Web-Based Student Academic Performance Evaluation System Using Data Mining

Afifah Ilham[1]*, Iqra Aswad[1], Muhammad Niswar[1], Ady W Paundu[1] and Nisa Mardhatillah[1]

[1]Department of Informatics, Faculty of Engineering, Hasanuddin University, Makassar, Indonesia

*E-mail: afifahlhm@gmail.com

Abstract. This research aims to retrieve student academic performance profiles using data mining technology. The clustering process is implemented to the dataset of student academic performance of Informatics Undergraduate Program in Hasanuddin University, Indonesia, from the intake year 2008 to 2013. Using k-means algorithm with k = 5, the algorithm retrieves two identical profiles. The profile 1 shows the student academic performance clusters based on GPA (Grade Point Average) and length of study while the profile 2 shows the student academic performance clusters based on GPA, length of study, and total credits earned. The SSE (Sum Square Error) formula is used to measure the variation within a cluster in each profile. The smaller the SSE, the better the similarity within a cluster. The SSE of the profile 1 (4.19628) is smaller than the SSE of the profile 2 (6.34025). This results shows that the profile 1 has better similarity within a cluster compared to the profile 2.

1. Introduction

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, the Web, other information repositories, or data that are streamed into the system dynamically [1]. One of the uses of data mining technology is in the field of education. As an example, data mining is used to detect content similarity in student final projects [2]. This helps the lecturer in detecting the final projects that may contain plagiarism. Another use of data mining in education is to evaluate student academic performance.

The use of data mining technology is increasingly developing in processing data stored in academic information systems becomes knowledge. The aim is to help educational institutions determine the right policies and programs to improve the academic quality of graduates. The use of data mining technology to evaluate student academic performance is carried out by grouping students' academic performance based on certain parameters.

Grouping of student academic performance based on parameters of class test, mid test, and final test is carried out with the aim of reducing the ratio of the number of students who fail to pass a course [3]. In another study, the analysis of student academic performance using a clustering algorithm was carried out by grouping students based on historical grades, history of the number of classes graduated, number of absences, and age of students [4]. Aggregate and attendance parameters were used to classify student academic performance into three groups, namely “poor”, “average”, and

"good" [5]. In clustering students' academic performance, the k-means, k-medoids, and x-means algorithms are used to compare the results of their accuracy [6].

Case studies at the Informatics Undergraduate Program in Hasanuddin University show that the use of student academic data stored in the Academic Information System database is only limited to the use of data for individual student needs, such as evaluating student study continuity after four semesters and evaluating student eligibility to participate in academic activities such as community service program, proposal seminars, and final exams.

To help institutions determine the appropriate study program capacity building programs and policies based on the evaluation of student academic performance, in this study, knowledge extraction was carried out from data stored in the database in the form of student academic performance profiles using the k-means clustering algorithm.

2. Research Method

2.1. Preprocessing

The data source came from the Hasanuddin University Academic Information System. The data taken is the academic performance data of Informatics Engineering students from class 2008 to class 2013. Preprocessing is carried out on data from students who are unable to complete their studies within 14 semesters. Students who drop out for this reason are considered to have a study period of 20 semesters. The number 20 is chosen so that students who fall into this category can be grouped correctly.

The distribution of the dataset based on GPA (Grade Point Average) and length of study is shown in Figure 1. The distribution of the dataset based on GPA, length of study, and total credits earned is shown in Figure 2.

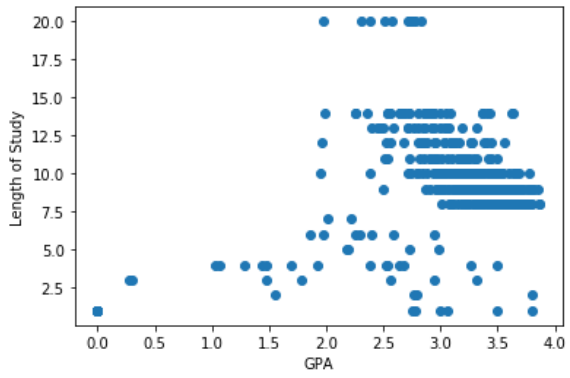

Figure 1. The distribution of the dataset based on GPA and length of study

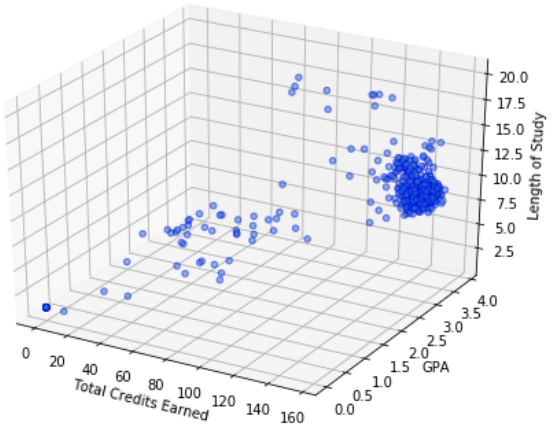

Figure 2. The distribution of the dataset based on GPA, length of study, and total credits earned

2.2. System Planning

This study will implement the k-means algorithm to cluster student academic performance. The system design begins with determining the academic performance parameters and determining the number of parameters that will be used as the basis for grouping using the k-means algorithm. The system is designed to retrieve 2 profiles of student academic performance, namely profile 1 and profile 2. Profile 1 classifies student academic performance based on parameters of GPA and length of study, while profile 2 classifies student academic performance based on GPA, length of study, and total credits earned by student.

The system is designed to be able to display the distribution of the dataset, perform the data normalization process and show the stages of the clustering process and visualization using the k-means algorithm. The k-means algorithm process starts from determining the number of clusters (*k*), determining the starting points of the cluster center (centroid) randomly as many as *k*, determining the initial membership of each centroid based on the closest distance from the data to the centroid using the Euclidean Distance equation, updating the coordinates of the centroid and membership for each centroid, and cluster assignment (determination of the end of membership for each centroid) [7].

2.2.1. Min-Max Normalization. Min-max normalization maps a value *v* from attribute *A* to *v'* into the range [*new_minA*, *new_maxA*] according to the following formula [8]:

$$v' = \frac{v - min_{\mathrm{A}}}{max_{\mathrm{A}} - min_{\mathrm{A}}}(new\_max_{\mathrm{A}} - new\_min_{\mathrm{A}}) + new\_min_{\mathrm{A}} \quad (1)$$

where:
v' = normalization result data
v = data to be normalized
min_A = the minimum value of attribute *A*
max_A = the maximum value of attribute *A*
new_min_A = the new minimum limit for attribute *A*
new_max_A = the new maximum limit for attribute *A*

2.2.2. Euclidean Distance. Euclidean distance is a distance calculation method used to measure the distance from 2 (two) points in Euclidean space (covering two-dimensional, three-dimensional, or even more euclidean planes). To measure the level of data similarity with the Euclidean distance formula, the following formula is used [9]:

$$ED(d, c) = \sqrt{\sum_{i=1}^{n}(d_i - c_i)^2} \quad (2)$$

where:
ED = distance between *x* and *y*
d = data on attribute
c = cluster center data
n = amount of data
d_i = value for each data *i*-th
c_i = the center value of the *i*-th cluster

2.2.3. Centroid. The centroid is the average of all data points in the group. Each centroid with a set of data elements is referred to as a cluster [7]. The centroid value can be determined by the following equation:

$$c\_updated = (\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i) \quad (3)$$

where:
c_updated = new centroid value
n = amount of data
x_i = *i*-th *x* data value
y_i = *i*-th *y* data value

2.3. System Implementation

The implementation of k-means is carried out in two ways, which are using the Python programming language and by using a library. The first way is to create a Python program that can show in detail and display a visualization of the stages in the clustering process using the k-means algorithm. The second way is to use the existing library in Scikit-learn. Scikit-learn is a machine learning library using the Python programming language. Flowchart of k-means algorithm implementation is shown in Figure 3.

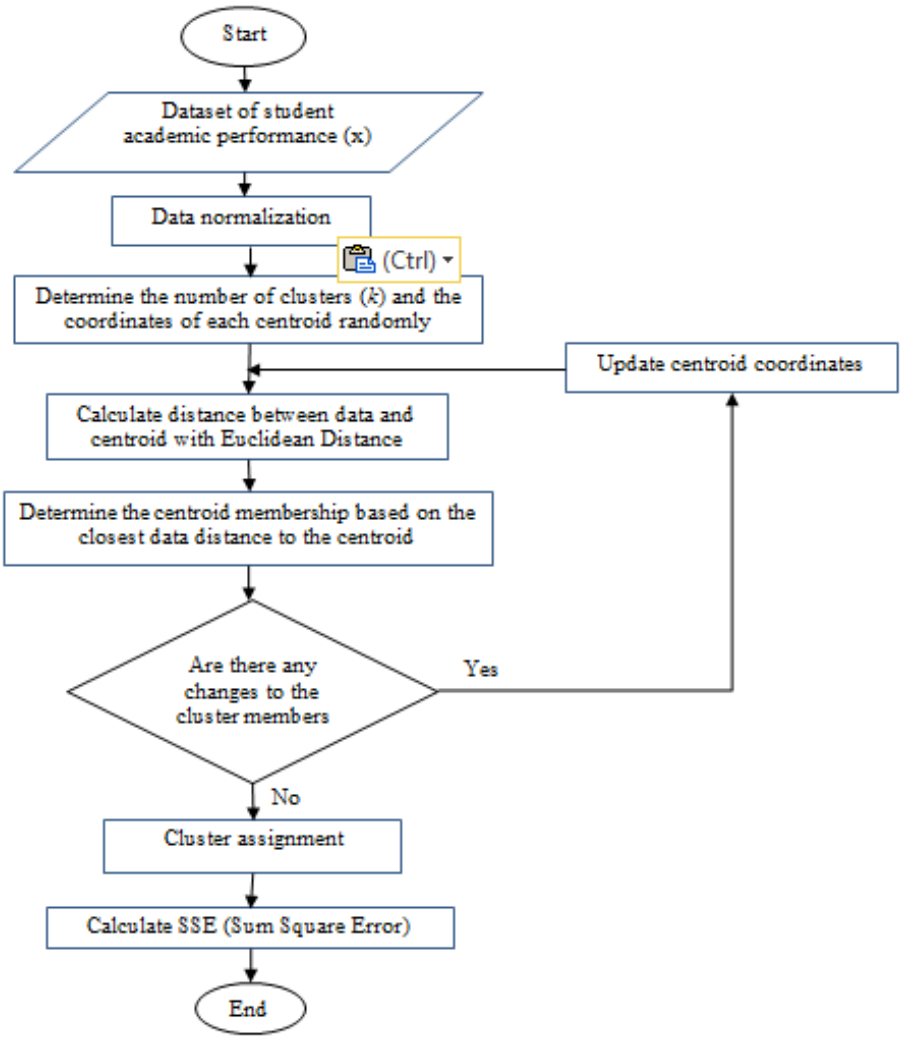

Figure 3. Flowchart of k-means algorithm implementation

2.4. Evaluation

The results of clustering can be evaluated by looking at the SSE value. The SSE value is the square value of the difference between the coordinates of the centroid to the respective data in the cluster member. Let n be the number of data in cluster d and c_i is the mean (average) value of each cluster. SSE is defined as follows [7].

$$SSE(d,c) = \sum_{i=1}^{n} (d_i - c_i)^2 \quad (4)$$

where:
d = data on attributes
c = cluster centroid data
n = amount of data
d_i = i-th data value
c_i = centroid value or average of the i-th cluster

The SSE value shows the level of similarity to each data in a cluster. A small SSE value indicates a strong similarity to each data in a cluster. The relationship between the SSE value and the k value is described in the curve of The Elbow Method [10].

2.5. Discussion

The discussion was carried out by analyzing basic statistics for each cluster. The calculated parameters are count (number of members), mean (average value), standard deviation, min (lowest value), 25% or lower quartile (Q1), 50% or median (Q2), 75% or upper quartile (Q3), and max (highest value). The evaluated statistical values can provide a more detailed description of a students' academic performance profile. For example, groups of students who graduate on time, groups of students who graduate but have a long study period, groups of students who do not graduate and leave their studies in the early stages of recovery, groups of students who are unable to complete their studies within a predetermined time, and so on. This kind of depiction of student academic performance provides better knowledge than simply describing the academic performance of students in the graduating or not graduating group.

3. Implementation and Result

Before clustering the dataset using the k-means algorithm, it is necessary to normalize the data. Data normalization is needed to get the same scale for the data used so that the calculation of the distance between data can be done correctly. Figure 4 and Figure 5 shows the distribution of normalized data.

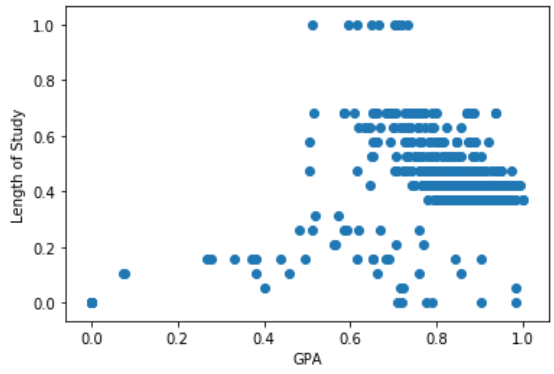

Figure 4. The distribution of normalized data of profile 1

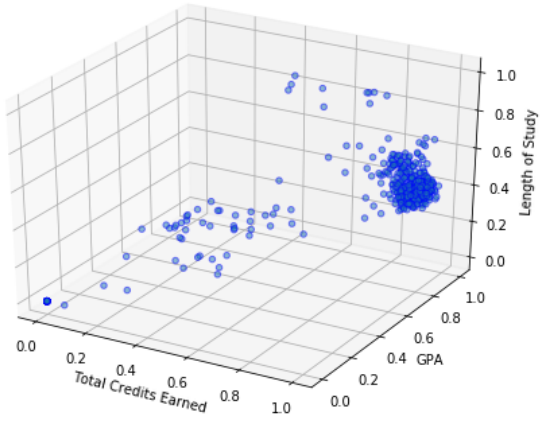

Figure 5. The distribution of normalized data of profile 2

Then, the k-means algorithm is used in grouping to create a profile of student academic performance. The profiling was done by clustering student academic performance based on GPA and length of study (profile 1) and clustering student academic performance based on GPA, length of study, and total credits earned (profile 2). Figure 6 and Figure 7 show the cluster formed in each profile with a value of k = 5, consisting of 5 groups marked with 5 different colors.

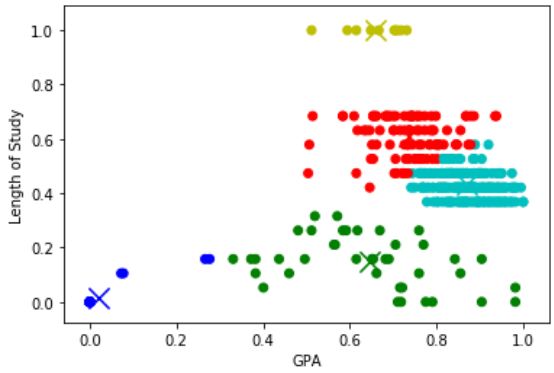

Figure 6. Clustering result of profile 1

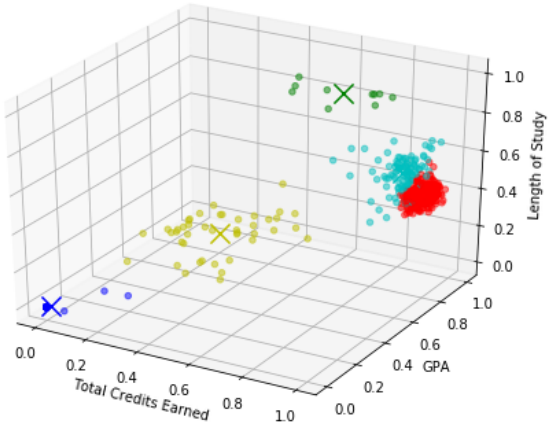

Figure 7. Clustering result of profile 2

The results of clustering are evaluated with the SSE value. For the value of $k = 5$ on profile 1 gives an SSE value of 4.196281. Meanwhile, profile 2 gives an SSE value of 6.34024956. The two profiles have given clustering results with a fairly low SSE value.

To obtain knowledge from the results of the clustering, a basic statistical analysis was carried out for each cluster. From the evaluated statistical value, an overview of the students' academic performance profile can be retrieve. The comparison of the two profiles retrieved is shown in Table 1.

Table 1. The comparison of student academic performance profile based on GPA and length of study (profile 1) with student academic performance based on GPA, length of study, and total credits earned (profile 2)

Profile 1		Profile 2		Status	Explanation
Cluster	Percentage of Members	Cluster	Percentage of Members		
1	23.0 %	5	23.0 %	Graduated at the end of study	Cluster 1 in profile 1 is identical to cluster 5 in profile 2
2	9.2 %	4	9.9 %	Leave after taking 1-7 semesters of lectures	Cluster 2 in profile 1 is identical to cluster 4 in profile 2
3	8.3 %	3	7.8 %	Leave after taking 1-4 semesters of lectures	Cluster 3 in profile 1 is identical to cluster 3 in profile 2
4	2.3 %	2	2.3 %	Leave after being unable to complete the study for 14 semesters	Cluster 4 in profile 1 is identical to cluster 2 in profile 2
5	57.2 %	1	57.0 %	Graduate on time	Cluster 5 in profile 1 is identical to cluster 1 in profile 2
Total	100 %	Total	100 %		

A web based user interface is made for the evaluation system so that it can be used by various parties easily. The application can process the dataset file according to the template file that has been prepared. The application receives input in the form of a dataset file, a choice of clustering parameters, and the number of clusters (k), as shown in Figure 8. As for the output results in the form of dataset distribution graphs, clustering results graphs, graphs of the relationship between SSE values and k values, and downloadable clustering results files, as shown in Figure 9.

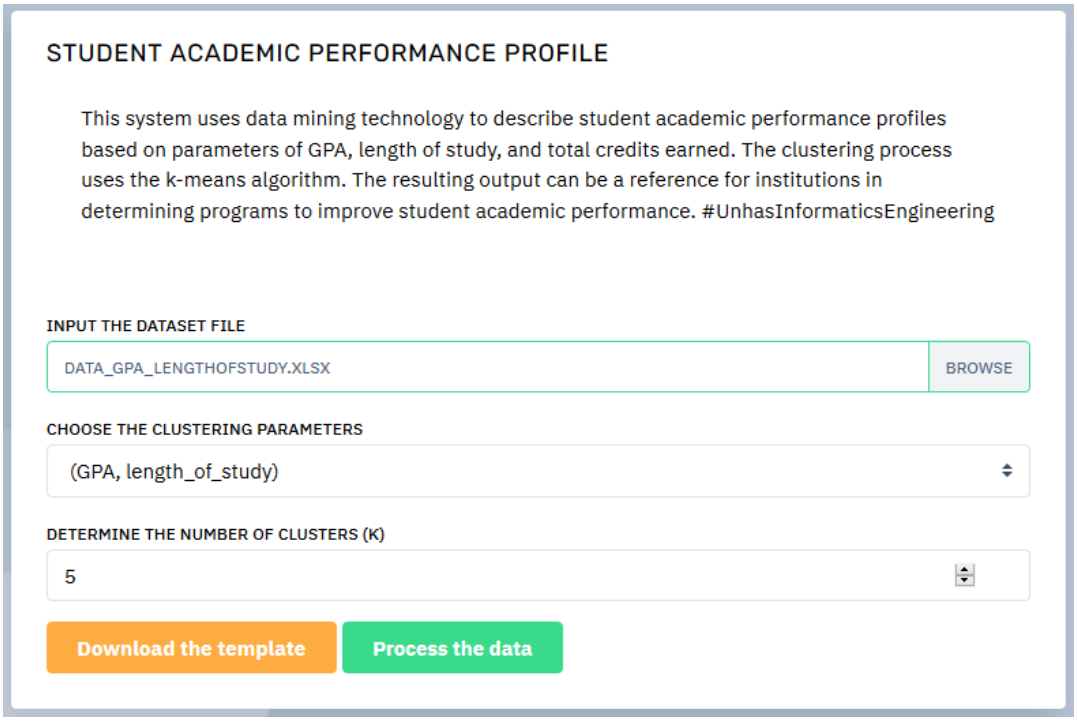

Figure 8. Application input interface

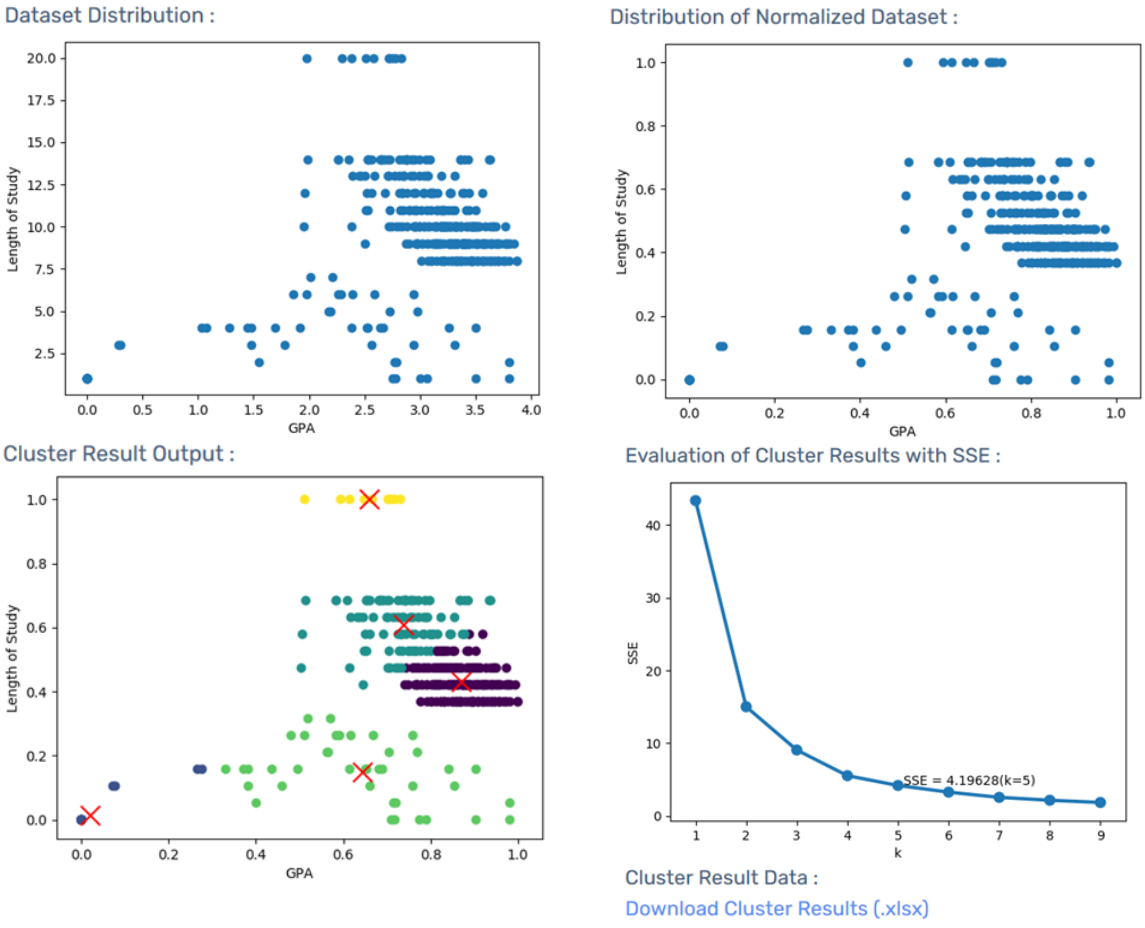

Figure 9. Application output

4. Conclusion

The use of the k-means algorithm with a value of $k = 5$ in the dataset of academic performance of Informatics Engineering students class 2008 to class 2013 describes the profile of academic performance into 5 clusters. The clusters consist of 2 clusters that consisting of graduating students and 3 clusters that consisting of students leaving or dropping out.

Clustering of student academic performance based on GPA and length of study (profile 1) provides identical results with clustering student academic performance based on GPA, length of study, and total credits earned (profile 2). Based on the SSE value, profile 1 has better similarity between cluster members than profile 2. This is indicated by the SSE value of profile 1 (SSE = 4.196281) which is smaller than the SSE value of profile 2 (SSE = 6.34024956) for the value of $k = 5$.

Acknowledgment

This research is funded by Lab Based Education Grant 2020, Faculty of Engineering, Hasanuddin University.

References

[1] Han J, Kamber M and Pei J 2011 *Data Mining : Concepts and Techniques* (Elsevier)

[2] Ilham A A, Bustamin A, Aswad I and Armin F 2020 Implementation of Clustering and Similarity Analysis for Detecting Content Similarity in Student Final Projects *IOP Conf. Ser.: Mater. Sci. Eng.* **875** 012039

[3] Borgavakar S P and Shrivastava A 2017 Evaluating Student's Performance using K-Means Clustering *Int. J. Eng. Res. Technol.* **6** pp 114-6

[4] Singh I, Sabitha A S and Bansal A 2016 Student Performance Analysis using Clustering Algorithm *2016 6th Int. Conf. - Cloud Syst. Big Data Eng.* pp 294-9.

[5] Nagesh A S and Satyamurty C V S 2018 Application of Clustering Algorithm for Analysis of Student Academic Performance *Int. J. Comput. Sci. Eng*. **6** pp 381-4

[6] Humamuddin and Nafis M 2017 Students Academic Performance Using Partitioning Clustering Algorithms *International Journal of Advanced Research in Computer Science* **8** pp 640-4

[7] Santosa B and Umam A 2018 *Data Mining dan Big Data Analytics : Teori dan Implementasi Menggunakan Python & Apache Spark* (Yogyakarta: Penebar Media Pustaka)

[8] Junaedi H, Budianto H, Maryati I and Melani Y 2011 Data Transformation pada Data Mining *Prosiding Konferensi Nasional "Inovasi dalam Desain dan Teknologi"* (Surabaya: Sekolah Tinggi Teknik Surabaya) pp 93-9

[9] Nishom M 2019 Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square *Jurnal Informatika: Jurnal Pengembangan IT* **4** pp 20-4

[10] Bholowalia P and Kumar A 2014 EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN *International Journal of Computer Applications* **105** pp 17-24