School Clustering Using Fuzzy C Means Method

Dwi Vernanda
Subang State Polytechnic
Subang, Jawa Barat, Indonesia
yogurt.nda@gmail.com

Nunu Nugraha Purnawan
Subang State Polytechnic
Subang, Jawa Barat, Indonesia
nunu@polsub.ac.id

Tri Herdiawan Apandi
Subang State Polytechnic
Subang, Jawa Barat, Indonesia
h.apandi@gmail.com

Abstract— Subang State Polytechnic is one of the tertiary institutions established in 2014. As a new tertiary institution, this institution certainly competes with other tertiary institutions in obtaining prospective students. At present, Subang State Polytechnic determines several schools to be visited for promotion and socialization activities for new students based on the number of students in schools in Subang Regency. However, that does not prove that it has influenced students to enroll in the Subang State Polytechnic. This research highlights the problems involved in each school including the number of graduates, graduates who go to college, graduates who go to the Subang State Polytechnic, average national school exam scores, average report card grades, number of counseling guidance teachers, number of universities that come for socialization activities to schools, and the distance from schools to the Subang State Polytechnic. School clustering uses the Fuzzy C Means method for 40 high schools in Subang Regency. Determination of the number of clusters using Modified Partition Coefficient (MPC), the results of the MPC calculation revealed that there are 3 clusters, each cluster has a cluster center and members. In cluster I, there are 9 schools, cluster II has 16 schools, and cluster III consists of 15 schools. The results of the clustering assisted the new students admissions committee in determining potential schools visited in the context of promotion activities and socialization of new student admissions, namely schools in cluster I.

Keywords — Admision of new students; Cluster; Fuzzy C Means; School; Socialization;

I. Introduction

The higher education experience rapid development which has an impact on competition among state universities in the search for prospective new students, as well as Subang State Polytechnic. This institution is a new tertiary higher education established in 2014, for this reason Subang State Polytechnic needs to implement a marketing strategy to get excellent prospective students. One of the strategies undertaken is to conduct socialization to secondary schools, especially in the Subang Regency (Ayu & Wulaning, 2016).

At present, Subang State Polytechnic determines some schools to be visited for socialization activities for new student admissions based on the large number of students in the schools around Subang district, but it does not prove that the large number of students at the schools will enroll in Subang State Polytechnic. Based on the report of the 2016-2017, not all of those who registered and were accepted as students at Subang State Polytechnic come from the schools that had been socialized (PMB Polsub, 2016).

To solve the problems faced by the committees, a solution is needed, namely by segmenting the right schools for socialization (Saputra & Riksakomara, 2018), determination of schools that are targeted for socialization by using the Fuzzy C Means (FCM) method with several input attributes or parameters (Xue et al., 2018) (Hardiani, 2018). In making school decisions which are potential for holding socialization, data mining is needed in the form of data from schools such as the number of graduates, the number of students continuing to college, the number of students continuing to o.

The function of the Decision Making System by applying the FCM method is to do clustering or grouping (Agustian, Hartati, & Musdholifah, 2018) of schools that have the potential to become targets for the socialization of new student admissions (Javadi, Rameez, Dahl, & Pettersson, 2018). In addition, FCM is an easy algorithm and is often used in grouping data (Gomes, Blanco, & Pessoa, 2019) because it makes an estimate that is efficient and does not require many parameters (Muhardi & Nisar, 2015).

The results of clustering using the fuzzy c means method are used by the subang state polytechnic in determining potential schools to be visited in the context of promotion activities and socialization of new student admissions.

II. Literature Review

2.1 Acceptance of New Students

Some routine activities every year that are always carried out by a tertiary institution namely the New Student Admission process. The admission process is a starting point for the search for quality students (Kurniawan, 2016). In the last few decades, some developed countries have stated that there was a significant increase in the case of young people graduating from secondary education who continued their tertiary education (Heinesen, 2018).

2.2 Marketing and Outreach Strategies

One of the ways to outperform competition is to utilize marketing and socialization strategies, marketing strategies are the basis for the overall planning of a company or institution (Blankenau, 2014).

Planning regarding marketing can be used as a guide for companies and institutions in carrying out their activities (Ratnawati et al., 2017).

2.3 Data Mining

The study of methods for finding patterns from data is data mining. Data mining is a series of processes to explore previously undefined values, the data sourced from a database (X, Kumar, & Q, 2008). Knowledge Discovery from Data is a structured process as follows (Han, Kamber, & Pei, 2011):

1. Data Cleaning is cleaning data from inconsistent data
2. Data Integration is the process of combining data from several different sources
3. Data Selection is the process of selecting data from a database in accordance with the objectives
4. Data Transformation is the process of changing the form of data into data suitable for the mining process
5. Data Mining is an important process that uses a certain method to obtain a pattern from the data
6. Pattern Evaluation is the process of identifying patterns
7. Knowledge Presentation is that which can represent the information needed, the process by which the information that has been obtained is then used by the data owner.

The grouping data mining has resulted several parts, namely description, estimation, and prediction. (Nariya & Kim, 2017).

2.4 Fuzzy C Means Clustering

Clustering is a technique that can divide a set of data into groups, where each group has a degree of equality. The clustering technique is possible to do with two approaches namely hard grouping and soft grouping. Soft grouping one of them is Fuzzy C Means with a degree of membership between zero and one.

In 1973, Dunn discovered a method, then this method was developed by Bezdek in 1981. This method centers on fuzzy logic that is similar to the K-Means method and the naming of this method is Fuzzy C Means Clustering. The work process of this method is that the data collected has been entered into the data group, the data entry into groups based on the membership value (L. Zhang & Luo, 2018). The FCM algorithm is as follows (Memon & Lee, 2018):

1) Determination of the number of groups (*c*), maximum iteration (*MaxIter*), fuzzifier (*m*), then determining the smallest objective value, the expected objective value (ε). Determination of initial objective function (*P0* = 0) and initial iteration (*t* = 1)
2) Increasing the random number uik, a lot of data is symbolized by *i*, then the number of groups is symbolized by *k*. The elements of membership of *U* are *i* and *k*.
3) Calculating the center of the to-*i* group with the equation:

$$P_i = \frac{\sum_{k=1}^{N} (u_{ik})^m X_k}{\sum_{k=1}^{N} (u_{ik})^m} \qquad (1)$$

4) The formula for calculating object functions on the *t* iteration with an equation:

$$J\,(P,U,X,c,m) = \sum_{i=1}^{C}\sum_{k=1}^{N}(u_{ik})^m\, d_{ik}^2(X_k,P_1) \qquad (2)$$

Information:

- *c* is the expected number of groups,
- *N* is a lot of research objects,
- uik is the membership value of the *k*-specific object in the *i* group (part of the matrix)
- *U*, *m* are *fuzzifiers*, and $d_{ik}^2(x_k,p_i)$ is the distance between the *k* approach and the center of the *k-i* group.

5) Calculation of changes in the membership matrix with the equation:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left[\frac{d_{ik}^2}{d_{jk}^2}\right]^{\frac{1}{m-1}}} \qquad (3)$$

the uik is the membership value of the k-th object with the center of the group, *i*, d_{ik}, d_{ik}^2 is the distance between the k-th object and the center of the i-th group, d_{ik}^2 is the distance between the *k*-object with the center of the *i* group *j*, and *m* is the *fuzzifier*.

6) Condition checking

- If |Jt−Jt−1|<ε or $t>MaxIter$ then stop;
- If not :t = t + 1, repeat to step 3;

The use of Fuzzy C-Means algorithm is often used for grouping data that is used to estimate something, estimation with FCM is an efficient estimate and does not require many parameters (Stetco, Zeng, & Keane, 2015). Several studies have produced statements that the Fuzzy CMeans method can be used to classify data based on certain attributes (Muhardi & Nisar, 2015).

2.5 Modified Partition Coefficient

Testing the results of the cluster use Modified Partition Coefficient (MPC). MPC is a method used to test the number of clusters that are right and valid (Y. Zhang, Wang, Zhang, & Li, 2008). MPC is a modification of the Partition Coefficient (PC) method which is able to reduce monotonous changes on the PC. The MPC formula is as follows:

$$MPC\,(c) = 1 - \frac{c}{c-1}\big(1 - PC(c)\big) \qquad (4)$$

MPC is used as a new way to calculate clusters in Fuzzy C Means, the results are known by calculating the distance between the degree of membership and the center of the cluster. The MPC value is between 0 to 1, in general to determine the optimal number of clusters with the largest MPC value. So by knowing the MPC index, it can validate the right number of clusters (Suleman, 2015).

2.6 Decision Making

One of the computer systems that can be utilized for decision making is the Decision Making System (SPK), SPK is used for decision makers to solve problems that are not fully structured (Rai, Sharma, & Lohani, 2019), namely by using data which can then be processed become information. The amount of information resulting from data processing can be used as a reference towards a particular decision making (Rohayani, 2013).

III. Proposed Method

This study uses the fuzzy c means method in determining the school clusters to be visited for the socialization of new student admission activities. The stages of this research process can be seen in Fig. 1. Completion Methodology.

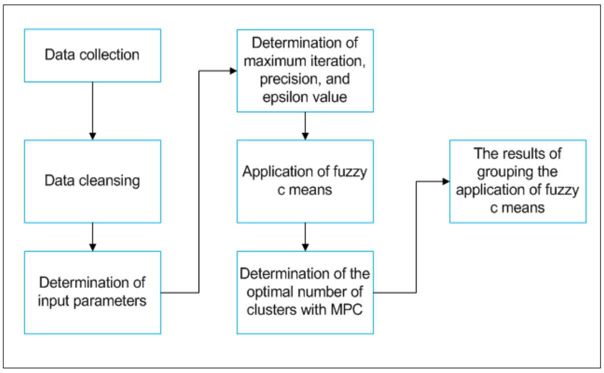
Data collection
Data cleansing
Determination of input parameters
Determination of maximum iteration, precision, and epsilon value
Application of fuzzy c means
Determination of the optimal number of clusters with MPC
The results of grouping the application of fuzzy c means

Fig 1. Completion Methodology

a. Data collection
The data collection is obtained from the students data for the 2017-2018 academic year from high schools in Subang Regency.
b. Data cleansing
Data cleansing is done to eliminate fields that do not need to be used during the data mining process.
c. Determination of input parameters
Determination of input parameters that will be used in research
d. Determination of maximum iteration, precision, and epsilon value
e. Application of fuzzy c means
f. Determination of the optimal number of cluster with modified partition coficient
g. The output result will determine the potential schools for outreach activities

IV. RESULT AND DISCUSSION

4.1 The Used Data

The table 1 is a breakdown of the used input parameters and school data. The data obtained is stored in the form of .xls and the total school data used is 40.

Table 1. Data of schools

Parameters	School			
	SMAN 1	SMKN 1 SUBANG	SMA PGRI	SMK RADITYA
the number of graduates from each school	314	203	381	84
the number of students continuing to tertiary institutions (both public and private) in percentage	88	20	40	22
the number of students studying at the Subang State Polytechnic	8	0	0	0
the average national exam scores of each school	82	90	71	77
the average grade report cards	90	86	77	70
number of college socialization	11	5	7	5
the number of counseling guidance teachers at the school	4	2	4	1
the distance of the school to the Subang State Polytechnic campus	1.3	43.4	3.7	15.6

4.2 *Clustering Fuzzy C Means (FCM)*

The FCM method grouping is done on a number of school data for two to five clusters. The parameters used are the number of students graduating, the number of students continuing to college, the number of students enrolling Subang State Polytechnic, the average score of school's national examination, the average grade of report cards, the number of counseling guidance teachers at school, the number of colleges attending socialization to school, and the distance from the schools to Subang State Polytechnic. The maximum iteration is 100 and the epsilon value is 0.000001.

The first calculation is started by determining the degree of membership randomly. After determining the degree of membership randomly, it starts with an iteration process, calculation μ squared and determination μ *cluster*. Determination of the cluster center of the amount μ *cluster*. Furthermore, the value of the objective function is calculated and the difference between the objective function is determined. In the first iteration, the difference in objective function value is greater than the epsilon value. Consequently, it must be continued to the next iteration. The iteration is continued until the difference in objective function value is smaller than the epsilon value, 0.000001.

4.3 Modified Partition Coefficient

Cluster determination is done using Modified Partition Coefficient (MPC). The result after clustering, starting from two to five clusters, MPC determines the most optimal number of clusters. It also can validate whether the data that has been in the cluster is appropriate (Kon & Kuwano, 2013). The validity index obtained for each cluster is in Table 2. Cluster Results Validity Index.

Table 2. Cluster Result Validity Index

No	Cluster	MPC
1	2	0.8679
2	3	0.9902
3	4	1.0023
4	5	1.0340

The optimal number of clusters can be seen from the highest MPC value and it is between ≥ 0 to ≤ 1. So that there are 3 optimal clusters in the grouping of schools that are potential for the socialization of the New Student Admissions.

4.4 Distribution Data

Data sets have a degree of membership to be grouped into three clusters. Schools that are closer to the center of the cluster will have a higher level of membership and vice versa. The distribution of degrees of membership based on each parameter is illustrated in Fig. 2-9.

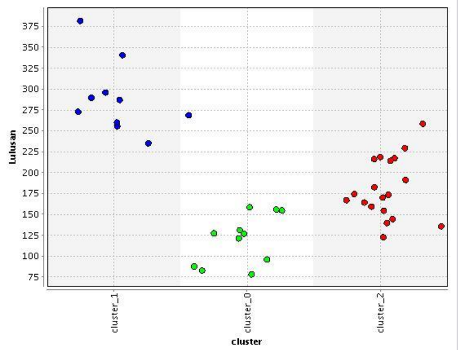

Fig 2. Graduate Distribution

From Fig 2, it can be seen that schools in cluster I tend to have more graduates compared to cluster I and cluster II, which is between 200-370. Schools in cluster II tend to have fewer graduates below 150 graduates, while in cluster III the number of graduates is in grades 125-270.

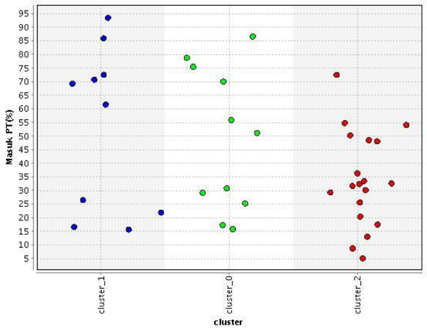

Fig 3. Distribution of Higher Education Entrance

Fig 3 illustrates that in cluster I, graduates who went to tertiary institutions tended to be above 80%, although there were some schools where students went to tertiary institutions only below 20%. In cluster II the number of students continuing to tertiary education varied from 15% to 80%. In cluster III, graduates who go on to college tend to have lower grades, below 40%.

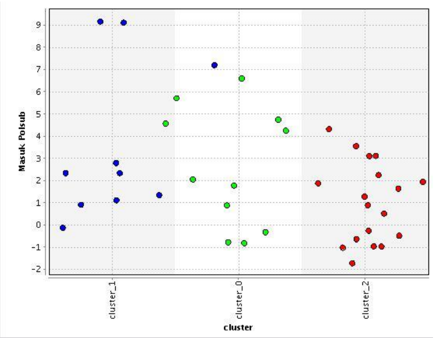

Fig 4. Distribution Students Who Proceed Subang State Polytechnic

Fig 4 shows the distribution of school students who proceed to Subang State Polytechnic, the top value is in cluster I. From Fig 4 it can be seen that the grouping with the fuzzy c means data method in one cluster can be a member of another cluster. This is because the level of data presence in a cluster is determined by the degree of membership.

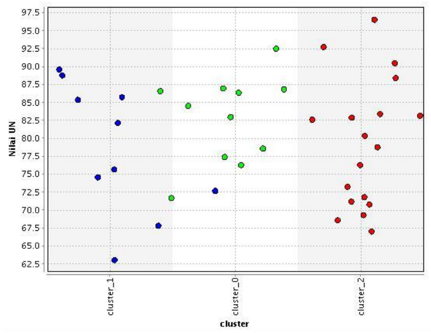

Fig 5. Distribution of National Exam Scores

Fig. 5 illustrates the distribution of school national averages, in cluster I the National Examination scores are below 90 with the lowest score being 60, cluster II the average National Examination scores tend to converge between 70-80. For cluster II the value is between 70 - 80 and cluster III the value tends to gather at values below 70.

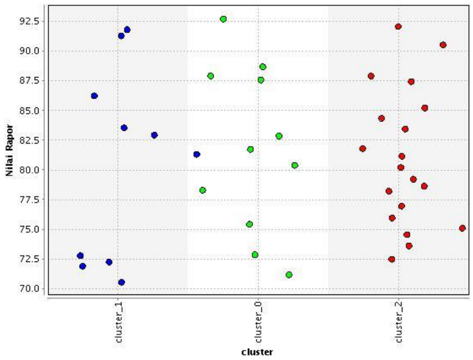

Fig 6. Distribution Card Report

From Fig 6, it can be seen that for cluster I the average distribution of report cards is at the top value that is 90 and the bottom value is 70. In cluster II, the average report value is spread from 70 to 92.5 and in cluster III the average value of report cards tends to gather at a value between 72.5 and 80.

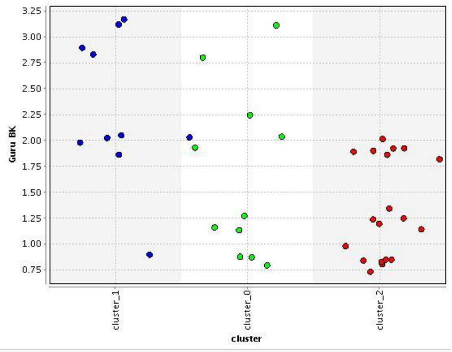

Fig 7. Distribution of Counseling Guidance Teachers

Figure 7 illustrates the distribution of BK teachers in each school, in cluster I the number of BK teachers is at the top level, namely more than three, cluster II and III the number of BK teachers tends to be below 2.

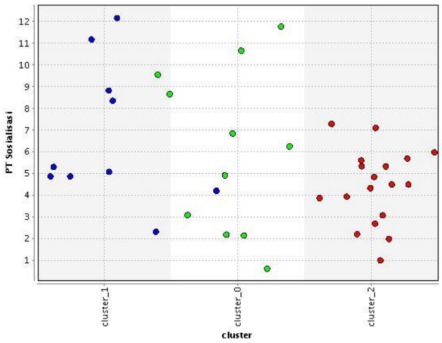

Fig 8. Distribution of Higher Education Socialization

Fig 8 shows the distribution of the number of universities that have come to school, in cluster I, the number of universities that came to school is above 5 to 12, for cluster II, the number of universities that come to school is very varied in numbers from 0 to 11. In cluster III tends to gather in the numbers below 6.

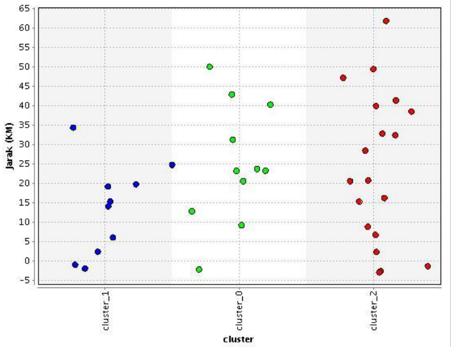

Fig 9. Distribution of School Distance to Subang State Polytechnic

From Fig 9, it can be seen that for cluster I the distance distribution between schools and Subang State Polytechnic is in the range below 20 KM, while for cluster II the spread is between 10 - 50 KM, and for cluster III the distance from schools to Subang State Polytechnic reaches the furthest distance which is above 60 KM.

4.5 Discussion

After the clustering process, the cluster center is obtained for each parameter as shown in Table. 2 Cluster Center.

Table. 2 Cluster Center

	Cluster I	Cluster II	Cluster III
Number of graduation	320.77	210.75	134.20
Number of students continuing on to collage	59.88	40.375	35.20
Number of admission to Subang State Polytechnic	3.66	1.93	1.40
Average National Examination Score	80.0	78.0	82.13
Average report	79.77	82.625	80.53
Number of teachers guidance counseling	7.77	5.1875	4.60
Number of campus socialization	2.88	1.6875	1.46
Distance from Subang State Polytechnic to school	9.05	24.57	26.8

Data sets that have a membership level closer to the center of the cluster will have a greater membership level and a dataset that has the highest membership level in one particular cluster is representative for that cluster. In cluster I with cluster center 70.47 there are 9 schools and the representatives are State Senior High School 1 Subang, State Senior High School 3 Subang, State Senior High School 1 Kalijati, State Senior High School 1 Pabuaran with the number of graduates continuing to tertiary and private tertiary institutions above 70% and marked by the distance from schools to Subang State Polytechnic ≤ 5 KM.

The center of cluster II is 55.64 and there are 16 schools with representatives of State

Senior High School 2 Subang, State Senior High School 1 Kalijati, State Senior High School 1 Ciasem. Cluster II was marked by the number of graduates who went to college < 70% with the distance from schools to Subang State Polytechnic of KM 20 KM. In cluster 3 cluster center 45.79 there are 15 schools with representatives, namely State Senior High School 1 Pusakagara, State Senior High School 1 Tanjungsiang, Vocational School of Bakti Kencana. The hallmark of cluster III is the number of graduates continuing to tertiary institutions below 30%, for the distance from schools to the Subang State Polytechnic ≥ 30 KM.

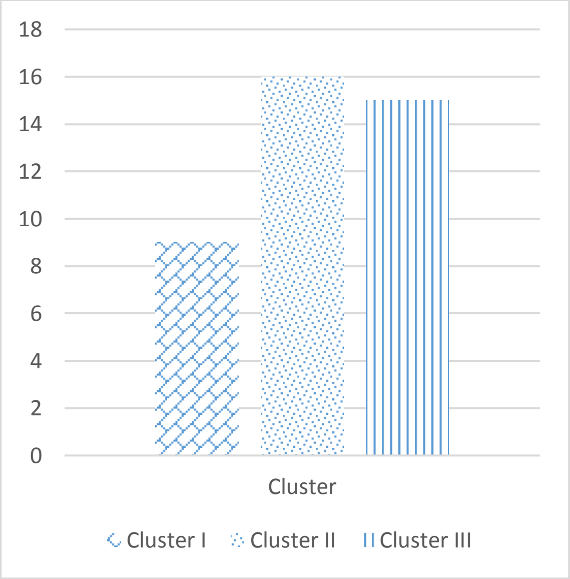

Fig 10. School Cluster Distribution

Fig 10 shows the school distribution diagram by cluster. These results can provide input to the Subang State Polytechnic Students Admission committee in determining potential schools for socialization. The potential schools are schools in cluster I with a total of 9 schools, the center of the cluster for graduates who go to tertiary institutions, both public and private, reaches 59.88, this is the highest data compared to cluster II and cluster III. Likewise, the cluster center on the parameters of other tertiary institutions that came to the school reached 2.88. While the center of the cluster the distance from the schools to Subang State Polytechnic is at the smallest value which is 9.05.

V. Conclusion and Suggestion

5.1 Conclusion

Based on research that has been done, it can be seen that the fuzzy c method means it can be used as a tool to conduct an analysis of approved school assessments for the socialization of admission of new students to the Subang State Polytechnic. By using the Modified Partition Coefficient, the most optimal result is the validity of three clusters with a value of 0.9902. Of the 40 existing schools, there are 9 schools included in cluster I, 16 schools are in cluster II, and 15 schools are in cluster III. The results of this grouping are used by the new admissions committee in making potential school decisions to be visited in the context of promotion activities and socialization of new student admissions, namely schools in cluster I.

5.2 Suggestion

1. This research can be continued by identifying the attributes that have a big influence on the clustering process.
2. Data that is processed should be more and not only in one batch.
3. This research can be continued using other clustering methods or can be

combined between fuzzy c means method with other methods.

VI. References

Agustian, H., Hartati, S., & Musdholifah, A. (2018). Two level clustering untuk analisis kuesioner akademik di STTA Yogyakarta. *Jurnal Ilmiah Bidang Teknologi, Angkasa, X Nomor 1*, 29–40.

Ayu, & Wulaning, P. (2016). Perancangan Sistem Pendukung Keputusan Pemasaran STIKOM Bali. *Sistem Dan Informatika*, *10*(2).

Blankenau, W. (2014). Admission standards, student effort, and the creation of skilled jobs. *Economic Modelling*, *43*, 209–216.

Gomes, E. P., Blanco, C. J. C., & Pessoa, F. C. L. (2019). Identification of homogeneous precipitation regions via Fuzzy c-means in the hydrographic region of Tocantins–Araguaia of Brazilian Amazonia. *Applied Water Science*, *9*(1), 1–12. http://doi.org/10.1007/s13201-018-0884-6

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques 3rd Edition* (3rd ed.).

Han, & Kamber, M. (2011). Data mining: Concepts and Techniques. 3rd Edition. *San Francisco: Morgan Kaufmann Publisher*, *1*.

Hardiani, T. (2018). Segmentasi Nasabah Simpanan Menggunakan Fuzzy C Means Dan Fuzzy RFM (Recency , Frequency , Monetary) Pada BMT XYZ. *Jurnal Ilmiah NERO*, *3*(3), 185–192.

Heinesen, E. (2018). Admission to higher education programmes and student educational outcomes and earnings – evidence from Denmark. *Economics of Education Review*, *10*, 3–9. http://doi.org/10.1016/j.econedurev.2018.01.002

Javadi, S., Rameez, M., Dahl, M., & Pettersson, M. I. (2018). Vehicle Classification Based on Multiple Fuzzy C-Means Clustering Using Dimensions and Speed Features. *Procedia Computer Science*, *126*, 1344–1350. http://doi.org/10.1016/j.procs.2018.08.085

Kon, M., & Kuwano, H. (2013). On sequences of fuzzy sets and fuzzy set-valued mappings. *Fixed Point Theory and Applications*, *327*, 1–19. http://doi.org/10.1186/1687-1812-2013-327

Kurniawan, E. (2016). Metode TOPSIS untuk Menentukan Penerimaan Mahasiswa Baru Pendidikan Dokter di Universitas Muhammadiyah purwokerto. *Bachelor Thesis*, (Universitas Muhammadiyah Purwokerto).

Memon, K., & Lee, D.-H. (2018). Generalised kernel weighted fuzzy C-means clustering algorithm with local information. *Fuzzy Sets and Systems*, *340*, 91–108.

Muhardi, & Nisar. (2015). Penentuan Penerimaan Beasiswa dengan Algoritma Fuzzy C Means. *Universitas Megow Pak Tulang Bawang, Tim Darmajaya*, *1*(2).

Nariya, M., & Kim, J. H. (2017). Comparative Characterization of

Crofelemer Samples Using Data Mining and Machine Learning Approaches With Analytical Stability Data Sets. *Pharmaceutical Sciences*, *106*(11), 3270–3279.

PMB Polsub, P. (2016). *Laporan PMB POLSUB Tahun 2016-2017. Politeknik Negeri Subang*.

Rai, S. P., Sharma, N., & Lohani, A. K. (2019). Novel approach for issues identification in transboundary water management using fuzzy c ‑ means clustering. *Applied Water Science*, *9*(1), 1–11. http://doi.org/10.1007/s13201-018-0889-1

Ratnawati, A. Y., Kom, S., Susena, M. M. E., Kom, S., Kom, M., & Terdahulu, P. (2017). KESEJAHTERAAN PEDAGANG BATIK DI KOTA SURAKARTA. *SAINSTECH*, *4*(2), 58–66.

Rohayani, H. (2013). Analisis Sistem Pendukung Keputusan Dalam Memilih Program Studi Menggunakan Metode Logika Fuzzy. *Jurnal Sistem Informasi (JSI)*, *5*(1), 530–539.

Saputra, D. B., & Riksakomara, E. (2018). Implementasi Fuzzy C-Means dan Model RFM untuk Segmentasi Pelanggan. *Jurnal Teknik ITS*, *7*(1), 1–6.

Stetco, A., Zeng, X., & Keane, J. (2015). Expert Systems with Applications Fuzzy C-means ++ : Fuzzy C-means with effective seeding initialization. *EXPERT SYSTEMS WITH APPLICATIONS*, *42*(21), 7541–7548. http://doi.org/10.1016/j.eswa.2015.05.014

Suleman, A. (2015). A new perspective of modified partition coefficient ☆. *Pattern Recognition Letters*, *56*, 1–6. http://doi.org/10.1016/j.patrec.2015.01.008

X, W., Kumar, & Q, R. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14(1)*, 1–37.

Xue, M., Zhou, L., Kojima, N., Sarmento, L., Machimura, T., & Tokai, A. (2018). Application of fuzzy c-means clustering to PRTR chemicals uncovering their release and toxicity characteristics. *Science of the Total Environment*, *622–623*, 861–868. http://doi.org/10.1016/j.scitotenv.2017.12.032

Zhang, L., & Luo, M. (2018). Diverse fuzzy c-means for image clustering. *Pattern Recognition Letters*, *1*.

Zhang, Y., Wang, W., Zhang, X., & Li, Y. (2008). A cluster validity index for fuzzy clustering. *Information Sciences*, *178*(4), 1205–1218. http://doi.org/10.1016/j.ins.2007.10.004