Jurnal Statistika Dan Komputasi (STATKOM), Vol. 4 No. 1 (June 2025)
ISSN : 2963-038X (print), ISSN : 2963-0398
DOI : https://doi.org/10.32665/statkom.v4i1.4825

STATKOM
JURNAL STATISTIKA DAN KOMPUTASI

Clustering Analysis of School Student Distribution in Bojonegoro Regency with Kernel K-Means

Ahmad Zakki Mubarok[1]

[1]Statistics Study Program, Universitas Nahdlatul Ulama Sunan Giri
E-mail: azakimuba@gmail.com[1]

Diajukan 4 Juni 2025 ***Diperbaiki*** 18 Juni 2025 ***Diterima*** 25 Juni 2025

Abstract

Background: *Education is a fundamental aspect that plays an important role in determining the progress of a nation. In Indonesia, education equity remains a major challenge, particularly in relation to the gap between urban and rural areas. Bojonegoro District, with its diverse geographical and social characteristics, reflects this issue of uneven educational access.*

Objective: *This study aims to examine the 2022 distribution of study groups at different education levels (kindergarten, elementary school, junior high school, high school, and vocational school) in Bojonegoro using descriptive statistics, normality testing, and the Kernel K-Means clustering algorithm.*

Methods: *Data were tested for normality using the Kolmogorov-Smirnov method. The clustering was performed by comparing four kernel types (Dot, Polynomial, Gaussian, and Sigmoid) to determine the most effective approach based on the Average Within Cluster Distance (AWCD) and the elbow method principle.*

Results: *The findings indicate that the distribution of study groups at the kindergarten and elementary school levels is relatively even and follows a normal pattern. In contrast, the distributions at the junior high school, high school, and vocational school levels remain uneven. The Kernel K-Means algorithm with the Dot kernel produced the most optimal results, identifying five main clusters that reflect regional disparities in educational participation.*

Conclusion: *This study demonstrates the novelty of applying Kernel K-Means in the educational context to uncover spatial disparities. The resulting clusters offer valuable insights into education inequality in Bojonegoro. These insights can inform policymakers in designing more targeted, equitable, and data-driven education policies.*

Keywords : *Education, Clustering, Kernel K-Means, Study Group Distribution.*

Abstrak

Latar Belakang: Pendidikan merupakan aspek fundamental yang berperan penting dalam menentukan kemajuan suatu bangsa. Di Indonesia, pemerataan pendidikan masih menjadi tantangan utama, terutama terkait dengan kesenjangan antara daerah perkotaan dan pedesaan. Kabupaten Bojonegoro, dengan karakteristik geografis dan sosial yang beragam, mencerminkan isu ketidakmerataan akses pendidikan ini.

Tujuan: Penelitian ini bertujuan untuk mengkaji distribusi rombongan belajar pada tahun 2022 di berbagai jenjang pendidikan (TK, SD, SMP, SMA, dan SMK) di Kabupaten Bojonegoro dengan menggunakan statistik deskriptif, uji normalitas, dan algoritma pengelompokan Kernel K-Means.

Metode: Data diuji normalitasnya dengan menggunakan metode Kolmogorov-Smirnov. Pengklasteran dilakukan dengan membandingkan empat jenis kernel (Dot, Polynomial, Gaussian, dan Sigmoid) untuk menentukan pendekatan yang paling efektif berdasarkan Average Within Cluster Distance (AWCD) dan prinsip metode Elbow.

Hasil: Temuan ini menunjukkan bahwa distribusi rombongan belajar di tingkat TK dan SD relatif merata dan mengikuti pola yang normal. Sebaliknya, distribusi di tingkat SMP, SMA, dan SMK masih belum merata. Algoritma K-Means dengan kernel Dot memberikan hasil yang paling optimal, dengan mengidentifikasi lima kelompok utama yang mencerminkan kesenjangan regional dalam partisipasi pendidikan.

Kesimpulan: Penelitian ini menunjukkan kebaruan penerapan Kernel K-Means dalam konteks pendidikan untuk mengungkap kesenjangan spasial. Cluster yang dihasilkan menawarkan wawasan yang berharga tentang ketidaksetaraan pendidikan di Bojonegoro. Wawasan ini dapat menjadi masukan bagi para pembuat kebijakan dalam merancang kebijakan pendidikan yang lebih tepat sasaran, adil, dan berbasis data.

Kata kunci: Pendidikan, Klasterisasi, Kernel K-Means, Distribusi Grup Belajar

INTRODUCTION

Education is a fundamental aspect of human life and a key driver of national progress. Every individual has the right to obtain proper and equitable education. However, educational inequality remains a persistent issue in Indonesia, particularly between urban and rural areas (Fitri, 2021). espite various policy efforts, disparities in access, resources, and outcomes still hinder the goal of equitable education.

Numerous studies have criticized Indonesia's education system as rigid and ineffective, resulting in low performance compared to other countries (Kurniawati, 2022). The problem does not lie in the system itself, which is structurally similar to systems used in other nations, but in its inconsistent implementation at the regional level (Fitri, 2021). These implementation gaps have created a mismatch between educational goals and actual outcomes in the field.

Bojonegoro Regency is a representative case of this issue, facing challenges such as limited educational infrastructure, unequal resource allocation, and disparities between rural and urban schools. Addressing these challenges requires a data-driven approach to better understand how educational resources and student populations are distributed across regions.

Clustering methods, especially Kernel K-Means, offer a promising solution for analyzing complex and non-linear patterns in educational data. Previous studies have also demonstrated the effectiveness of clustering techniques, such as K-Means, in mapping regional disparities based on economic indicators, for example, grouping Indonesian cities by Consumer Price Index (CPI) variations during the COVID-19 pandemic using K-Means and multiple regression methods (Azizah & Athoillah, 2021).

Unlike traditional K-Means, Kernel K-Means can project data into higher dimensions using kernel functions, allowing for more accurate grouping of sub-districts with similar characteristics. Prior work by Nurdiansyah et al. (2024), demonstrated the effectiveness of Kernel K-Means in analyzing population-based clustering, outperforming other approaches such as Fast K-Means, K-Medoids, and Fuzzy C-Means.

Most recent studies still rely on classic K-Means or other linear-based clustering methods, such as Sholikhah (2022), Khan et al. (2023), Nurdiansyah et al. (2023), Fitriyah et al. (2023), Muiz (2024) and Safitri (2024), which may fail to capture the true complexity of educational disparities. This presents a research gap in applying non-linear clustering algorithms, such as Kernel K-Means, to the education sector, especially in Indonesia.

This study aims to address that gap by analyzing the distribution of study groups across various education levels in Bojonegoro Regency. Specifically, it evaluates the statistical distribution of school group data, tests for normality using the Kolmogorov-Smirnov method, and performs clustering using the Kernel K-Means algorithm implemented via RapidMiner. The algorithm's performance is evaluated using the Average Within Cluster Distance and the Elbow method.

The dataset used is novel, as it has not been previously analyzed using statistical distribution tests nor clustering based on student group numbers across education levels.

The urgency of this research lies in the need for accurate student distribution analysis to prevent educational inequality. The clustering results will be useful for academics in advancing education analytics, for policymakers in planning equitable resource distribution, and for the public in promoting transparency and equal access to education. By applying Kernel K-Means, this research supports more effective, fair, and data-driven decision-making in education planning.

METHOD

Research Design

This research uses a descriptive quantitative approach with an explorative method. The aim is to map school group data, test the normal distribution of the data, and cluster sub-districts in Bojonegoro Regency based on educational characteristics using the Kernel K-Means algorithm, which is one of the unsupervised learning techniques in data mining. The statistical methods used include a normality test and a non-linear clustering algorithm.

Population and Sample

The population in this study comprises all sub-districts in Bojonegoro Regency that have education data from kindergarten to senior high school. The year 2022 was chosen because it represents the most recent and complete dataset available from the Education Office and reflects the latest post-pandemic educational conditions in the region. The sample used is secondary data obtained from the Satu Data Bojonegoro Website, specifically from the Education Office, consisting of the number of study groups across five levels of education: kindergarten, elementary school, junior high school, senior high school, and vocational high school. These five levels were selected because they represent the complete spectrum of formal education in Indonesia and align directly with the objective to assess distributional equity.

Sampling Techniques

As this study uses secondary data, no primary data collection or sampling technique was applied. The dataset is considered comprehensive and representative of the education conditions in Bojonegoro Regency in 2022.

Research Subjects

The variables used in this study are presented in Table 1 below.

Table 1. Definition of Research Variables.

Variable	Description	Measurement Scale
X_1	Number of kindergarten students in 2022	Ratio
X_2	Number of elementary school students in 2022	Ratio
X_3	Number of junior high school students in 2022	Ratio
X_4	Number of senior high school students in 2022	Ratio
X_5	Number of vocational high school students in 2022	Ratio

Data Analysis Techniques

Data analysis techniques were carried out through the following stages:

1) Descriptive Statistics: calculating the mean and standard deviation of the number of students in each school group from 28 sub-districts in Bojonegoro Regency.
2) Normality Test: applying the Kolmogorov-Smirnov statistical test to assess whether the data is normally distributed.
3) Clustering: using the Kernel K-Means algorithm, which extends conventional K-Means by applying a kernel function that transforms data into a higher dimension to capture non-linear patterns. The performance of the clusters is measured using Within-Cluster Sum of Squares (WCSS).
4) Descriptive Statistics of Formed Clusters: calculating the mean and standard deviation of student numbers for each cluster of sub-districts.

In applying the clustering method in In applying the clustering method via RapidMiner software, the process design is shown in Figure 1, involving several operators such as: Retrieve, Normalize, Multiply, Clustering (K-Means Kernel), Data to Similarity, Multiply (second), and Performance (Sholikhah, 2022) (Nurdiansyah et.al, 2023).

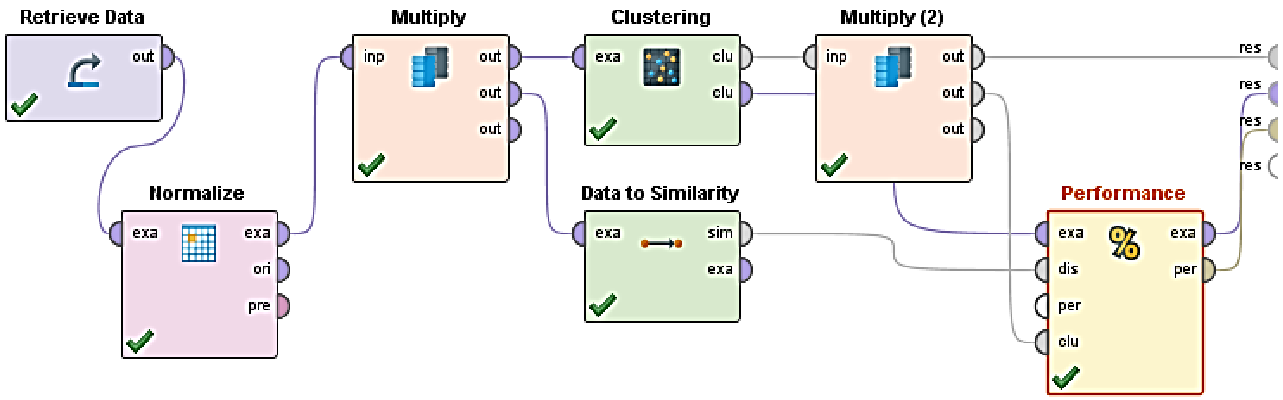

Figure 1. Process Design of Clustering method

In Figure 1, the tools or operators used in this study are illustrated. "Retrieve" is used to input the dataset. "Normalize to Standardize" adjusts the data to a standard normal distribution. "Multiply" splits the process into parallel workflows. "Clustering" applies the Kernel K-Means algorithm. "Data to Similarity" computes pairwise similarity values. "Performance" evaluates the resulting clusters. The "Clustering" operator supports several kernel types: Dot Product Kernel (Linear), Polynomial Kernel, Radial Basis Function (RBF) or Gaussian Kernel, and Sigmoid Kernel (Tanh Kernel).

The process of comparing clustering methods is carried out by analyzing the performance vector values and identifying the optimal number of clusters (k) based on the Average Within Cluster Distance (AWCD) for each kernel type. The selection of the best model follows the Elbow method principle (Fitriyah et.al, 2023), by identifying the point at which the rate of decrease in AWCD begins to plateau, indicating an optimal clustering solution.

RESULTS AND DISCUSSION

Descriptive Statistics

Based on the results of observations in 2022, descriptive statistics were obtained regarding the number of study groups in Bojonegoro Regency, including kindergarten, primary, junior high, senior high, and vocational levels. A summary of these results is presented in Table 2.

Table 2. Descriptive Statistics of Study Groups by Education Level in Bojonegoro (2022)

Strudy Group	N	Minimum	Maximum	Mean	Std. Deviation
X_1	28	7	50	23.86	11.414
X_2	28	11	48	25.86	9.571
X_3	28	1	17	3.89	3.370
X_4	28	0	8	1.75	1.669
X_5	28	0	12	2.18	2.480

Based on Table 2, the number of study groups varies greatly between education levels. At the kindergarten and primary school levels, the average number of study groups per sub-district is quite high, at 23.86 and 25.86 respectively, with a relatively large standard deviation (11.414 and 9.571), indicating a wide distribution. Meanwhile, the number of study groups at the junior high school, senior high school and vocational school levels is much smaller with an average of 3.89; 1.75; and 2.18 respectively. There are even sub-districts that have no study groups at the senior high school and vocational school levels (minimum value of 0). This indicates that the distribution of primary education is more equitable than secondary education, which still faces challenges in equalizing access across sub-districts.

Normality Test

The Kolmogorov-Smirnov test of the Normal distribution was conducted to determine whether the distribution of data on the number of study groups in Bojonegoro Regency, which includes kindergarten, primary school, junior high school, senior high school, and vocational school, is evenly distributed or follows the Normal distribution. The results of this test are summarized in Table 3.

Table 3. Summary of Kolmogorov-Smirnov Test Results for Normality of Data on the Number of Study Groups.

Study Group	Statistic of χ^2	P-value
X_1	0.125	0.200
X_2	0.120	0.200
X_3	0.273	0.000
X_4	0.245	0.000
X_5	0.207	0.003

Based on the Kolmogorov-Smirnov test results summarized in Table 3, it is known that the data on the number of study groups at the kindergarten and primary school levels has a p-value of 0.200, which is greater than the 0.05 significance level. This indicates that the distribution of data at both levels follows a normal distribution. In contrast, the data at the junior high, high school, and vocational school levels showed p-values of 0.000, 0.000, and 0.003, respectively, which are smaller than 0.05, so they are not normally distributed. Thus, it can be concluded that only kindergarten and primary school data fulfill the assumption of normality, while the other levels show deviations from the normal distribution, which supports the use of non-parametric or non-linear statistical analysis such as Kernel K-Means clustering.

Clustering

Comparison of clustering methods is done by analyzing the average value of the distance between the best clusters obtained from each method. The evaluation results of Kernel K-Means clustering with different types of kernels are summarized in Table 4 below.

Table 4. Results of Clustering Method Comparison Process

	Average Within Cluster Distance			
	K-Means Kernel			
k	Dot	Polynomial	Gaussian	Sigmoid
2	-33.111	-56.480	-29.644	-24.731
3	-12.738	-32.383	-23.447	-15.705
4	-10.137	-16.213	-18.776	-11.964
5	-6.716	-15.858	-10.509	-16.030
6	-5.450	-7.539	-7.337	-11.964
7	-4.918	-13.303	-6.712	-10.444
8	-3.551	-8.679	-4.911	-9.970
9	-3.945	-3.635	-6.616	-5.795
10	-3.294	-4.110	-4.211	-5.795

Table 4 shows that the Dot kernel generally gives the best results compared to other kernels. At almost all values of k (number of clusters from 2 to 10), the Dot kernel produces average distance values closest to zero, especially at k ≥ 5. For example, at k=10, the Dot kernel has a value of -3.294, closest to zero compared to Polynomial (-4.110), Gaussian (-4.211), and Sigmoid (-5.795). In contrast, the Polynomial Kernel, despite providing the lowest values (e.g. -56.480 at k=2), performs poorly as its values are far from zero, especially for small to medium cluster counts. The Gaussian and Sigmoid kernels have fluctuating performance, but tend to produce larger (more negative) values, making them less optimal in this context. Therefore, based on this evaluation, the Dot Kernel can be considered as the most consistently effective clustering method, as it produces values closest to zero for almost all variations in the number of clusters.

In Figure 2, based on the principle of the Elbow method, the Dot Kernel also shows the clearest elbow pattern. The within cluster distance value of the Dot Kernel decreases sharply from k=2 to k=4, from -33.111 to -10.137, but after k=4, the decrease slows down significantly. This supports the selection of k=5 as the most optimal cluster count.

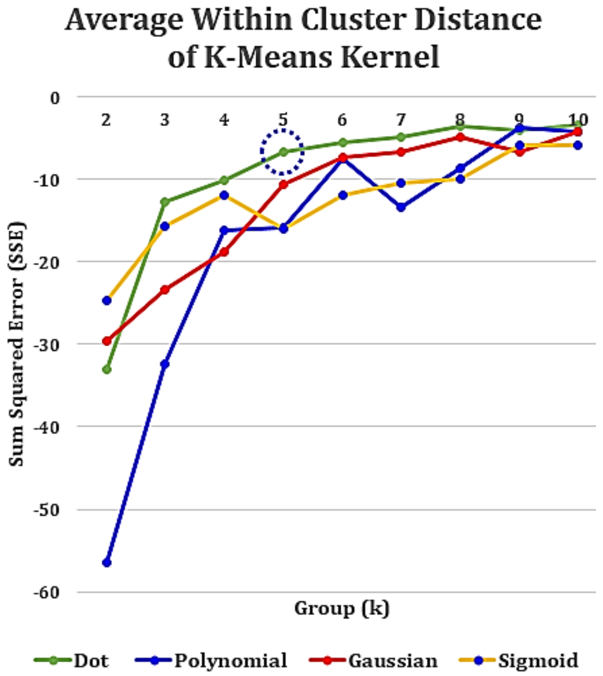

Figure 2. Process Design of Clustering method

Descriptive Statistics for Each Formed Cluster

The clusters formed can be used as a reference for splitting the data. This process helps in determining the distribution of each cluster. By understanding the distribution, the specific characteristics of each cluster can then be identified.

Based on Table 5, which presents the results of descriptive statistics for each cluster formed, an overview of the unique characteristics of each cluster based on the five levels of education can be obtained: Kindergarten, elementary school, junior high school, senior high school, and vocational school. Cluster 0 shows moderate mean values for kindergarten and primary school (17.40 and 20.80 respectively), but very low for junior high school and above. This indicates that this group is dominated by individuals with a strong primary education background but few who continue to higher levels, especially SMK (average of only 1.20). Cluster 1 is similar, but has higher mean scores for primary and kindergarten (22.67 and 33.33), indicating that this group tends to be stronger in primary education, but has almost no representation in senior high school and vocational school.

Cluster 2 stands out with the highest averages across all levels (e.g. TK: 42.40; SD: 39.80; SMP: 9.60), suggesting this group has very high and evenly distributed education participation from TK to SMK. This could reflect areas or populations with good access and continuity of education. Cluster 3 has medium scores across all levels, with kindergarten and primary school averages above 25, and junior high school to vocational school in the low to medium range, indicating a group with fairly equitable access to education, although not as high as Cluster 2.

Finally, Cluster 4 shows low averages across all levels, especially from junior high school onwards (SMA: 0.38; SMK: 0.75), depicting a group with very limited educational participation, especially at the secondary level. This finding is critical for stakeholders and policymakers, as it highlights the need for targeted intervention in areas represented by Cluster 4.

Table 5. Results of Descriptive Statistics for Each Formed Cluster

Cluster		N	Minimum	Maximum	Mean	Std. Deviation
cluster_0	X_1	5	12	23	17.40	4.278
	X_2	5	18	25	20.80	3.421
	X_3	5	2	4	3.00	.707
	X_4	5	1	2	1.40	.548
	X_5	5	0	4	1.20	1.643
cluster_1	X_1	3	17	29	22.67	6.028
	X_2	3	31	37	33.33	3.215
	X_3	3	3	5	3.67	1.155
	X_4	3	1	1	1.00	.000
	X_5	3	0	1	.67	.577

Cluster		N	Minimum	Maximum	Mean	Std. Deviation
cluster_2	X_1	5	37	50	42.40	4.775
	X_2	5	34	48	39.80	5.215
	X_3	5	6	17	9.60	4.393
	X_4	5	3	8	4.20	2.168
	X_5	5	4	12	6.00	3.464
cluster_3	X_1	7	25	35	28.43	3.552
	X_2	7	24	32	27.86	2.673
	X_3	7	2	4	3.43	.787
	X_4	7	1	3	2.14	.900
	X_5	7	2	3	2.43	.535
cluster_4	X_1	8	7	18	12.75	4.464
	X_2	8	11	23	15.75	4.496
	X_3	8	1	2	1.38	.518
	X_4	8	0	1	.38	.518
	X_5	8	0	2	.75	.707

CONCLUSION

Conclusion

Based on the results of descriptive statistical analysis and clustering process on the number of study groups in Bojonegoro Regency in 2022, it is found that the distribution of basic education (kindergarten and primary school) tends to be more evenly distributed and higher than that of secondary education (junior high school, senior high school, and vocational school). This can be seen from the higher average number of study groups and the normal distribution of data at the kindergarten and primary school levels. In contrast, the junior secondary level and above shows a low average value and non-normal data distribution, indicating inequality in access to further education in some sub-districts. Through the clustering process using the Kernel K-Means method, the Dot Kernel proved to be the best method based on the average distance value within clusters that was closest to zero and a clear elbow pattern at k=4. The clustering results also identified five groups of areas with different educational characteristics, ranging from groups with high and equitable education participation to groups with limited access to education, especially at the secondary level. The application of the Kernel K-Means method in this study demonstrates its novelty and significance in identifying educational disparities, particularly due to its ability to handle non-linear patterns in multidimensional data. This provides added value in the context of educational planning and regional development.

Suggestions

Based on the findings of this study, it is recommended that relevant parties, especially local governments and education agencies, use the clustering results as a reference in planning and equitable access to education. Areas classified as clusters with low secondary education participation (such as Cluster 4) need to receive special attention, both in the form of educational infrastructure development, improving teaching staff, and affirmative policies to improve education continuation. Additionally, the Dot Kernel-based clustering approach should be considered a strategic tool in routine education data monitoring and evaluation. This method offers a more robust and adaptable framework for supporting data-driven, equitable, and region-specific educational policymaking. In the future, similar research can also be developed by considering other variables such as student numbers, teacher ratios, or socioeconomic conditions to enrich analysis and policy recommendations.

BIBLIOGRAPHY

Azizah, F., & Athoillah, M. (2021). Analisis dampak Covid-19 terhadap indeks harga konsumen dengan K-Means dan regresi berganda. *Indonesian Journal of Applied Statistics*, 4(1), 21–33. https://doi.org/10.13057/ijas.v4i1.46329

Fitri, S. F. N. (2021). Problematika Kualitas Pendidikan di Indonesia. *Jurnal Pendidikan Tambusai*, 5(1), 1617–1620. https://jptam.org/index.php/jptam/article/view/1148

Fitriyah, H., Safitri, E. M., Muna, N., Khasanah, M., Aprilia, D. A., & Nurdiansyah, D. (2023). Implementasi Algoritma Clustering dengan Modifikasi Metode Elbow untuk Mendukung Strategi Pemerataan Bantuan Sosial di Kabupaten Bojonegoro. *Jurnal Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, 4(3), 1598-1607. https://doi.org/10.46306/lb.v4i3.453

Khan, A. S. S., Fatekurohman, M., & Dewi, Y. S. (2023). Perbandingan Algoritma K-Medoids Dan K-Means Dalam Pengelompokan Kecamatan Berdasarkan Produksi Padi Dan Palawija Di Jember. *Jurnal Statistika Dan Komputasi*, 2(2), 67–75. https://doi.org/10.32665/statkom.v2i2.2301

Kurniawati, F. N. A. (2022). Meninjau Permasalahan Rendahnya Kualitas Pendidikan Di Indonesia Dan Solusi. *Academy of Education Journal*, 13(1), 1–13. https://doi.org/10.47200/aoej.v13i1.765

Muiz, R. A. (2024). Comparison of K-Means and Fuzzy C-Means for Optimizing Tuberculosis Management and Healthcare Service Allocation in Bojonegoro. *Jurnal Statistika Dan Komputasi*, 3(2), 80–91. https://doi.org/10.32665/statkom.v3i2.3532

Nurdiansyah, D., Ma'ady, M., Sukmawaty, Y., Utomo, M., & Mutiani, T. (2024). Clustering Analysis For Grouping Sub-Districts In Bojonegoro District With The K-Means Method With A Variety Of Approaches. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 18(2), 1095-1104. https://doi.org/10.30598/barekengvol18iss2pp1095-1104

Nurdiansyah, D., Saidah, S., & Cahyani, N. (2023). Data Mining Study For Grouping Elementary Schools In Bojonegoro Regency Based On Capacity And Educational Facilities. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 17(2), 1081-1092. https://doi.org/10.30598/barekengvol17iss2pp1081-1092

Quraisy, A. (2020). Normalitas Data Menggunakan Uji Kolmogorov-Smirnov dan Saphiro-Wilk. *Journal of Health, Education, Economics, Science, and Technology*, 3(1), 7-11. https://doi.org/10.36339/

Safitri, E. M. (2024). Clustering Study Of Hospitals In Bojonegoro Based On Health Workers With K-Means And K-Medoids Methods. *Jurnal Statistika Dan Komputasi*, 3(2), 92–102. https://doi.org/10.32665/statkom.v3i2.3592

Sholikhah, N. A. (2022). Studi Perbandingan Clustering Kecamatan di Kabupaten Bojonegoro Berdasarkan Keaktifan Penduduk Dalam Kepemilikan Dokumen Kependudukan. *Jurnal*

Statistika Dan Komputasi, 1(1), 42–53. https://doi.org/10.32665/statkom.v1i1.443