Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 23, No. 1, November 2023, pp. 143∼154
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: 10.30812/matrik.v23i1.3341

Feature Selection on Grouping Students Into Lab Specializations for the Final Project Using Fuzzy C-Means

Indradi Rahmatullah , Gibran Satya Nugraha , Arik Aranta
Universitas Mataram, Mataram, Indonesia

Article Info

Article history:

Received September 02, 2023
Revised September 14, 2023
Accepted November 01, 2023

Keywords:

Cluster
Fuzzy C-Means
Silhouette Coefficient
Pearson Correlation
Principal Component Analysis

ABSTRACT

The student's Final Project is critical as a requirement to graduate from the University. In the PSTI at Mataram University, each student is required to choose a specialization lab to focus on the final project topic that they will work on. From the questionnaire, 57.7% of students answered that it is difficult to select a lab, and others answered that they prefer to determine the labs based on the grades of the courses that represent each lab. This research aimed to group and analyze students in the final project specialization lab by using the main method, namely Fuzzy C-Means (FCM). The methods used were FCM for clustering, Silhouette Coefficient for analysis of cluster quality results, Pearson Correlation, and Principal Component Analysis for the feature selection processing. The results of this study showed that the FCM method followed by a method for feature selection has better results than previous studies that used the K-Means method without feature selection; with this research result using 131 data, the cluster validation result is 0.501, after feature selection using Pearson correlation is 0.534. Thus, Fuzzy C-Means followed by the right feature selection method can group students into specialization laboratories with good results and can be further developed.

Corresponding Author:

Gibran Satya Nugraha, +6281325259291,
Department of Informatics Engineering,
University of Mataram, Mataram, Indonesia,
Email: gibransn@unram.ac.id

How to Cite:
I. Rahmatullah, G. Nugraha, and A. Aranta, "Feature Selection on Grouping Students Into Lab Specializations for the Final Project Using Fuzzy C-Means", *MATRIK: Jurnal Managemen,Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 1, pp. 143-154, Nov, 2023.

1. INTRODUCTION

Universities are one example of an educational institution that is a place for scientists to work to encourage economic, social, cultural, and technological development. Higher Education is expected to be able to print its graduates by providing not only knowledge but also expertise, skills, and competencies so that they can have a creative, innovative spirit, and entrepreneurial spirit in the era of globalization [1]. The Informatics Engineering Department (PSTI) is one of Mataram University's departments tasked with implementing higher Education in the field of Informatics Engineering. PSTI should produce computer graduates with integrity, entrepreneurial insight, and competence experience in their areas so they can later compete at the national, regional, and even international levels. To accomplish this, PSTI provides students with various facilities and infrastructure, especially laboratories (Labs). PSTI has 3 Research Laboratories: Artificial Intelligence and Its Application Laboratory, Embedded System Laboratory, and Enterprise System Laboratory. In the research lab, students work on research or thesis as a condition for their graduation.

The Final Project (TA) is one of the requirements that students must take to obtain a Bachelor of Strata-1 (S1) degree, as stated in the University of Mataram Rector Regulation Number 3 of 2020 concerning Academic Guidelines for the University of Mataram Article 23. The final project must be completed so students can implement the theory in solving a problem by their respective scientific studies. In preparing this final project, students are expected to be able to carry it out with good preparation and by the abilities possessed by the students themselves [2]. The selection of the specialization of the final project topic determines the smoothness of the final project completion process [3]. For example, students must know their concentration path or lab major. They only realized a specific concentration after arriving at the end of the lecture. Whereas at the beginning of the college, it is an important part of determining the actual concentration [4].

Several problems are encountered when choosing a concentration or major in the lab, and some students sometimes have difficulty choosing to take or enter the appropriate lab. From a questionnaire that researchers distributed to 52 random PSTI student respondents from the Class of 2017 - 2019, the results were that 57.7% felt difficulty in choosing a lab compared to those who had little difficulty (34.6%) or had no difficulty at all (7.7%) in choosing a lab. Another thing that researchers got from the results of the questionnaire is that PSTI students tend to choose labs according to the results of courses related to these labs, amounting to 67.3% compared to lecturer recommendations 57.6%, based on passion 53.9%, and following friends' suggestions 25.1%. This proves that choosing a lab based on the grades of courses related to the lab can be an option for students to determine final project topics.

However, it becomes difficult for students if they choose a lab that aligns with their courses. The reason is that there are no guidelines for subjects required to enter specific labs, and many issues have various values. Research is needed to analyze and find patterns or knowledge from a set of problems to provide recommendations to lab students that should be chosen based on the value of the courses that have been taken. Finding patterns or knowledge in the case (unsupervised) as described above can be solved by grouping students who have taken the final project to explore the patterns and knowledge for the subjects taken so that it will be seen what courses are related to these labs and the distribution of their values can be analyzed.

This research uses the Fuzzy C-Means (FCM) method to recommend lab placement based on subject grades. FCM is a data clustering technique in which the existence of each data point in a cluster is determined by the degree of membership. This allows data to be included in several clusters according to the degree of membership that represents its eligibility [5]. Validation of this algorithm will use external information in the form of subjects related to each lab and determined by each lab head. Research [6] used the FCM method to cluster students' final project specialization with a dataset of 126 students divided into 4 clusters and obtained from FCM result from Informatic Engineering is 16 students in cluster 1, 11 students in cluster 2, 14 students in cluster 3, 22 students in cluster 4 and from Informatic System is 22 students in cluster 1, 17 students in cluster 2, 16 students in cluster 3, 8 students in cluster 4. Research [7] using the K-Means method to determine major concentration for informatics students using the k-means method at the Asian Institute of Malang using a sample of student data from 2017 graduates, data divided into 3 clusters and obtained the initial and final centroid of the first attribute is 5.83%, the second attribute is 31.44%, the third attribute is 35.89%. Research [8] used the FCM method to determine the specialization of primary selection in high school using 42 student data divided into 3 clusters and obtained an accuracy of 78.6%.

Research [9] using the K-Means method to group final project recommendations using a dataset in the form of transcripts of students who have graduated totaling 488, which are divided into 3 clusters and obtained accuracy results using the silhouette coefficient method of 0.5852%. Research [10] uses the Fuzzy AHP method to create a decision support system for thesis topic recommendations using a dataset of transcripts of students entering semester 6. The most significant weight result is 1.410, and the smallest is 0.08. Research [11] used the FCM method to cluster the basketball player position with parameter data from the height, weight, age, and body mass index of 23 players. It used a manual combination of featured selection where the accuracy result obtained with all features is the accuracy result is 0.8696; after feature selection the accuracy result is 0.9565.

The difference between this research and the previous research is that the previous research discussed the outline of the implementation of FCM along with comparing accuracy before and after the implementation of FCM. Meanwhile, this research will focus

on adding additional trials or further testing by performing feature selection on the data parameters used and analyzing the cluster results generated from the FCM process before and after performing feature selection using a method. Based on previous research references, this research aims to group and analyze students in the final project specialization lab by building a model that implements the Fuzzy C-Means method based on course grades and feature selection. This research is carried out in order, starting with the introduction section discussing the background of the problem and the research method section discussing the methods used to find solutions. The results and analysis section discusses the system design results, and the closing section discusses the conclusions and suggestions from this research.

2. RESEARCH METHOD

This research aims to create a clustering model to group students into specialization labs at PSTI using FCM from the graduated student transcript dataset, select features from the parameters used, namely mandatory courses, and compare the cluster results using the Silhouette Coefficient method. The method used in this research can be seen in Figure 1.

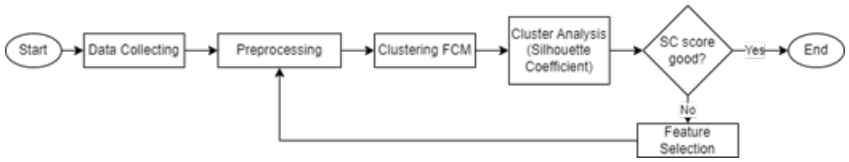

Figure 1. Research Flowchart

2.1. Data collecting

The research begins by collecting data on compulsory courses that represent each lab through a questionnaire made using Google form, the google form contains a selection of mandatory courses based on the 2020 Mataram University Informatics Engineering education guidelines, which each lab head fills indirectly, which compulsory courses represent the lab they are in charge. Later, these courses will become parameters or features of the model to be built. Then, the processing data in the form of transcripts of students who have graduated from the study program totals 331 data. The course feature can be seen in Table 1.

Table 1. Courses Per Lab

Lab 1	Lab 2	Lab 3
Logika Informatika	Pengantar Teknologi Informasi	Tata Tulis Karya Ilmiah
Sistem Digital	Sistem Digital	Proyek Perangkat Lunak
Algoritma dan Pemrograman	Organisasi dan Arsitektur Komputer	Pengantar Teknologi Informasi
Probabilitas dan Statistika	Sistem Operasi	Logika Informatika
Metode Numerik	Jaringan Komputer	Algoritma dan Pemrograman
Matematika Diskrit	Keamanan Teknologi Informasi	Probabilitas dan Statistika
Aljabar Linier	Big Data	Metode Numerik
Algoritma dan Struktur Data	Internet of Things (IoT)	Matematika Diskrit
Pengolahan Citra Digital	Pemodelan dan Simulasi	Algoritma dan Struktur Data
Kecerdasan Buatan	Sistem Terdistribusi	Pengolahan Citra Digital
Riset Teknologi Informasi	Pemrosesan Paralel	Sistem Informasi
Teori Bahasa dan Automata	Pemrograman Bergerak	Sistem Basis Data
Big Data		Jaringan Komputer
Pemodelan dan Simulasi		Analisis dan Perancangan Berorientasi Objek
Logika Fuzzy		Rekayasa Perangkat Lunak
Jaringan Syaraf Tiruan		Pemrograman Web
		Kecerdasan Buatan
		Pemrograman Berorientasi Objek
		Keamanan Teknologi Informasi
		Riset Teknologi Informasi
		Teori Bahasa dan Automata
		Big Data
		Pemrograman Visual
		Komputer dan Masyarakat
		Jaringan Syaraf Tiruan

2.2. Preprocessing

Before clustering, data preprocessing is carried out, converting letter grades from student transcripts into numerical values based on the conversion of grades at the University. The example can be seen in Table 2 and Figure 2. Filling in empty values with the average grade per course, the example can be seen in Figure 3.

Table 2. Grade Conversion

Letter grades	Number Grades
A & B+	4
B & C+	3
C & D+	2
D	1
E	0.0

Pemrosesan Paralel	Pengolahan Citra Digital	Kecerdasan Buatan	Pemrograman Berorientasi	Keamanan Teknologi Informasi	Riset Teknologi Informasi	Teori Bahasa dan Otomata	Big Data
-	A	B+	A	A	A	A	A
-	B+	A	B+	A	A	A	A
A	A	A	A	A	A	B+	A
A	A	A	A	A	A	B+	A
NA	B	A	C+	B	B+	B	B+
NA	B+	B+	B	A	B	C+	B+
B+	B	B	C+	C+	B+	C+	B+
NA	A	A	A	B+	B+	B+	B+
B+	B+	A	C	B	B+	D+	B+
C+	B+	B	D+	B	B+	C+	B+
A	A	A	B+	A	A	A	A
A	A	B+	B	-	B	-	A

Raw Data

Pemrosesan Paralel	Pengolahan Citra Digital	Kecerdasan Buatan	Pemrograman Berorientasi Objek	Riset Teknologi Informasi	Big Data	Pemodelan dan Simulasi	Logika Fuzzy	Sistem Terdistribusi
#N/A	4	4	4	4	4	4	4	2
#N/A	4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	4	4
#N/A	3	4	3	4	4	4	2	2
#N/A	4	4	3	3	4	3	#N/A	3
4	3	3	3	4	4	4	#N/A	2
#N/A	4	4	4	4	4	4	3	4
4	4	4	2	4	4	1	3	4
3	4	3	2	4	4	4	#N/A	2
4	4	4	4	4	4	4	4	4

Converted Data

Figure 2. Illustration Example Conversion Data

Pemrosesan Paralel	Pengolahan Citra Digital	Kecerdasan Buatan	Pemrograman Berorientasi Objek	Riset Teknologi Informasi	Big Data	Pemodelan dan Simulasi	Logika Fuzzy	Sistem Terdistribusi	Pemrograman Visual	Proyek Perangkat Lunak
3,851852	4	4	4	4	4	4	4	2	3	4
3,851852	4	4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	4	4	4	4
3,851852	3	4	3	4	4	4	2	2	3	3
3,851852	4	4	3	3	4	3	3,452	3	2	4
4	3	3	3	4	4	4	3,452	2	3	4
3,851852	4	4	4	4	4	4	3	4	3	3
4	4	4	2	4	4	1	3	4	4	3
3	4	3	2	4	4	4	3,452	2	3	1
4	4	4	4	4	4	4	4	4	4	3
4	4	4	3	3	4	3,579755	3,452	4	3,217125	3,429003
4	4	4	4	4	4	4	3,452	4	3,217125	4

Figure 3. Illustration Example Fill Missing Value

2.3. Clustering FCM

Fuzzy C-Means (FCM) is one of the clustering methods that is part of the Hard K-Means method. FCM uses a fuzzy clustering model so that data can be a member of all classes or clusters formed with different degrees or membership levels between 0 and The level of data presence in a class or cluster is determined by its membership degree [5].

2.4. Cluster analysis

Cluster analysis using the Silhouette Coefficient method, one of the internal criteria-based validation measures. Silhouette coefficient will evaluate the placement of each object in each cluster by comparing the average distance of objects in one cluster and the distance between objects with different clusters [12]. This is the equation for the silhouette coefficient [13]:

$$SC = \frac{1}{n}\Sigma_{i=1}^{n} s(i) \quad (1)$$

The criteria for silhouette coefficient measurement can be seen in Table 3.

Table 3. Silhouette Coefficient Criteria

Score	Interpretation
0,71 1,0	Structure strong
0,52 0,70	Structure good
0,26 0,50	Structure weak
< 0,25	Structure bad

2.5. Feature selection

Feature selection is done if the result of the silhouette coefficient is not good or below the value of 0.50, such as the process in Figure 1. This feature selection uses two methods to find and compare which features can produce optimal clusters: Pearson correlation and principal component analysis. Pearson Correlation Coefficient is just a measure of linear correlation between two sets of data [14]. The resulting value in Pearson's lies in [-1;1], for a value of -1, which means a perfect negative correlation (as one variable increases, the other decreases), +1 means a perfect positive correlation, and 0, which means there is no linear correlation between the two variables [15]. Additionally, Principal Component Analysis (PCA) is an algorithm to reduce dimensions by converting a collection of correlated dimensions into uncorrelated dimensions. This algorithm will produce a value called the Principal Component (PC) [16].

3. RESULT AND ANALYSIS

This research was performed using Python language with Jupyter Notebook. Implementation in Python using skfuzzy for FCM clustering process and sklearn for Silhouette Coefficient, Pearson Correlation, Principal Component Analysis process, and other supported libraries.

3.1. Full Data Result

The clustering results of using full data with FCM and the original student lab choices can be seen in Table 4, where cluster 0 is Lab 2 and Lab 3, cluster 1 is Lab 2 and Lab 1, and cluster 2 is Lab 2 and Lab 1. The mapping is done by finding the highest average value per cluster in each lab, and the max value can be seen in Table 5.

Table 4. FCM Full Data Clustering Result

Students	Clustering Result		Original Lab
	Cluster	Lab Recommendations	
1	0	Lab 2, Lab 3	Lab 3
2	2	Lab 2, Lab 1	Lab 1
3	1	Lab 2, Lab 1	Lab 2
4	1	Lab 2, Lab 1	Lab 2
5	2	Lab 2, Lab 1	Lab 3
6	0	Lab 2, Lab 3	Lab 3
7	2	Lab 2, Lab 1	Lab 1
8	1	Lab 2, Lab 1	Lab 1
9	2	Lab 2, Lab 1	Lab 1
10	1	Lab 2, Lab 1	Lab 2
11	0	Lab 2, Lab 3	Lab 3
12	2	Lab 2, Lab 1	Lab 3
13	1	Lab 2, Lab 1	Lab 2

Students	Clustering Result		Original Lab
	Cluster	Lab Recommendations	
14	2	Lab 2, Lab 1	Lab 2
15	1	Lab 2, Lab 1	Lab 2
⋮	⋮	⋮	⋮
330	2	Lab 2, Lab 1	Lab 3
331	1	Lab 2, Lab 1	Lab 2

Table 5. Average Value Per Lab Courses Full Data

Cluster	Lab 1	Lab 2	Lab 3
Cluster 0	2.792228	2.972934	2.868786
Cluster 1	3.289496	3.452791	3.246770
Cluster 2	3.784215	3.871287	3.747643

From Table 4, each data has two lab recommendations. This can be attributed to cluster analysis results using the silhouette coefficient method with a result of 0.49. This value falls into the third category, which means that according to the explanation in Table 1, the FCM cluster results have a weak structure.

Table 6. Pearson Correlation Feature Selection Result

Lab 1	Lab 2	Lab 3
Logika Informatika	Sistem Digital	Proyek Perangkat Lunak
Sistem Digital	Organisasi dan Arsitektur Komputer	Logika Informatika
Algoritma dan Pemrograman	Jaringan Komputer	Algoritma dan Pemrograman
Metode Numerik	Keamanan Teknologi Informasi	Metode Numerik
Matematika Diskrit	Pemodelan dan Simulasi	Matematika Diskrit
Algoritma dan Struktur Data	Pemrograman Bergerak	Algoritma dan Struktur Data
Kecerdasan Buatan		Sistem Informasi
Teori Bahasa dan Automata		Sistem Basis Data
Pemodelan dan Simulasi		Jaringan Komputer
Logika Fuzzy		Analisis dan Perancangan Berorientasi Objek
		Pemrograman Web
		Kecerdasan Buatan
		Pemrograman Berorientasi Objek
		Keamanan Teknologi Informasi
		Teori Bahasa dan Automata
		Pemrograman Visual

Table 6 is the result of feature selection using Pearson correlation in the form of a list of courses from each lab that have been selected using a threshold of 0.4. Using the features listed in Table 6, clustering is done again using FCM. The silhouette coefficient result is still the same as before the feature selection, which is 0.49. The cluster data results are mostly the same.

Table 7. PCA Feature Selection Result

PC	Lab 1	Lab 2	Lab 3
1	-	-	-
2	Pengolahan Citra Digital (-0.49), Big Data (-0.54), Pemodelan dan Simulasi (0.47)	Pemodelan dan Simulasi (0.45), Pemrosesan Paralel (-0.68)	Big Data (-0.46), Pemrograman Visual (0.40)
3	Riset Teknologi Informasi (0.57), Jaringan Syaraf Tiruan (-0.43)	Big Data (-0.57), Pemrosesan Paralel (0.41)	Jaringan Syaraf Tiruan (-0.50)
4	Pengolahan Citra Digital (-0.48)	Sistem Operasi (-0.50), Sistem Terdistribusi (-0.55)	Pengolahan Citra Digital (-0.53)

PC	Lab 1	Lab 2	Lab 3
5	Jaringan Syaraf Tiruan (-0.76)	Big Data (-0.60), Internet of Things (IoT) (0.62)	Riset Teknologi Informasi (0.56), Jaringan Syaraf Tiruan (-0.46)
6	Probabilitas dan Statistika (-0.82)	Pengantar Teknologi Informasi (0.84)	Pengantar Teknologi Informasi (-0.49), Probabilitas dan Statistika (-0.62), Komputer dan Masyarakat (-0.40)
7	Big Data (0.67)	Sistem Digital (-0.56), Sistem Operasi (-0.49),	Tata Tulis Karya Ilmiah (-0.63)
8	Teori Bahasa dan Automata (-0.75)	Pemrograman Bergerak (0.47) Sistem Terdistribusi (0.62)	Komputer dan Masyarakat (-0.53)
9	Aljabar Linier (-0.54)	Organisasi dan Arsitektur Komputer (-0.43), Jaringan Komputer (-0.48),	Jaringan Syaraf Tiruan (-0.42)
10	Sistem Digital (-0.58), Logika Fuzzy (0.42)	Internet of Things (IoT) (0.44) Pemodelan dan Simulasi (-0.59), Pemrosesan Paralel (-0.45)	-
11	Logika Informatika (-0.58)	Sistem Digital (0.59), Keamanan Teknologi Informasi (-0.64)	Proyek Perangkat Lunak (-0.46)
12	Matematika Diskrit (-0.41), Pemodelan dan Simulasi (0.45)	Jaringan Komputer (-0.64)	Rekayasa Perangkat Lunak (0.54)
13	Algoritma dan Pemrograman (0.50), Aljabar Linier (0.49)		Teori Bahasa dan Automata (0.57)
14	Kecerdasan Buatan (0.52)		Pengantar Teknologi Informasi (-0.56)
15	Metode Numerik (0.74)		Tata Tulis Karya Ilmiah (-0.40), Logika Informatika (-0.42)
16	Matematika Diskrit (0.62)		Sistem Informasi (-0.52), Pemrograman Web (0.66)
17			Matematika Diskrit (0.43)
18			Algoritma dan Pemrograman (0.49), Algoritma dan Struktur Data (-0.42)
19			Proyek Perangkat Lunak (0.42), Pemrograman Visual (0.41)
20			Logika Informatika (0.57), Sistem Basis Data (-0.43)
21			Sistem Basis Data (0.42)
22			Matematika Diskrit (-0.54)
23			Kecerdasan Buatan (0.46), Keamanan Teknologi Informasi (-0.42)
24			Metode Numerik (0.46), Jaringan Komputer (-0.42),
25			Pemrograman Berorientasi Objek (0.62) Jaringan Komputer (-0.58)

Table 7 is the result of feature selection using principal component analysis in the form of a list of courses from each lab that have been selected using a threshold of 0.4, the same as the Pearson correlation method. From Table 6, it can be seen that a collection of principal components (PC) has been formed containing courses and their weight values from the PCA feature selection process. Lab 1 has 16 PCs, Lab 2 has 12 PCs, and Lab 3 has 25 PCs. Then, the course that will be used in FCM clustering is taken. The course taken is a course that has a positive weight value from each PC, so the course is obtained in Table 8. Using the courses listed in Table 8, the cluster validation result using the silhouette coefficient is 0.43, which is a poor result.

Table 8. PCA Feature Selection Full Data Result

Lab 1	Lab 2	Lab 3
Permodelan dan Simulasi	Sistem Digital	Pemrograman Visual
Riset Teknologi Informasi	Organisasi dan Arsitektur Komputer	Riset Teknologi Informasi
Big Data	Jaringan Komputer	Rekayasa Perangkat Lunak
Logika Fuzzy	Keamanan Teknologi Informasi	Teori Bahasa dan Automata
Algoritma dan Pemrograman	Pemodelan dan Simulasi	Pemrograman Web
Aljabar Linear	Pemrograman Bergerak	Matematika Diskrit
Kecerdasan Buatan		Algoritma dan Pemrograman
Metode Numerik		Proyek Perangkat Lunak
Matematika Diskrit		Logika Informatika
		Sistem Basis Data
		Kecerdasan Buatan
		Metode Numerik
		Pemrograman Berorientasi Objek

3.2. Reduced data result

Due to the poor results of FCM clustering before and after selecting parameter features using Pearson Correlation and Principal Component Analysis, according to the flowchart of this research, a trial will be carried out in the form of preprocessing again, namely data reduction by reducing the number of datasets with the criteria that the student data has many empty valuesreducing the dataset, which initially amounted to 331 data, to 131 data, namely data from generation 12 to 14 students who have similarities in empty grades in several courses due to differences with the curriculum used today. Table 9 is the result of FCM clustering after data reduction.

Table 9. FCM Reduced Data Clustering Result

Students	Clustering Result		Original Lab
	Cluster	Lab Recommendations	
1	0	Lab 2, Lab 1	Lab 1
2	1	Lab 2, Lab 3	Lab 2
3	1	Lab 2, Lab 3	Lab 3
4	2	Lab 2, Lab 1	Lab 3
5	2	Lab 2, Lab 1	Lab 1
6	2	Lab 2, Lab 1	Lab 1
7	2	Lab 2, Lab 1	Lab 2
8	2	Lab 2, Lab 1	Lab 1
9	0	Lab 2, Lab 1	Lab 1
10	2	Lab 2, Lab 1	Lab 1
⋮	⋮	⋮	⋮
130	0	Lab 2, Lab 1	Lab 2
131	0	Lab 2, Lab 1	Lab 1

Table 10. Average Value Per Lab Courses Reduced Data

Cluster	Lab 1	Lab 2	Lab 3
Cluster 0	2.792228	2.972934	2.868786
Cluster 1	3.289496	3.452791	3.246770
Cluster 2	3.784215	3.871287	3.747643

Just like in the initial cluster results, where each data has two lab recommendations, the results of cluster validation using the silhouette coefficient increased to 0.50, a difference that is not too significant, but there is a change. Next, feature selection will be repeated using Pearson correlation and principal component analysis. For the Pearson Correlation process using the reduced data, the Pearson correlation results carried out with the reduced data have the same results as the previous Pearson correlation process, which can be seen in Table 6. For the cluster validation results with the silhouette coefficient using the selected courses, the reduced data obtained a result of 0.53, including the third category in the silhouette coefficient criteria. This means there is an improvement, and the cluster structure is good enough. Similar to the previous principal component analysis process from the reduced data, principal components (PC) are formed in each lab with the same number. Table 10 is a course that has been selected based on the results of PCA based on positive weight values and will be used in FCM clustering. Using the courses listed in Table 11, the cluster validation result using the silhouette coefficient is 0.41, which is a poor result.

Table 11. PCA Feature Selection Reduced Data Result

Lab 1	Lab 2	Lab 3
Big Data	Internet of Things (IoT)	Big data
Teori Bahasa dan Automata	Big Data	Probabilitas dan Statistika
Aljabar Linear	Sistem Digital	Komputer dan Masyarakat
Probabilitas dan Statistika	Keamanan Teknologi Informasi	Tata Tulis Karya Ilmiah
Metode Numerik	Permodelan dan Simulasi	Pemrograman Visual
Pengolahan Citra Digital		Proyek Perangkat Lunak
Kecerdasan Buatan		Pengolahan Citra Digital
		Matematika Diskrit
		Sistem Informasi

3.3. Result comparison

This section will compare the results between the full amount of data and the reduced data and the feature selection process based on the cluster validation results using the silhouette coefficient. The comparison between using full data and reduced can be seen in Table 12. In using full student data (331 data) and full course parameters, the cluster validation result is 0.487. After parameter feature selection using Pearson correlation is 0.485, feature selection using principal component analysis is 0.430. In using data that has been reduced (131 data), the cluster validation result is 0.501. After parameter feature selection using Pearson correlation is 0.534, feature selection using principal component analysis is 0.413.

Table 12. Comparison Cluster Result

Total Data	Original Value	After Pearson Correlation	After Principal Component Analysis
Full Data (331 Data)	0.487	0.485	0.430
Reduced Data (131 Data)	0.501	0.534	0.413

3.4. Analysis and Discussion

In building a model for grouping students into lab specializations, supporting datasets are collected as graduating students transcripts containing the value of each course taken, and the dataset totals 331 data. This research uses the FCM method as the main clustering method by taking grades per course for the training process. This FCM model has also been widely used in other studies with the same case and has good results [6, 8]. The advantages of the FCM method are that it is more flexible, can handle fuzzy or uncertain data, and can describe complex relationships between variables compared to the Fuzzy AHP and K-Means methods [7, 9, 10]. The next process after FCM clustering is to calculate the accuracy of the clustering results. For accuracy in this study, it is replaced by a cluster result validation test using the silhouette coefficient method like research [9]. The next process is feature selection; in contrast to previous related research without feature selection [6, 8] and also with feature selection of manual parameter combinations [11], this research focuses on adding methods in feature selection, namely the Pearson Correlation and Principal Component Analysis methods.

From the test results that have been done using full student data (331 data) along with course features totaling 35 courses, which are then divided into each lab, the initial original cluster result is 0.487. Feature selection is carried out using Pearson Correlation, which results in a new combination of course features for the FCM clustering test with a result of 0.485. Feature selection is also carried out using Principal Component Analysis as a comparison with the Pearson Correlation method, and the result is 0.430. Due

to the test results with 331 data getting unfavorable results, testing was carried out again by reducing student data to 131 data, followed by the next process, namely clustering using FCM with a result of 0.501, then using the Pearson Correlation course feature selection results obtained a result of 0.534. The last feature selection using Principal Component Analysis obtained a result of 0.413. Compared to research [9], which used the K-Means method and the same case with a silhouette coefficient score of 0.4591, this study had slightly similar results before feature selection. In contrast, after feature selection, the results were better. Referring to these results, the combination of the FCM method and the feature selection method used can produce cluster groupings that are quite good or developed based on the results of the silhouette coefficient analysis.

4. CONCLUSION

Based on the research conducted on the creation of a student lab specialization feature selection model from the clustering results using the FCM method with data in the form of course grades, it can prove that the hypothesis based on the results of the questionnaire, namely choosing a lab based on the value of related courses, is possible. This research has implications; namely, the system has contributed to or helped students determine the selection of labs, thus enabling students to determine the focus in the final project and the University to produce graduates on time. This research develops compared to previous research using the K-Means method without applying feature selection. In this study, several research limitations cause the results obtained to be optimal, such as the compulsory courses used are from the old curriculum based on the 2020 academic guide, the number of datasets is relatively small, and the variation of many values is empty, courses are limited to compulsory courses only. Therefore, further research can overcome these limitations, such as using the latest academic guide references, adding datasets because there are undoubtedly new graduates, and not limited to compulsory courses, namely adding special elective courses so that it is expected to produce more optimal research results.

5. ACKNOWLEDGEMENTS

The author would like to express my deepest gratitude to our friends who supported us every step of the way, providing encouragement and motivation throughout this research journey, and very grateful to my dedicated lecturers for their guidance, expertise, and valuable feedback, improving the quality of this research.

6. DECLARATIONS

AUTHOR CONTIBUTION

The first author, the researcher, contributed to making this article, starting with the field research, ideas, experimental system, results, and conclusion. The second and third authors are an informatics engineering lecturer and a supervisor who guided the preparation of this article.

FUNDING STATEMENT

-

COMPETING INTEREST

All authors reported this article with no competing financial interests or personal conflicts.

REFERENCES

[1] E. Erdisna, M. Ridwan, and H. Syahputra, "Developing Digital Entrepreneurship Learning Model: 4-D Competencies-Based for Millennial Generation in Higher Education," *Utamax : Journal of Ultimate Research and Trends in Education*, vol. 4, no. 2, pp. 84–100, 2022.

[2] C. D. Safitri, S. Azisah, and M. J. Annur, "The Analysis of Students' Challenges To Thesis Writing at UIN Alauddin Makassar," *English Language Teaching for EFL Learners*, vol. 3, no. 2, pp. 41–53, sep 2021.

[3] J.-C. Chang, Y.-T. Wu, and J.-N. Ye, "A Study of Graduate Students' Achievement Motivation, Active Learning, and Active Confidence Based on Relevant Research," *Frontiers in Psychology*, vol. 13, no. June, pp. 1–10, jun 2022.

[4] D. Nirmala and E. P. Hendro, "Problema Dalam Memilih Judul Penelitian Kebahasaan Bagi Pemula," *Jurnal Harmoni*, vol. 5, no. 1, pp. 15–19, 2020.

[5] N. Nurfaizah and F. Fathuzaen, "Clustering Customer Data Using Fuzzy C-Means Algorithm," *PIKSEL : Penelitian Ilmu Komputer Sistem Embedded and Logic*, vol. 9, no. 1, pp. 1–14, 2021.

[6] N. A. Ningrum and M. A. Syaputra, "Penerapan Metode Fuzzy C-Means untuk Penentuan Kompetensi Mahasiswa di STMIK Dharmawacana Metro," *Journal Computer Science and Informatic Systems:J-Cosys*, vol. 2, no. 1, pp. 1–7, 2022.

[7] P. Subekti, T. D. Andini, and M. Islamiyah, "Sistem Penentuan Konsentrasi Jurusan Bagi Mahasiswa Informatika Menggunakan Metode K-Means Di Institut Asia Malang," *Jurnal Manajemen Informatika (JAMIKA)*, vol. 12, no. 1, pp. 25–39, 2022.

[8] H. K. Candra, M. Bahit, and B. Sabella, "Penerapan Metode Klustering Fuzzy C-Means Untuk Penentuan Peminatan Pemilihan Jurusan Pada Sekolah Menengah Tingkat Atas," *POSITIF : Jurnal Sistem dan Teknologi Informasi*, vol. 7, no. 2, pp. 108–119, 2021.

[9] H. Haviluddin, S. J. Patandianan, G. M. Putra, N. Puspitasari, and H. S. Pakpahan, "Implementasi Metode K-Means Untuk Pengelompokkan Rekomendasi Tugas Akhir," *Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer*, vol. 16, no. 1, pp. 13–18, mar 2021.

[10] A. Abdullah and S. Sucipto, "Sistem Pendukung Keputusan Untuk Rekomendasi Topik Skripsi Dengan Metode Fuzzy AHP," *Jurnal Transformatika*, vol. 18, no. 2, pp. 231–239, jan 2021.

[11] Antoni Erga and Y. Nataliani, "Seleksi Fitur Pada Pengelompokkan Posisi Pemain Basket Menggunakan Fuzzy C-Means," *Journal Of Information Technology and Computer Science*, vol. 3, no. 1, pp. 77–84, 2021.

[12] J. Mantik, D. Hartama, and M. Anjelita, "Analysis of Silhouette Coefficient Evaluation with Euclidean Distance in the Clustering Method (Case Study: Number of Public Schools in Indonesia)," *Jurnal Mantik*, vol. 6, no. 3, pp. 3667–3677, 2022.

[13] Winarno, "Comparison Of Clustering Levels Of The Learning Burnout Of Students Using The Fuzzy C-Means And K-Means Methods," *Jurnal Teknologi Informasi dan Pendidikan*, vol. 16, no. 1, pp. 38–53, 2023.

[14] D. E. Parson, "Research Commons at Kutztown University Hawk Mountain Raptor Migration Phenology ' s Relation to Weather Hawk Mountain Raptor Migration Phenology ' s Relation to Weather," 2023.

[15] P. Chen, F. Li, and C. Wu, "Research on Intrusion Detection Method Based on Pearson Correlation Coefficient Feature Selection Algorithm," *Journal of Physics: Conference Series*, vol. 1757, no. 1, pp. 1–10, jan 2021.

[16] R. Pujianto, Adiwijaya, and A. A. Rahmawati, "Analisis Ekstraksi Fitur Principle Component Analysis pada Klasifikasi Microarray Data Menggunakan Classification And Regression Trees," *eProceedings . . .* , vol. 6, no. 1, pp. 2368–2379, 2019.

[This page intentionally left blank.]