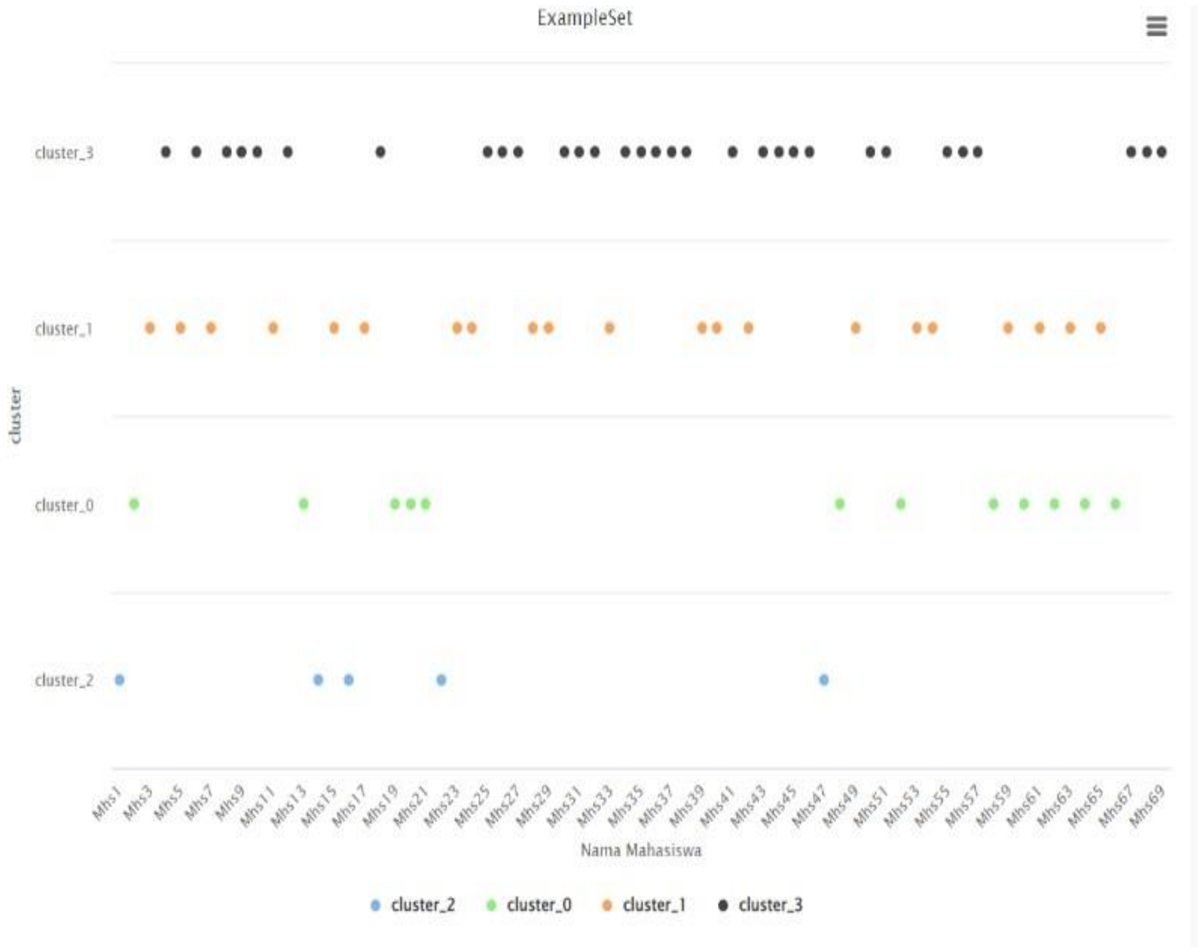


**Figure 7.** Scatter Plot Clustering Graph

## 5. CONCLUSSION

Based on the results and discussion that has been done above, it can be concluded that the predictiondata for grouping student expertise can be grouped into 4 clusters, namely:

1. Cluster 1: 12 students have Software Engineer skills.
2. Cluster 2: 21 students have System Analyst skills.
3. Cluster 3: 5 students have Database Administrator skills.
4. Cluster 4: 31 students have IT Entrepreneur skills.

The result is greatly influenced by the initial centroid value, so if you have larger and more data, you need the right centroid value. The results of grouping students can then be used as a reference for decisions for students and the campus in developing this area of expertise. For further research, it is necessary to expand the criteria for the field of study for other semesters toget good decision results. The K-Means algorithm is considered quite accurate in grouping, but it can be developed by implementing it with other clustering algorithms so that there is a comparison of results in order to obtain even more accurate data.

## REFRENCES

[1] Yusra, D. Olivita, and Dkk., “Perbandingan Klasifikasi Tugas Akhir Mahasiswa Jurusan Teknik Informatika Menggunakan Metode Naïve Bayes Classifier dan K-Nearest Neighbor,” *J. Sains, Teknol. dan Ind.*, vol. 14, no. 1, pp. 79–85, 2016.

[2] D. Kuswandi, E. Surahman, Z. Zufar At Thaariq, and M. Muthmainnah, “K-Means Clustering of Student Perceptions on Project-Based Learning Model Application,” in *2018 4th International Conference on Education and Technology (ICET)*, 2018, pp. 9–12, doi: 10.1109/ICEAT.2018.8693932.

[3] F. Istighfar, A. B. P. Negara, and T. Tursina, “Klasifikasi Bidang Keahlian Mahasiswa Menggunakan Algoritma Naive Bayes,” *J. Sist. dan Teknol. Inf.*, vol. 11, no. 1, p. 77, 2023, doi: 10.26418/justin.v11i1.52402.

[4] N. Fuad, “Algoritma Fuzzy Naive Bayes Untuk Mengklasifikasikan Bidang Keahlian Mahasiswa Teknik Informatika Universitas Islam Lamongan,” *Joutica*, vol. 4, no. 2, p. 302, 2019, doi: 10.30736/jti.v4i2.330.

[5] C. Nas, “Data Mining Pengelompokan Bidang Keahlian Mahasiswa Menggunakan Algoritma K-Means (Studi Kasus : Universitas Cic Cirebon),” *Syntax J. Inform.*, vol. 9, no. 1, pp. 1–14, 2020, doi: 10.35706/syji.v9i1.3472.

[6] H. Sulastri and A. I. Gufroni, “Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia,” *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 299–305, 2017, doi: 10.25077/teknosi.v3i2.2017.299-305.

[7] Y. D. Darmi and A. Setiawan, “Penerapan Metode Clustering K-Means Dalam Pengelompokan Penjualan Produk,” *J. Media Infotama*, vol. 12, no. 2, pp. 148–157, 2017, doi: 10.37676/jmi.v12i2.418.

[8] S. M. Mr and A. L. Caroline, “The Study on Clustering Analysis in Data Mining,” *Int. J. Data Min. Tech. Appl.*, vol. 8, no. 1, pp. 46–49, 2019, doi: 10.20894/ijdmta.102.008.001.011.

[9] F. P. Dewi, P. S. Aryni, and Y. Umaidah, “Implementasi Algoritma K-Means Clustering Seleksi Siswa Berprestasi Berdasarkan Keaktifan dalam Proses Pembelajaran,” *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 7, no. 2, pp. 111–121, 2022, doi: 10.14421/jiska.2022.7.2.111-121.

[10] M. R. Muttaqin and M. Defriani, “Algoritma K-Means untuk Pengelompokan Topik Skripsi Mahasiswa,” *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 121–129, 2020, doi: 10.33096/ilkom.v12i2.542.121-129.

[11] D. Zakiyah, N. Merlina, and N. A. Mayangky, “Penerapan Algoritma K-Means Clustering Untuk Mengetahui Kemampuan Karyawan IT,” *Comput. Sci.*, vol. 2, no. 1, pp. 59–67, 2022, doi: 10.31294/coscience.v2i1.623.

[12] A. E. Wibowo and T. Habanabakize, “K-Means Clustering untuk Klasifikasi Standar Kualifikasi Pendidikan dan Pengalaman Kerja Guru SMK di Indonesia,” *J. Din. Vokasional Tek. Mesin*, vol. 7, no. 2, pp. 152–163, 2022, doi: 10.21831/dinamika.v7i2.53848.

[13] H. Shen and Z. Duan, “Application Research of Clustsering Algorithm Based on K-Means in Data Mining,” in *2020 International Conference on Computer Information and Big Data Applications (CIBDA)*, 2020, pp. 66–69, doi: 10.1109/CIBDA50819.2020.00023.

[14] A. R. Lubis, S. Prayudani, Y. Fatmi, and O. Nugroho, “Classifying News Based on Indonesian News Using LightGBM,” in *2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, Nov. 2022, pp. 162–166, doi: 10.1109/CENIM56801.2022.10037401.

[15] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, “The feature extraction for classifying words on social media with the Naïve Bayes algorithm,” *IAES Int. J. Artif. Intell.*, vol. 11, no. 3, pp. 1041–1048, 2022, doi: 10.11591/ijai.v11.i3.pp1041-1048.

[16] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, “Spelling Checking with Deep Learning Model in Analysis of Tweet Data for Word Classification Process,” in *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2022, pp. 343–348, doi: 10.23919/EECSI56542.2022.9946476.

[17] T. Widiyaningtyas, M. I. W. Prabowo, and M. A. M. Pratama, “Implementation of K-means clustering method to distribution of high school teachers,” in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2017, pp. 1–6, doi: 10.1109/EECSI.2017.8239083.

[18] M. Mardi and M. R. Keyvanpour, “GBKM: A New Genetic Based K-Means Clustering Algorithm,” in *2021 7th International Conference on Web Research (ICWR)*, 2021, pp. 222–226, doi: 10.1109/ICWR51868.2021.9443113.

[19] W. Purba, S. Tamba, and J. Saragih, “The effect of mining data k-means clustering toward students profile model drop out potential,” *J. Phys. Conf. Ser.*, vol. 1007, no. 1, 2018, doi: 10.1088/1742-6596/1007/1/012049.

[20] R. Baruri, A. Ghosh, R. Banerjee, A. Das, A. Mandal, and T. Halder, “An Empirical Evaluation of k-Means Clustering Technique and Comparison,” in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019,

pp. 470–475, doi: 10.1109/COMITCon.2019.8862215

[21] S. D. Prasetiani and N. Rochmawati, “Penerapan Data Mining Untuk Clustering Menu Favorit Menggunakan Algoritma K-Means (Studi Kasus Kedai Expo),” *J. Informatics Comput. Sci.*, vol. 3, no. 03, pp. 278–286, 2022, doi: 10.26740/jinacs.v3n03.p278-286.

[22] R. Ahuja, A. Solanki, and A. Nayyar, “Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor,” in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2019, pp. 263–268, doi: 10.1109/CONFLUENCE.2019.8776969.

[23] Yahya and W. P. Hidayanti, “Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Efektivitas Penjualan Vape ( Rokok El ektrik ) pada ‘ Lombok Vape On ,’” *J. Inform. dan Teknol.*, vol. 3, no. 2, pp. 104–114, 2020.