ELINVO (Electronics, Informatics, and Vocational Education), November 2021; vol 6 (2): 166-173
ISSN 2580-6424 (printed), ISSN 2477-2399 (online,) DOI: https://doi.org/10.21831/elinvo.v6i2.44486

A Comparison of K-Means and Agglomerative Clustering for Users Segmentation based on Question Answerer Reputation in Brainly Platform

Puji Winar Cahyo[1*], Landung Sudarmana [2]
[1,2] Department of Informatics, Universitas Jenderal Achmad Yani Yogyakarta
[*]E-mail: pwcahyo@gmail.com

ABSTRACT

Brainly is a question and answer (Q&A) site that students can use as a media for questions and answers. Students can also use Brainly to find and share educational information that helps students solve their homework problems. In Brainly, users can answer questions according to their interests. However, it could be that the interest is not necessarily following the competencies possessed. It causes many answers to the questions given not to have a high rating because the answers given are of low quality to be prioritized as the main answer. This study aims to apply the K-Means and Agglomerative Clustering methods to segment users based on the reputation of the answerers by conducting clustering based on track records in answering questions on mathematics subjects. This study used the number of the brightest scores and the number of answers that did not get a rating as the basic features for clustering. The comparison between the two methods used is based on the Silhouette Score, representing the quality of the clustering results, calculated by applying the Silhouette Coefficient method. This study result indicates that the K-Means method gives better results than the Agglomerative Clustering. The Silhouette Score generated by the K-Means method is higher at 0.9081 than the Agglomerative Clustering method, which is 0.8990, which produces two clusters or two segments.

Keywords*:* K-Means, Agglomerative Clustering, Question Answerer, Brainly, Clustering

INTRODUCTION

Question and answer (Q&A) sites have developed into one of the significant sources of information and knowledge emerging from Internet-mediated social practices [1]. Users can publish questions, provide answers and comments, and discuss with other users [2]. Even today, question and answer sites have developed into one of the components of educational resources that appear to support the student learning process [3], [4] which students can use to find and share academic information. Moreover, the question and answer sites are web-based to be accessed by students online to support the online learning process, especially during the current COVID-19 pandemic [5].

In the last few years, research on question and answer sites has been conducted several times. A study [1] aimed to identify user criteria and data-based features, both textual and non-textual, to assess the quality of answers published on the question and answer sites. One of the identification processes was done by observing how important user features are and the emphasis on answering expertise. Identifying relationships or gaps between user quality criteria and data features across knowledge domains can help better understand user evaluation behavior for their preferred answers and response quality evaluations. A study [6] aimed to identify essential indicators and improve the quality of solutions on the question and answer sites. The analysis results showed that several important factors affect the quality of answers, which could be used as indicators to improve the quality of solutions and guide the publication of more high-quality answers.

Brainly is a question and answer site that students can use as a medium to ask and answer questions about school subjects [4]. Brainly is designed to use for students to find and share educational information that helps students solve their homework problems [7]. Students and even parents can use the site to ask questions related to their children's homework. Questions that

have been published can be answered by all users, whether students, material experts, or professional educators who master related fields. Brainly has also implemented gamification elements in the form of points and rankings to motivate users to be more actively involved in the community formed on the site.

In Brainly, each user has a particular interest in a specific subject. These users can answer questions according to their interests. However, it could be that the interest is not necessarily following the competencies possessed. It causes many answers to the questions given not to have a high rating. Answers with a low rating are indicated as answers that lack quality to be prioritized as the main answer. For this reason, the questioner needs to know the reputation of the answerer's profile on a particular subject, whether the answerer masters that subject.

Based on these problems, this study aimed to apply data mining methods to segment users based on the reputation of the answerers. We conducted clustering based on track records in answering questions on specific subjects to be used as a basis for users who ask questions to prioritize or not answers that come from that user. This study applied several data mining methods to cluster data as the basis for grouping user data. The clustering process aimed to form a cluster of users who answer based on the number of the brightest scores and the number of answers that do not get a rating on specific subjects.

Various studies have been conducted to improve the quality of the answers published on the Q&A site and improve the service to Q&A site users, especially Brainly. A study [8] aimed to identify how a Q&A site can enhance the quality of answers where one of the case studies is Brainly. The study conveyed that Brainly tried to maintain high-quality answers by recruiting moderators to participate massively in removing inappropriate questions. Only experienced users, such as moderators, are allowed to delete answers. If the answers are incomplete, incorrect, irrelevant, or spam, they probably will be deleted. Most of the answers were removed (30%) to keep the high site quality. A study was also conducted by [9] aimed to improve Brainly's features related to speed and accuracy of answers. The study proposes a model to help users get faster and more precise answers by forming student clusters. Student clusters are created based on a list of the brightest students each day, interest in the subject, and activeness in answering questions.

Based on the literature study described previously, research related to the application of data mining methods to determine the reputation of answering questions based on track records in answering questions on specific subjects has never been done by previous research. This study implemented several data mining methods to cluster data as the basis for grouping Brainly user data, including the K-Means and Agglomerative Clustering methods. It aimed to find out which way produces the best cluster quality. The agglomerative clustering method consists of several approaches, namely Single, Complete, and Average Linkage, which we would choose the best approaches from them and then compared with the results of clustering using the K-means method.

The K-Means method is one of the popular clustering methods used by several previous researchers, especially in the field of education [10]–[13]. The agglomerative clustering method has also been used to cluster data in several previous studies [14]–[16], which has better performance and accuracy than the K-Means method in particular case studies [17].

METHODS

This study consists of several stages: data collection, program design & implementation, and evaluation of the quality of clustering results as a basis for comparison.

A. Data Collection

In this study, the user's answering profile data collection was carried out by utilizing one of the web data extraction techniques called web

scraping. We could use this method to extract data from web pages [18]. The stages of the extraction process using this method consisted of analyzing web data structures and making crawl engines to parse HTML and XML documents [19] from Brainly web pages. After that, we used the user profile data from the results of this process for the clustering process.

B. Program Design and Implementation

At this stage, the program design and implementation were carried out. The data clustering program in this study consisted of several processes. The initial process is to filter the data by identifying the user profile data to answer questions on a subject on the Brainly platform only to get the needed data. We used the data from the filtering process for the clustering process using the K-Means and Agglomerative Clustering methods. As previously explained, the Agglomerative Clustering method consists of several approaches, namely single, complete, and average linkage. We would choose the best approaches from them. After that, we compared it with the results of clustering using the K-means method.

The results of the clustering process will produce users who have the type of answerer who has similar achievements. From the group of respondents who have identical achievements, users with the level of achievement of answers to questions that are less qualified can be identified. Therefore, we would know their reputation. In general, the sequence of the data clustering process in this study can be seen in Figure 1 Starting from the acquisition of user profile data (answering questions) through Brainly site. Data of user profile proceed to the data filtering stage, it process determines the cluster parameters to be used. The results of the parameterized data can then be forwarded to the clustering stage so as to produce user cluster groups. The results of the clustering are used to validate the optimal number of clusters to be applied.

Clustering is the process of grouping data into several groups so that objects in one group have many similarities and have many differences with objects in other groups [20]. The K-Means method is a clustering method that functions to group *N* objects into *K* classes based on distance from the cluster to group objects that are almost the same as a particular area [21]. The K-Menas method is one of the data mining methods used to cluster data. The K-Means method is a cluster analysis method that leads to partition *N* objects of observation into *K* clusters (clusters) where each object of observation belongs to a cluster with the closest mean or average value [22].

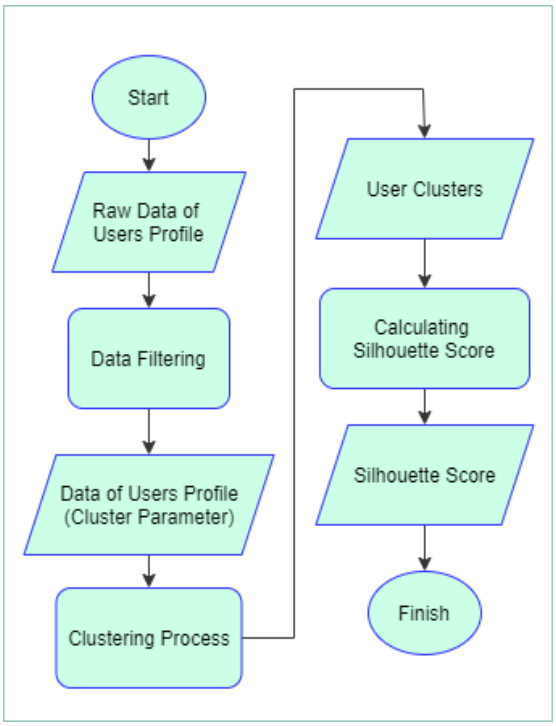

Figure 1. Program Design

In general, the K-Means method is carried out through several stages as follows [23]: (1) Determination of the number of clusters; (2) Determination of the cluster center (centroid) randomly according to the number of clusters that have been determined; (3) Determination of clusters of each data based on the proximity value between the data and the cluster center; (4) Calculate the latest cluster center from the data in each cluster by finding the average value; (5) Determination of each

data cluster based on the proximity value between the data and the new cluster center; (6) Repeat step 4 until no data moves cluster.

The calculation of K-Means is obtained from (1) [21],

$$d(x_j, c_j) = \sqrt{\sum_{j=1}^{n}(x_j - c_j)^2} \quad (1)$$

where:
d : distance
n : number of objects
j : (starting from 1 to n)
x_j : feature object j to x
c_j : centroid feature to j

The agglomerative clustering method is a method of data clustering by grouping data objects into a bottom-up hierarchical group [24]. This method begins by considering every single object that exists as a cluster and then iteratively combines them to form larger clusters. This method has several approaches, namely Single, Complete, and Average Linkage.

The single linkage approach is a clustering procedure based on the smallest distance between objects. This grouping algorithm begins by selecting the smallest distance in the matrix $D = \{d_{ij}\}$, then combining the corresponding objects such as *P* and *Q* to obtain a cluster (*PQ*). The next step is to find the distance between (*PQ*) and other clusters, for example, *R* so that it can be written as in (2) [25] where dRP is the distance of the closest neighbors from clusters *R* and then *P* and *dRQ* is the distance of the closest neighbors from clusters *R* and *Q*.

$$d(R, P + Q) = \min\{d(R, P), d(R, Q)\} \quad (2)$$

The complete linkage approach is a clustering procedure based on the greatest distance between objects. This clustering algorithm begins by selecting the largest distance in the matrix $D = \{d_{ij}\}$, then combining the corresponding objects such as *P* and *Q* to obtain a cluster (*PQ*). The next step is to find the distance between (*PQ*) and other clusters, for example, *R* so that it can be written as in (3) [25] dRP is the distance of the farthest neighbor from clusters *R* dan then *P* and *dRQ* is the distance of the farthest neighbor from clusters *R* and *Q*.

$$d(R, P + Q) = \max\{d(R, P), d(R, Q)\} \quad (3)$$

The average linkage approach is a centralized grouping procedure based on the average between objects. The average linkage algorithm begins by defining a matrix $D = \{dij\}$ to obtain the closest object, for example, *P* and *Q*, then these objects are combined into clusters (*PQ*). Then the distance between (PQ) and other clusters *R* so that it can be written as in (4) [25] where n_P is the number of members in cluster *P* and n_Q is the number of members in cluster *Q*.

$$d(R, P + Q) = \frac{n_P\, d(R,P) + n_Q\, d(R,Q)}{n_P + n_Q} \quad (4)$$

The next stage is implementation. At this stage, the program code is made according to the program design made in the previous stage. The programming language used to create programs is the Python programming language. This research also utilizes several libraries in the Python programming language, including the Numpy, Pandas, Scikit-learn, and Matplotlib libraries.

C. Evaluation of Clustering Results

In this study, the evaluation of cluster quality from the clustering results using the K-Means method and Agglomerative Clustering was carried out using the Silhouette Coefficient method. It was done by averaging the distance between an object and all other objects in the cluster and the minimum average distance from an object to all other clusters [23]. If the Silhouette Score value of 0 is close to 1, then the cluster containing the objects is very dense, and the objects are far apart from other clusters, which shows the better quality of the cluster. On the other hand, if the Silhouette Score of 0 is close to -1, it means that the cluster that contains objects is not dense, and the object is very close to other clusters, which shows that the quality of the cluster is getting worse [26].

RESULT AND DISCUSSION

This study will discuss the comparison of clustering results using the K-Means and Agglomerative Clustering methods (single linkage, complete linkage, and average linkage) to determine the reputation of the answerers on the Brainly platform. In this study, users profile data was obtained by utilizing one of the web data extraction techniques called web scraping, which is a technique to extract data from the Brainly web page automatically. As a sample, the user's profile data used is user profile data based on mathematics. From this data collection process, 852 data were obtained where the data consisted of many features such as user names, joining dates, etc.

In this study, the features used as the basis for clustering were the number of the smartest scores and the number of answers that did not get a rating in mathematics. This feature is based on the answerer's historical rating profile, which other users can normally rate [27]. Through this, a data filtering process is carried out to delete data based on features that are not used. The results of data filtering in this study are shown in Table 1.

Table 1. Users Profile Data

No.	Username	Smartest Total	Total Answer With No Rating
1.	DB45	4320	19950
2.	Takamori37	2280	7221
3.	5dregensleyer	2252	13669
4.	newwiguna	1926	9602
5.	whongaliem	1649	7448
6.	tribeking22	1553	10217
7.	Kilos	1429	6910
	.....		
848.	nuris99	0	4
849.	Smartgirl06	0	1
850.	riaamriani	0	5
851.	crsh23	0	7
852.	khoulahikhsan	0	5

After data filtering was carried out, the process towards the clustering stage could be carried out using a predetermined method. Before completing the clustering stage using each method, looked for the optimal number of suitable clusters for 852 data with the two previously mentioned features. The search for the optimal number of clusters was carried out by applying the Silhouette Coefficient method to calculate using the Silhouette Score method by conducting nine experiments, ranging from 2 to 10 clusters.

Based on the experiment of the data clustering process using the K-Means method, the results obtained are shown in Figure 2. We tested the optimal number of clusters nine times with cluster variations ranging from 2, 3, 4, 5, 6, 7, 8, 9, 10. Of the nine tests, the optimal number of clusters was 2 clusters.

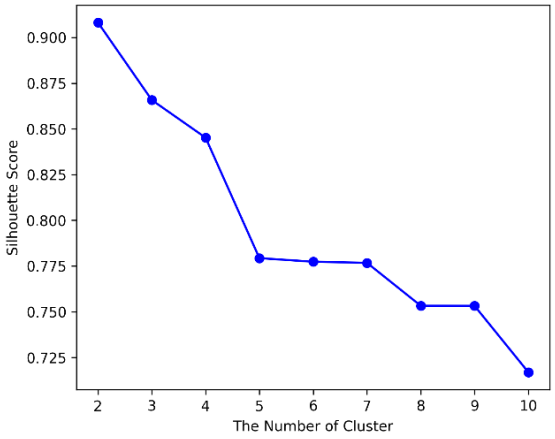

Figure 2. Silhouette Score of K-Means

Based on the data clustering process experiment using the Agglomerative Clustering method, each approach obtained results shown in Figure 3, Figure 4, and Figure 5. The optimal number of clusters was tested nine times with a varying number of clusters. From the results of the tests carried out, the optimal number of clusters is 3 clusters for the Single Linkage approach and 2 clusters for the Complete and Average Linkage approaches. The experiment, of course, used the same amount of Brainly platform user data so that it can be seen that the optimal cluster value in the Agglomerative Clustering method is 2 clusters.

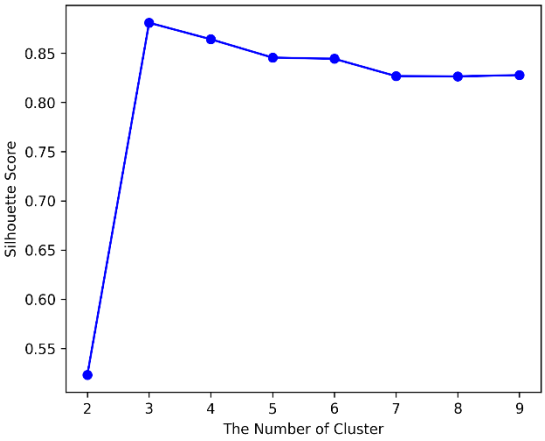

Figure 3. Silhouette Score of Agglomerative Clustering (Single Linkage)

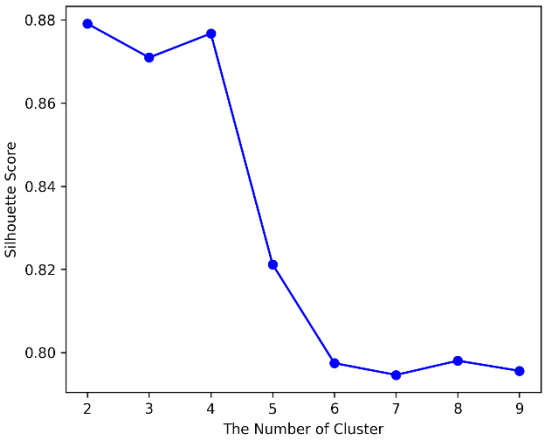

Figure 4. Silhouette Score of Agglomerative Clustering (Complete Linkage)

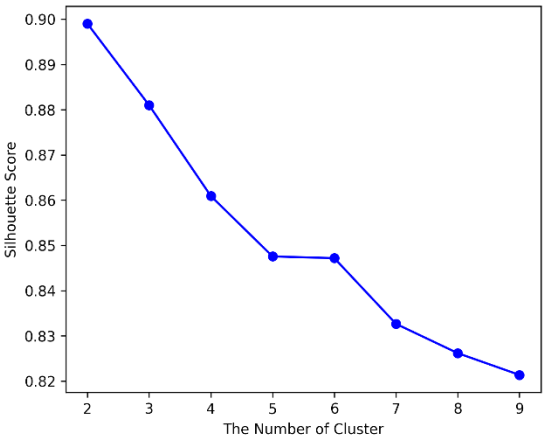

Figure 5. Silhouette Score of Agglomerative Clustering (Average Linkage)

In more detail, the comparison of the Silhouette Score values from the clustering results using the predetermined method is shown in Table 2.

Table 2. Silhouette Score Comparison

NC	KM	SL	CL	AL
2	0.9081	0.5233	0.8791	0.8990
3	0.8658	0.8810	0.8709	0.8810
4	0.8452	0.8642	0.8767	0.8610
5	0.7793	0.8456	0.8212	0.8476
6	0.7773	0.8444	0.7975	0.8472
7	0.7768	0.8267	0.7946	0.8326
8	0.7533	0.8265	0.7980	0.8262
9	0.7532	0.8277	0.7956	0.8214

In Table 2, there are five attributes, namely The Number of Clusters (NC), Silhouette Score of K-Means (KM), Silhouette Score of Agglomerative Clustering with Single Linkage (SL) approach, Silhouette Score of Agglomerative Clustering with Complete Linkage CL approach), and Silhouette Score of Agglomerative Clustering with Average Linkage (AL) approach.

Based on the results, the Average Linkage approach has the highest result compared to the two approaches in the other Agglomerative Clustering method. It was a score of 0.8990. Therefore, we used the Agglomerative Clustering method with the Average Linkage approach and compared it with the K-Means method. Based on Table 2, it can be seen that the results of clustering with the K-Means method get a score of 0.9081. These results are better when compared to the results of clustering using the Agglomerative Clustering method.

The results of determining the most optimal number of clusters from the K-Means method with the Agglomerative Clustering method with the Average Linkage approach are the same, namely the number of two clusters. The clustering process was carried out with two clusters for mathematics subjects to take a deeper look at the clustering results with the two methods. The results of the clustering process with the two methods are shown in Figure 6 and Figure 7.

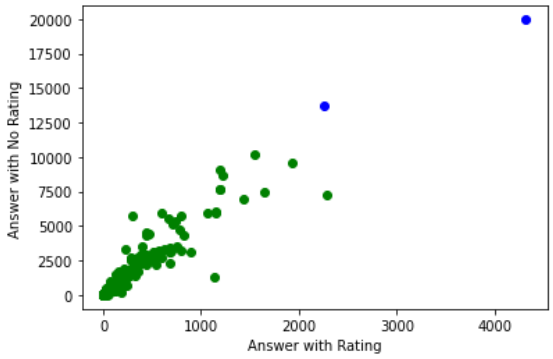

Figure 6. The result of clustering process Agglomerative Clustering (Average Linkage)

Based on Figure 6 and Figure 7, it can be seen that the blue color is the first cluster, and the green color is the second cluster. Although the Silhouette Score of the two methods is only slightly different, if viewed in more detail, the results of the clustering process using the K-Means method are much better than the Agglomerative Clustering method with Average Linkage approach. It can be seen from the results of the distribution of data from each cluster.

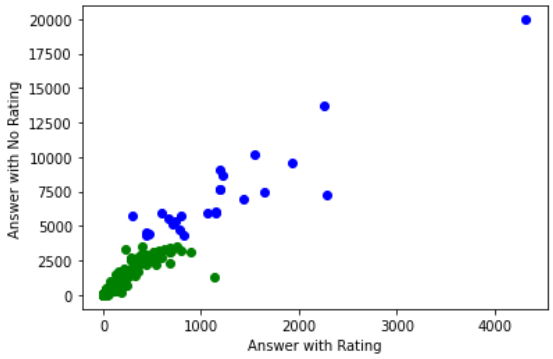

Figure 7. The result of clustering process K-Means

CONCLUSION

Based on the results of the comparison of methods that have been carried out, it can be concluded that the application of the K-Means Method for clustering user profile data is the best choice. This choice was based on testing using the Silhouette Coefficient method, which gave better results, namely 0.9081 compared to using the Agglomerative Clustering method with the Single Linkage approach getting results of 0.8810, Complete Linkage 0.8791, and Average Linkage 0.8990. Through clustering using the K-Means method, problems in grouping the reputation can be resolved more accurately. The optimal value of 0.9081 was obtained from 2 cluster numbers using the K-Means method.

ACKNOWLEDGEMENTS

The author thanks to the Faculty of Engineering and Information Technology, Universitas Jenderal Achmad Yani Yogyakarta, for the author's support in the form of internal research funding assistance in the 2021 Applied scheme and Brainly for granted access data.

REFERENCES

[1] H. Fu and S. Oh, "Quality assessment of answers with user-identified criteria and data-driven features in social Q&A," *Inf. Process. Manag.*, vol. 56, no. 1, pp. 14–28, Jan. 2019.

[2] Z. Liu and B. J. Jansen, "Identifying and predicting the desire to help in social question and answering," *Inf. Process. Manag.*, vol. 53, no. 2, pp. 490–504, Mar. 2017.

[3] I. Adaji and J. Vassileva, "Susceptibility of users to social influence strategies and the influence of culture in a Q&A collaborative learning environment," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10391 LNCS, pp. 49–64.

[4] K. Sagan, J. Colby, S. Y. Rieh, and E. Choi, "Beyond questioning and answering: Teens' learning experiences and benefits of social Q&A services," in *CSCW 2017 - Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017, pp. 295–298.

[5] A. Ayuningtyas, A. S. Honggowibowo, S. Mulyani, and A. Priadana, "A Web-Based Aircraft Maintenance Learning Media to Support Learning Process in Aerospace Engineering Education during the COVID-19 Pandemic," in *Proceeding - 2020 Sixth International Conference on e-Learning (econf)*, 2020, pp. 55–60.

[6] M. Li, Y. Li, Y. Lu, and Y. Zhang, "Evaluating Indicators of Answer Quality in Social Q&A Websites," *PACIS 2019 Proc.*, Jun. 2019.

[7] E. Choi *et al.*, "Utilizing content moderators to investigate critical factors for assessing the quality of answers on brainly, social learning Q&A platform for students: A pilot study,"

Proc. Assoc. Inf. Sci. Technol., vol. 52, no. 1, pp. 1–4, Jan. 2015.

[8] L. T. Le, C. Shah, and E. Choi, "Assessing the quality of answers autonomously in community question–answering," *Int. J. Digit. Libr.*, vol. 20, no. 4, pp. 351–367, 2019.

[9] P. W. Cahyo, K. Kusumaningtyas, and U. S. Aesyi, "A User Recommendation Model for Answering Questions on Brainly Platform," *J. Infotel*, vol. 13, no. 1, pp. 7–12, 2021.

[10] Y. Christian and J. Jimmy, "Perancangan Model Predıksı Performa Akademık Mahasıswa Menggunakan Algorıtma K-Means Clusterıng (Studı Kasus: Unıversıtas Xyz)," Mar. 2021.

[11] Y. H. Chrisnanto and G. Abdullah, "The uses of educational data mining in academic performance analysis at higher education institutions (case study at UNJANI)," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 11, no. 1, pp. 26–35, Mar. 2021.

[12] S. Dewi, S. Defit, and Y. Yunus, "Akurasi Pemetaan Kelompok Belajar Siswa Menuju Prestasi Menggunakan Metode K-Means (Studi Kasus SMP Pembangunan Laboratorium UNP)," *J. Sistim Inf. dan Teknol.*, pp. 28–33, Sep. 2020.

[13] P. W. Cahyo, "Klasterisasi Tipe Pembelajar Sebagai Parameter Evaluasi Kualitas Pendidikan Di Perguruan Tinggi," *Teknomatika*, vol. 11, no. 1, pp. 49–55, 2018.

[14] D. Exasanti and A. Jananto, "Analisa Hasil Pengelompokan Wilayah Kejadian Non-Kebakaran Menggunakan Agglomerative Hierachical Clustering di Semarang," *J. Tekno Kompak*, vol. 15, no. 2, pp. 63–75, Aug. 2021.

[15] L. Zahrotun, "ANALISIS PENGELOMPOKAN JUMLAH PENUMPANG BUS TRANS JOGJA MENGGUNAKAN METODE CLUSTERING K-MEANS DAN AGGLOMERATIVE HIERARCHICAL CLUSTERING (AHC)," *J. Inform.*, vol. 9, no. 1, Jan. 2015.

[16] R. O. Pratikto and N. Damastuti, "Klasterisasi Menggunakan Agglomerative Hierarchical Clustering Untuk Memodelkan Wilayah Banjir," *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 6, no. 1, pp. 13–20, Jan. 2021.

[17] Z. Arifin, S. Santosa, and M. A. Soeleman, "KLASTERISASI GENRE CERPEN KOMPAS MENGGUNAKAN AGGLOMERATIVE HIERARCHICAL CLUSTERING- SINGLE LINKAGE," *J. Cyberku*, vol. 13, no. 2, pp. 2–2, Dec. 2017.

[18] R. C. Pereira and T. Vanitha, "Web Scraping of Social Networks," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 3, no. 7, pp. 237–240, 2015.

[19] Fatmasari, Y. N. Kunang, and S. D. Purnamasari, "Web Scraping Techniques to Collect Weather Data in South Sumatera," in *Proceedings of 2018 International Conference on Electrical Engineering and Computer Science, ICECOS 2018*, 2019.

[20] M. R. E. Waluyo, P. Y. Saputra, and H. E. Dien, "KLASTERISASI WILAYAH TANAH LONGSOR BERDASARKAN DAMPAK WILAYAH DAN GEOGRAFIS MENGGUNAKAN METODE K-MEANS (Studi Kasus : Kabupaten dan Kota di Jawa Timur)," *Semin. Inform. Apl. Polinema*, Oct. 2020.

[21] U. A. Nasron and M. Habibi, "Analysis of Marketplace Conversation Trends on Twitter Platform Using K-Means," *Compiler*, vol. 9, no. 1, pp. 51–62, May 2020.

[22] A. I. Abdullah, E. Winarko, and A. Musdholifah, "Metode Boost-K-means untuk Clustering Puskesmas berdasarkan Persentase Bayi yang Diimunisasi," *JRST (Jurnal Ris. Sains dan Teknol.*, vol. 4, no. 2, p. 89, Nov. 2020.

[23] Suyanto, *Data Mining Untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika, 2017.

[24] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Fourth Edi. Cambridge, MA: Morgan Kaufmann Publishers, 2016.

[25] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, Fifth Edit. Springer, 2019.

[26] A. I. Abdullah, A. Priadana, M. Muhajir, and S. N. Nur, "Data Mining for Determining The Best Cluster Of Student Instagram Account As New Student Admission Influencer," *Telemat. J. Inform. dan Teknol. Inf.*, vol. 18, no. 2, pp. 255–266, Oct. 2021.

[27] H. L. Sun, K. P. Liang, H. Liao, and D. B. Chen, "Evaluating user reputation of online rating systems by rating statistical patterns," *Knowledge-Based Syst.*, vol. 219, p. 106895, 2021.