Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering for Grouping Students' Abilities in Online Learning Process

Indah Purnama Sari[1*], Al-Khowarizmi[2], Ismail Hanif Batubara[3]
[1,2]Department of Information Technology, Universitas Muhammadiyah Sumatera Utara, Indonesia
[3]Department of Mathematics Education, Universitas Muhammadiyah Sumatera Utara, Indonesia

ABSTRACT

The time of the Covid-19 pandemic, which is still unknown when it will end, has resulted in the learning and teaching process having to be carried out online. The impact of the learning process carried out online is a less satisfactory result than the learning process carried out offline. This study aims to group students based on their learning outcomes in statistics courses and measured probability based on the variables of attendance scores, assignments, midterm exams (UTS), and final semester exams (UAS) which are then used to evaluate learning for subjects that require analytical skills. good quantitative. This study uses cluster k-means analysis and fuzzy C-Means in grouping students into three groups based on their learning outcomes.

Keywords: Covid 19, Cluster Analysis, K-Means, Fuzzy C-Means, Online, Offline.

Corresponding Author:
Indah Purnama Sari
Department of Information Technology
Universitas Muhammadiyah Sumatera Utara, Indonesia
Jalan Kapten Mukhtar Basri No 3 Medan 20238.
Email : indahpurnama@umsu.ac.id

Article history:
Received Jan 1, 2021
Revised Feb 14, 2021
Accepted Mar 12, 2021

1. INTRODUCTION

One of the branches of mathematics which is no less important in its application in the field of computers is statistics. Mathematics and statistics require good quantitative skills and structured thinking skills so that if you are familiar with it, it will make it easier to make algorithms in computer science which also requires the ability to work in a very well structured manner.

The statistics course is a compulsory subject for the information technology study program at the Muhammadiyah University of North Sumatra (UMSU) programmed in the third semester. Student statistics learning is still relatively low, this is evidenced by the acquisition of an average score of 114 students, namely 69.87. Seeing that statistical learning has not been maximized, researchers feel the need to classify student abilities based on things that are assessed for further action based on the group.

There are several methods that can be used to classify students' abilities, one of which is the k-means cluster analysis. The k-means cluster method identifies objects that have certain characteristics in common, and then uses these characteristics as centroids (Nasari, 2015). This method has been widely used in various fields, such as what was done by Edmira Rivani (2010) who grouped the provinces of production of rice, maize, soybeans, and green beans. J. Ong (2013) uses cluster analysis to determine President University's marketing strategy. In addition, there is also E. Muningsih (2015) who uses k-means for online shop product clustering in determining stock items. Abroad this method is also often used,

In addition, one of the clustering methods that can be used to classify data is Fuzzy C-Means (FCM). FCM is a data grouping technique where the existence of each data point in a group (cluster) is determined by the degree of membership. By repeating the cluster center and membership value of each data, it will be found that the cluster center is headed to the right location (Kusumadewi, 2010).

K-Means Clustering is a method that attempts to partition existing data into two or more groups (Prasetyo, 2012). This method partitions data into groups (clusters) so that data with the same characteristics are included in the same cluster and different data are grouped into other clusters. The iterative concept of the FCM method is the same as the K-Means method, which is based on minimizing the objective function. With this FCM method, we will analyze the number of optimum clusters that will group corporate bond data into different clusters using the Xie Beni Index.

After analyzing with the FCM method and obtaining the optimum number of clusters based on the Xie Beni index, then conducting the evaluation stage by comparing the results of the grouping with the K-Means method. The stage of determining the best method is done by comparing the ratio of standard deviation within the cluster (Sw) to the ratio of standard deviation between clusters (Sb) in each method. The smallest Sw / Sb ratio will be selected as the best method.

2. RESEARCH METHOD

A. K-Means Algorithm

This study uses primary data derived from the record scores of third semester students in statistics and probability courses, the UMSU Information Systems study program for the 2019/2020 academic year. The number of students who became the object of research was 50 students. Some of the independent variables used in the study are in Table 1.

Table 1. Research Variables

Variable	Variable Name	Data Type
X_1	Attendance Value	Continuous
X_2	Average Assignment Score	Continuous
X_3	UTS scores	Continuous
X_4	UAS Value	Continuous

This study will classify students based on student scores in statistics and probability courses using the K-Means algorithm. K-Means is a non-hierarchical algorithm where the clustering process is based on the closest distance to the specified center point. One of the frequently used distances is Euclidean, which can be obtained by the equation two data distances that are calculated and p is the dimension of the data used. Determination of the cluster center point can be seen from the equation below.

Where :

Cm (q): center of the p-th group of variables

m	: 1, 2,……., k
nm	: Number of objects in the m-th group
k	: Number of clusters
q	: 1, 2,….., p
xi	: The observed value of the ith object of the q-variable
i	: 1, 2,…, nm

B. Fuzzy C-Means (FCM) Algorithm

FCM algorithm steps (Kusumadewi, 2010):

1. Input data to be grouped (X), in the form of a matrix of size nxp (n = number of data samples, p = variable for each data).
2. Determine the number of clusters (c), weight power (m> 1), maximum iteration (MaxIter), smallest error expected (ε), the initial objective function (P0 = 0).
3. Generate a random number Uik, where i = 1,2,3, ..., n; k = 1,2, ..., c;
4. Calculating the k-th cluster center in the j-variable; vkj, where k = 1,2, ..., c; j = 1,2, ..., p;

$$V = [v_{kj}] = \frac{\sum_{i=1}^{n}(u_{ik})^m x_{ij}}{\sum_{i=1}^{n}(u_{ik})^m} \tag{1}$$

5. Calculating the objective function value in the t-iteration (Pt) with the formula:

$$Pt = \sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik})^{m} d_{ik}\left(x_i, v_{kj}\right) \qquad (2)$$

6. Counting change matrix membership uik with formula:

$$u_{ik} = \left[\frac{\left[\sum_{j=1}^{p} d_{jk}\right]^{\frac{1}{m-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{p} d_{jk}\right]^{\frac{1}{m-1}}}\right]^{-1} \qquad (3)$$

7. Checking stop condition: If (| Pt - Pt-1 |) <ε) or (t> MaxIter) then iteration stops Otherwise, then t = t + 1, go back to step 4.

3. RESULTS AND DISCUSSION

A. Descriptive

SI FIKTI UMSU students who take statistics and probability courses are used as research objects which consist of 2 classes, namely class 2A and 2B. Descriptive statistics are used to find out how the general description of the distribution of statistics scores and student probabilities in each. In addition, it can also be seen a description of the student's ability to master the material obtained. Descriptive statistics for statistical value and probability for each class can be seen in tables 2 and 3. While table 4 is descriptive statistics to see the overall score of students.

Table 2. Descriptive Statistics Statistical Score Class 2A

	Presence	Duty	UTS	UAS
N	25	25	25	25
N missing	0	0	0	0
Average	84.64	66.25	75.51	63.86
Mode	93.75	68.33	77.25	60
Median	93.75	68.33	77.25	60
Variance	416.57	163.33	194.88	567.39
Range	100	82.67	90	100

Based on table 2, the mean statistical value and probability of class 2A is 84.64 for attendance, 66.25 for assignments, UTS is 75.51, and 63.86 for UAS. The most scores obtained in class 2A for attendance, assignments, UTS, and UAS were 93.75, 68.33, 77.25, and 60. UAS and attendance in class 2A are more diverse than assignments and UTS, because the resulting variance is very high, namely 567.39 for UAS and attendance of 416.57.The value of the assignment is more homogeneous when compared to the UAS, attendance and UTS scores, because of the variance generated the value of the assignment is the smallest, namely 163.33.

Table 3. Descriptive Statistics of Class 2B Statistical Value

	Presence	Duty	UTS	UAS
N	26	26	26	26
N missing	0	0	0	0
Average	91.25	67.48	76.41	65.93
Mode	100	68.67	78.75	60
Median	93.75	68.33	78.5	60
Variance	251.86	112.15	161.04	262.12
Range	87.5	76.04	86.75	100

Table 3 shows that the highest average of the 26 students in class 2B is for attendance scores, namely 91.25, followed by a UTS score of 76.41, assignments of 68.67, and finally a UAS of 65.93. The most scores obtained in class 2B were 100 for attendance, assignments were 68.67, UTS was 78.75, and UAS

60. 60. The UAS scores in class 2B are more diverse than the other values, because the variance for UAS is the highest, which is 262.12. Next is attendance with a variance of 251.86, UTS of 161.04, and student scores for assignments that are more homogeneous, because they have the smallest variance, namely 112.15.

Table 4. Descriptive Statistics of Statistical Value and Probability of SI UMSU Students

	Presence	Duty	UTS	UAS
N	51	51	51	51
N missing	0	0	0	0
Average	89.47	70.30	66.97	59.41
Mode	100	90	75	60
Median	93.75	68.33	77	60
Variance	319.69	275.89	382.20	592.43
Range	100	75	80	100

Table 4 shows the descriptive statistics for all classes, namely classes 2A and 2B. The mean score of statistics and probability for 51 students was 89.47 for attendance, 70.30 for assignments, UTS 66.97, and UAS 59.41. The highest scores for all students in grades 2A and 2B who take Statistics and probability courses are 100 for attendance, 90, for assignments, UTS 75, and UAS are 60.The mean score of all students obtained is 93.75 for attendance , 68.33 for assignments, UTS is 77, and UAS is 60. assignment amounting to 275.89.

B. Value Cluster Class 2A and 2B

Cluster analysis of the statistical value and probability of the UMSU Information System (SI) semester III students using the K-means algorithm with the number of clusters is 3. Cluster 1 is for the low value group (below average), cluster 2 is for the approximate average value based on the distance used from its centroid, and cluster 3 is for the high value group (above average). The results of group formation based on the Statistics and Probability scores of SI students in class 2A and 2B can be seen in tables 5 and 6.

Table 5. Cluster Centroids of Each Value

Variable	Cluster 1	Cluster 2	Cluster 3	Grand Centroid
Presence	4.17	91.44	91.89	89.47
Duty	19.13	68.92	80.29	70.30
UTS	0	44.61	76.55	66.97
UAS	0	31.48	70.5	59.41

Table 5 contains the value of the average value of the variables in each cluster of each value, namely attendance, assignment, UTS, and UAS. It can be seen that cluster 1 has a very low score, because the resulting average is very small, namely 4.17 for attendance, 19.13 for assignments, and 0 for UTS and UAS. Whereas for cluster 2 and 3 it is higher than cluster 1, but cluster 3 is the highest. Grand centroid is the average of each score, where 89.47 for attendance, 70.30 for assignments, UTS is 66.97, and UAS is 59.41. Based on the value of the cluster centroid generated for each cluster of each value, cluster 1 is the low value group, cluster 2 is the value group that is around the average,

Table 6. Statistical Value Cluster and Probability

Cluster	Number of Students	Cluster distance from Centroid
Low	3	19.75
About average	27	28.88
High	24	18.41

The number of students included in cluster 1, cluster 2, and 3 can be seen in the table 6. Table 6 shows that the number of students belonging to the low score group is 3 people, 1 each from class 2A and 2B. There are 27 students who have scores around the average, namely 2 in class 2A and 25 in class 2B. While those in the high score group were 24 students, namely 10 class 2A and 14 class 2B. Based on grouping based on the value of attendance, assignments, UTS, and UAS, it can be seen that class 2B is the best grade when compared to class 2A.

C. FCM Grouping Results

Table 7. FCM Grouping Results

Number of Clusters	Iteration	Objective Functions	Sw / Sb ratio	Xie Beni Index
1	17	17.78	3.85	0.79
2	27	10.33	1.17	0.71
3	63	14.55	1.04	0.63

D. Discussion

The results of descriptive statistical analysis show that class 2B SI FIKTI UMSU is the class that gets the highest average for statistics and probability courses and class 2A is the lowest. The scores obtained by students in class 2A and 2B vary widely, because the variance generated from the two classes for attendance, assignment, UTS and UAS scores is high. However, the highest diversity value of the two classes was class 2B and the lowest was class 2A. This can be seen from the resulting variance based on tables 1,2 and 3. The results of the cluster analysis which consisted of 3 groups showed that cluster 1 was a group of students with low scores, there were 3 students who came from classes 2A and 2B. This means that from 51 SI FIKTI UMSU students who took the Statistics course and the probability was 2.63% who received low scores. There are 27 students from class 2A 2 and 24 in class 2B who are in cluster 2, that is, the values are around the average. So there are 23.68% of SI students who get Statistics and Probability scores that are around the average. Meanwhile, those who received high scores were 73.68% or as many as 24 students from 51 SI FIKTI UMSU students, of which 27 were class 2A, and 24 were class 2B. Research on cluster analysis using the k-means algorithm has also been conducted by Poerwanto and Fa'rifah (2016), namely looking at the Discrete Mathematics scores obtained by TI FTKOM UNCP students. Overall,

4. CONCLUSION

Based on the results of the analysis that has been discussed, it is known that the highest diversity of scores in the Statistics and Student Probability subjects is class 2B and the lowest is class 2A. In general, the highest average score seen from attendance, assignments, UTS, and UAS is class 2A, and the lowest is class 2B. The results of clustering using the k-means algorithm show that the understanding of class 2B towards Statistics and Probability subjects is higher when compared to class 2A. It can be seen that of the 40 SI FIKTI UMSU class 2B students who are in cluster 3 (high score) there are 27 students, and those in cluster 1 (low score) there are 1 student.

REFERENCES

Al-Khowarizmi, A. K., Fauzi, F., Sari, I. P., & Sembiring, A. P. (2020). The effect of indonesian and hokkien mobile learning application models. *Journal of Computer Science, Information Technology and Telecommunication Engineering*, *1*(1), 1-7.

Borkowska-Niszczota, M., 2015. Tourism Clusters in Eastern Poland - Analysis of Selected Aspects of the Operation. Procedia - Social and Behavioral Sciences, 213, pp. 957–964. Available at: http://linkinghub.elsevier.com/retrieve/pii/ S1877042815058668.

Muningsih, E. & Kiswati, S., 2015. Application of the K-Means Method for Online Shop Product Clustering Determination of Stocks of Goods. Bianglala Informatics Journal, 3 (1), pp.10–17.

Nasari, F., Darma, S. & Information, S., 2015. Implementation of K-Means Clustering in Student Admission Data New. National Seminar on Information and Multimedia Technology 2015, 10 (2), pp. 73–78.

Ong, JO, 2013. Implementation of the K-Means Clustering Algorithm to Determine the Marketing Strategy. Journal Ilimiah Industrial Engineering, 12 (1), pp. 10–20.

Rivani, E., 2010. Application of K-Means Cluster for Grouping Provinces based on Production of Rice, Maize, Soybeans, and Mung Beans 2009. Mat Stat, 10 (2), pp. 122–134.

Poerwanto, B. & Fa'rifah, RY, 2016. Cluster Analysis Using the K-Means Algorithm. d'ComputarE, 6 (2), pp. 62–77

Al-Khowarizmi, A. K., Nasution, I. R., Lubis, M., & Lubis, A. R. (2020). The effect of a SECoS in crude palm oil forecasting to improve business intelligence. *Bulletin of Electrical Engineering and Informatics*, *9*(4), 1604-1611.

Prayudani, S., Hizriadi, A., Lase, Y. Y., & Fatmi, Y. (2019, November). Analysis Accuracy Of Forecasting Measurement Technique On Random K-Nearest Neighbor (RKNN) Using MAPE And MSE. In *Journal of Physics: Conference Series*(Vol. 1361, No. 1, p. 012089). IOP Publishing.

Ramadhani, F., & Ilona, D. (2018). Determinants of web-user satisfaction: using technology acceptance model. In *MATEC Web of Conferences* (Vol. 248, p. 05009). EDP Sciences.

Ramadhani, F., Ramadhani, U., & Basit, L. (2020). Combination of Hybrid Cryptography In One Time Pad (OTP) Algorithm And Keyed-Hash Message Authentication Code (HMAC) In Securing The Whatsapp Communication Application. *Journal of Computer Science, Information Technology and Telecommunication Engineering*, *1*(1), 31-36.

Ramadhani, F., Zarlis, M., & Suwilo, S. (2020). Improve BIRCH algorithm for big data clustering. In *IOP Conference Series: Materials Science and Engineering* (Vol. 725, No. 1, p. 012090). IOP Publishing.

Syah, R., Nasution, M. K., & Elveny, M. (2021). Sensitivity of MAPE using detection rate for big data forecasting crude palm oil on k-nearest neighbor. *International Journal of Electrical & Computer Engineering (2088-8708)*, *11*(3).

Sari, I. P., Hutagalung, F. S., & Hutasuhut, B. K. (2020). Determination of Campus Promotion Policy Strategy Applied The Profile Matching Method. *Journal of Computer Science, Information Technology and Telecommunication Engineering*, *1*(1), 17-23.

Hutagalung, F. S., Mawengkang, H., & Efendi, S. (2019). Kombinasi Simple Multy Attribute Rating (SMART) dan Technique For Order Preference by Similarity To Ideal Solution (TOPSIS) dalam Menentukan Kualitas Varietas Padi. *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan*, *3*(2), 109-115.

Sari, I. P., Hutagalung, F. S., & Hutasuhut, B. K. (2020). Analisa Model Pemanfaatan Jaringan Komputer Yang Efektif untuk Peningkatan Produktivitas pada Jaringan LAN Universitas Muhammadiyah Sumatera Utara. *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan*, *5*(1), 193-197.

Hutagalung, F. S., Sari, I. P., & Hutasuhut, B. K. (2020). Analisa SWOT Strategi Perencanaan Pemasaran Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Muhammadiyah Sumatera Utara. *InfoTekJar: Jurnal Nasional Informatika dan Teknologi Jaringan*, *5*(1), 198-201.

Lubis, A. R., Lubis, M., & Listriani, D. (2019, August). Big Data Forecasting Applied Nearest Neighbor Method. In *2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC)* (pp. 116-120). IEEE.

Lubis, A. R., & Prayudani, S. (2020, October). Optimization of MSE Accuracy Value Measurement Applying False Alarm Rate in Forecasting on Fuzzy Time Series based on Percentage Change. In *2020 8th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-5). IEEE.

Qiao, W., & Yang, Z. (2019). An improved dolphin swarm algorithm based on Kernel Fuzzy C-means in the application of solving the optimal problems of large-scale function. *IEEE Access*, *8*, 2073-2089.