Cluster Analysis of School Data in Jakarta

Nathaniel Susianto Sutanto
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
nathaniel.sutanto@binus.ac.id

Marcell Kurniawan Sutanto
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
marcell.sutanto@binus.ac.id

Daniel Aditya Tumansery
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
daniel.tumansery@binus.ac.id

Karli Eka Setiawan
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
karli.setiawan@binus.ac.id

Muhammad Fikri Hasani
Computer Science Department,
School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
muhammad.fikri003@binus.ac.id

Abstract*—School clustering could potentially be a strategy to address the challenges of resource allocation and educational quality around the world. This study focuses on school clustering in Jakarta, the capital city of Indonesia, where the education quality has yet to meet the demand for high quality human resources. To the best of our knowledge, this is the first study to cluster schools in Jakarta. The objective of this research is to determine the effectiveness of school clustering in Jakarta when applied to government data available publicly. The clusters formed would then be analyzed and compared to other variables such as the official accreditation by the government. An experimental approach was used involving two individual methods, namely k-means and DBSCAN clustering to cluster both continuous and discrete data. Kmeans clustering produced 2 clusters with cluster 1 containing higher standard school compared to cluster 0. Meanwhile, the implementation of DBSCAN algorithm failed to produce any meaningful clusters even after parameter tuning. The failure of DBSCAN indicated that the school data is distributed evenly which caused the density-based algorithm to fail at clustering. The research concluded that school data of Jakarta are very highly distributed and it is challenging to effectively cluster the data. Future research could use more specific data according to city or education level which could potentially increase the effectiveness of cluster analysis.

***Keywords*—clustering, k-means, schools, clusters**

I. INTRODUCTION

The quality of education is a crucial factor to determine the future of a country and its people. To educate the country is one of the main goals in the 1945 constitution of Indonesia. Indonesia's government has been working to improve the education system, through various methods. Jakarta, as the capital city of Indonesia, has been a major focus of these efforts. A research by Zahroh [1] concluded that as of right now, the high Human Development Index in some regions in Indonesia does not reflect the educational quality. This means that education in Indonesia, even though having a good HDI index, still needs improvement. Dinatus Solichah mentioned in his research [2] that one of the supports for student interest is the facility, there is a significant relationship between school facilities and interest in learning with study habits. School facilities and the social environment have a significant effect on student learning outcomes [3]. Besides that, the learning atmosphere is also influenced by the existence of complete and adequate school facilities so that it can encourage students to actively develop their potential [3].

As part of this ongoing effort to improve education in Jakarta, the government collects and provides various data related to schools in the city, including student demographics, teacher qualifications, and academic performance. This data could be found in the ministry of education's data website. However, with many schools in the city, analyzing and understanding this data proves to be a daunting task. This is where clustering techniques can be useful.

According to Ghosal et al. clustering is a part of unsupervised learning and is more challenging than classification problems [4]. Clustering is a data analysis technique that aims to group data points into clusters to discover their natural groupings. In the context of school data, clustering can help identify patterns and trends that might not be obvious at first glance. For example, clustering can reveal which schools have similar student demographics or which schools have similar academic performance.

This paper discusses an analysis of school data in Jakarta using clustering techniques. Specifically, k-means clustering is utilized to group schools based on infrastructure numbers such as classrooms and libraries and also the number of staff and teachers from these schools. We then analyze the results to gain insights about the education sector in Jakarta and identify areas where further improvement is needed. Overall, our study highlights the potential of clustering techniques as a powerful tool for analyzing school data and improving education in Jakarta through improvements within the clusters.

This research would be done in four main stages, data collection, data preprocessing, data analysis and results. Data will be collected from the ministry of education's website and preprocessed accordingl. Data would then be analyzed by forming clusters with various clustering techniques, this cluster analysis would mainly be done with the python programming language. Finally results could be concluded with the clusters formed and recommendations could be made to improve the schools within each cluster.

II. RELATED WORKS

In the education sector, there has been numerous research utilizing cluster analysis methods to reveal information through subgroups as a result of the clustering methods. Clustering has been used to group both the students and the schools itself within a certain area. The most popular algorithm used in researches is the k-means algorithm, this is due to the flexibility amd simplicity of using k-means. K-means clustering also works efficiently with both small and large datasets.

One such research that utilizes k-means was the clustering of school building data to find the most efficient way to save energy conducted by Marrone et al. [5]. Their research led to 2 significant clusters of schools in the Lazio region, Central Italy, these clusters are then analyzed using an R2 and p-value. The centroid of those clusters was later compared to a real-life school model in the region for further analysis of the cost-saving measures to be taken. The 2 clusters formed are proven to be optimal because when clusters were added, the individual R2 scores would be lower, this is probably due to their low data count of only 60 data used out of 80 due to missing values. Another research that utilizes this method is the clustering of students' grades to determine a major that they should pick by Irawan Y. [6]. They picked out 141 student samples in SMA negeri 1 Pangkalan Kerinci for their research. The grades are picked out of 8 subjects from each student. The results consists of 2 clusters, 62 students in the high category score, and 79 students in the low category score. They mentioned that it still needs to be adjusted between the data and reports with the system since it is still a new system in SMA negeri 1 Pangkalan Kerinci.

Many previous researchers frequently employ the well-liked clustering algorithms k-means, BIRCH, and DBSCAN, such as the research by Nafuri et al [7]. These methods had benefits and drawbacks, but they may create clusters that perform satisfactorily. Finding excellent cluster results can be challenging since it depends on how the grouping algorithm's parameters are adjusted, which is one of the limitations mentioned by the reviewer above. Križani´c analyzed student conduct that was recorded in the e-learning system and could affect examination performance [8]. K-means and decision trees are two of the data mining techniques employed. Based on how the students behaved in e-learning, the cluster analysis divided the students into three groups, and three decision tree models were created as a result. The element that produced the greatest knowledge gain was midterm exam performance. Lower exam performance would result from infrequent use of online learning and lecture materials. Using hierarchical and k-means methods, PanduRanga Vital et al. examined student performance [9]. The methods used have been shown to be successful at predicting students' course success. Hierarchical clustering has offered major contributing factors that effects student results through relationships in dendrograms, such as extracurricular activity, attendance, and number of failed classes.

Similarly, a study by Govindasamy et al.[10] compared four clustering algorithms: fuzzy c-min, k-medoids, k-means, and expectation maximization (EM). The performance of the students on the final exam for the semester was predicted using the data of 1531 college students. The study's findings showed that, although the implementation time was longer, fuzzy c-means and EM had better clustering quality in terms of purity and normalized mutual information (NMI). Seven clustering techniques have been used by Navarro et al. [11] to analyze educational datasets with multiple student success grades. According to the study's results, the best division strategies were k-means and PAM, while the best hierarchy techniques was DIANA. Another research comparing clustering algorithms was done by Trivedi which compared four different algorithms: k-means, DBSCAN, hierarchical and affinity propagation [12]. The results from this research did not show the best cluster result, but instead the researchers suggested that the school could use any one of the resulting clusters to be applied to the students as seen fit. This means that the best clusters formed could only be judged by the teachers who have personal experience with the students themselves and thus, could utilize one of the created clusters.

III. METHODOLOGY

Dataset used for this research is primarily sourced from the Indonesian ministry of education’s website. Data from various schools in Jakarta was obtained by web-scraping and also through available datasheets. Afterwards, two types of data are combined, accreditation data and general school information data. The combined data can be seen in Table 3.1. Accreditation data consists of the school’s accreditation from the government. While the general school information contains general info of the school. Other relevant information are also added such as school district/city which can help for further analysis of which area has a better education in general. In total, there are 8936 schools from Jakarta ranging from pre-school to high school and even special needs school. But of all the schools in the compiled dataset, only some of these schools will be considered for clustering, for example, pre-schools will be excluded since pre-schools are not accredited by the government. Special needs schools are also excluded because they may have more staff than normal schools to attend to the special students’ needs.

Col Name	Description	Col Name	Description
District	Categorical	Last Sync	Data for government website
Subdistrict	Categorical	Sync amount	Data for government website
School Name	String	Students	Continuous count
School ID	String	Study Group	Continuous count
Education Level	Categorical	Teacher	Continuous count
Private/Public	Boolean categorical	Staff	Continuous count
Accreditation Year	Continuous integer	Classroom	Continuous count
Accreditation Score	Continuous integer from 0-100	Laboratory	Continuous count
Accreditation rank	Categorical (A,B,C)	Library	Continuous count

Table 3.1 Dataset description

Pre-processing of the data will mainly focus on data cleansing and feature selection. Data such as school name and school ID will not be used in the final clustering since it is not relevant. Categorical values will be encoded accordingly. District and city data will not be used in clustering but will instead be used to compare each city’s schools. It could determine whether schools in different districts or cities have different standards. Afterwards, data will be normalized using min-max scaler.

K-means clustering is a popular unsupervised learning algorithm that processes a given dataset into distinct groups, or clusters based on similarities. According to Sinaga et al., k-means clustering is a probability-based model and nonparametric [13]. This is because the algorithm finds clusters through updating probabilistic weights in the process of creating the predetermined cluster number. K-means must be initialized with a set amount of clusters. In order to find the optimal number of n clusters, there are some methods such as the elbow method, silhouette method and GridSearch [14]. All of these methods would usually return a similar value of k.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that defines clusters based on connected density points[15]. It does not require a predetermined number of clusters but relies on two inputs: Eps, which determines the maximum distance for points to be considered in the same cluster, and MinPts, the minimum number of points within Eps for a point to be classified as a core point. DBSCAN forms clusters by assigning each point a circle with a radius of Eps and evaluating its membership based on the number of points within that circle. The algorithm operates on the concept of data density, categorizing points as noise, core, or border based on their proximity to other points. The number of resulting clusters depends on the chosen Eps and MinPts parameters, and the Euclidean distance is commonly used to measure the distance between points during clustering.

The scoring method used to score clustering are different from the ones used in classification. This is due in part to clustering not having a set value to determine its correctness. Silhouette score is one scoring technique to determine if clustering is successful. Silhouette score ranges from -1 to 1 with a score of 1 indicating that clustering is highly successful and -1 meaning clusters are assigned incorrectly[16].

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)}$$

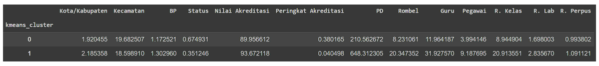
Figure 3.1 Silhouette score formula

Silhouette score utilizes inter cluster distance and intra cluster distance. In figure 3.1, ‘b’ represents the inter cluster distance while ‘a’ represents average inter cluster distance. A cluster with small intra cluster distance with large inter cluster distance should have a relatively high silhouette score.

The other scoring method used is Davies-Bouldin index. It evaluates how well the clustering is through internal calculations[17]. The formula calculates the average distance of every point to its centroid. The lower the value, the better its results since it means that the cluster are more tightly packed and indeed forms a cluster

IV. RESULT AND DISCUSSION

In our analysis of school data in Jakarta using clustering techniques, specifically k-means clustering, we identified distinct clusters based on infrastructure numbers such as classrooms and libraries, as well as the number of staff and teachers in each school. Through the clustering process, we were able to differentiate the schools in Jakarta into 2 different clusters.



	Kota/Kabupaten	Kecamatan	BP	Status	Nilai Akreditasi	Peringkat Akreditasi	PD	Rombel	Guru	Pegawai	R. Kelas	R. Lab	R. Perpus
kmeans_cluster													
0	1.920455	19.682507	1.172521	0.674931	89.956612	0.380165	210.562672	8.231061	11.964187	3.994146	8.944904	1.698003	0.993802
1	2.185358	18.598910	1.302960	0.351246	93.672118	0.040498	648.312305	20.347352	31.927570	9.187695	20.913551	2.835670	1.091121

Table 4.1 Data distribution of each cluster using k-means

In figure 4.1, it could be seen that cluster 1 includes the better schools compared to cluster 0. For instance, cluster 1 comprised schools with ample infrastructure and a high ratio of staff to students, suggesting a higher level of resources and potentially better overall educational quality. In contrast, cluster 0 consisted of schools with limited infrastructure and a lower staff-to-student ratio, indicating areas where improvement and investment may be needed.

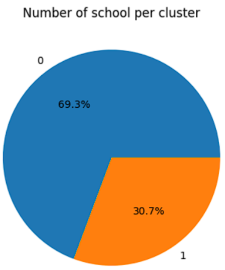


Figure 4.1 Proportion of cluster members using k-means

Our clustering method divides the 2 clusters into a proportion of 69.3% and 30.7% for cluster 0 and cluster 1 respectively. This indicates that there is still a high number of schools with relatively lower standards in Jakarta. The school proportion could later be used as the base proportion for comparing clusters with other attributes.

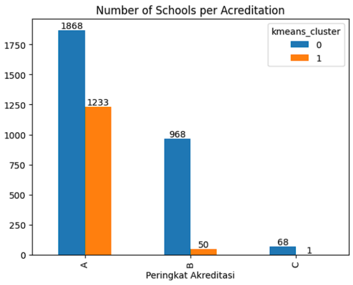


Figure 4.2 Comparison of k-means cluster members based on accreditation

From this chart, the clustering can successfully cluster schools according to their accreditation, with accredited A schools having the highest proportion of cluster 1 schools with 39%. While both accredited B and C schools are mostly cluster 0 schools with a proportion of 95% and 98.55% respectively.

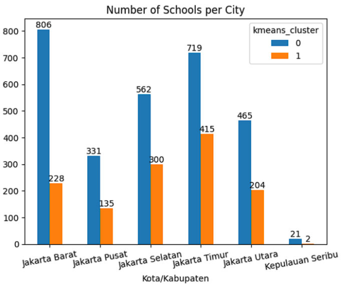


Figure 4.3 Comparison of k-means cluster members based on city

From figure 4.3, certain areas contain a higher proportion of cluster 1 schools than the base proportion. Cities such as Jakarta Selatan and Jakarta Timur both contain a proportion of 34.8% and 36.6%. While other areas still need to improve their school infrastructure or accreditation score in order to match or exceed the base proportion as found from the clustering.

Overall, the k-means clustering obtained a silhouette score of 0.3704 and a davies-bouldin index of 1.0951. These results indicate that the clustering still needs improvement since the silhouette score is near 0, while db index is far from 0, which is the ideal score for a perfect cluster.

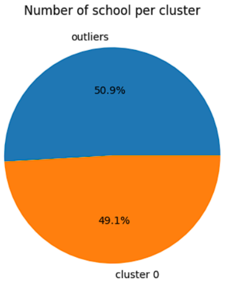


Figure 4.4 Proportion of clustered members to outliers using DBSCAN

As for the DBSCAN clustering, the algorithm failed to determine clusters for this dataset. Various epsilon and MinPts values have been used in order to obtain a good cluster, but the algorithm can only make 1 cluster (cluster 0) and keep the other values as -1 or outliers.The clustering proportion can be seen in Figure 4.4. This indicates that our dataset is not dense enough for the algorithm to create multiple clusters, instead it could only form 1 cluster that is within distance of each other while the rest of the data that are further away was not captured in a cluster.

One of the factors that caused the error on running DBSCAN clustering is the size of the data, currently the data holds about 40000 entries. Reducing the data by making clusters based on smaller groups such as education level, city, or even district could probably yield better silhouette scores and also have denser, more sparse data that could be used by DBSCAN to create clusters. Other preprocessing methods such as dimensionality reduction or standard scaler can also be used alongside outlier removal to create a standard distributed dataset, but it comes with a drawback of having to exclude some schools from the analysis.

Other than that, there are various other clustering algorithms that could effectively be used to cluster both categorical and continuous data such as twostep or Bernoulli mixture models that utilize a mixture of multiple machine learning models. Different distance metrics could also be utilized although their effects could be minimal towards the final cluster.

V. CONCLUSION

In conclusion, our analysis revealed that the clustering technique obtained a decent accuracy score, but fails to find true clusters based on density. This indicates that the data still has too much noise and could still be improved through more preprocessing. Overall, cluster analysis techniques could provide a powerful tool for understanding the education landscape in Jakarta and identifying areas for targeted improvement. By grouping schools based on shared attributes, we could gain a deeper understanding of the challenges and opportunities within each cluster. This information can assist them in making informed decisions to enhance the quality of education in Jakarta.

Furthermore, future research can help optimize this dataset to create more accurate models in clustering schools or find ways to optimize the ways that the data can be used. With the availability of other public data, researchers and government officials can hopefully work together in finding trends of Jakarta, or even Indonesia's current education trend. Other than that, there are also other clustering algorithms such as fuzzy c-means that could potentially be used to conduct school data cluster analysis. By utilizing the insights gained from the data and cluster analysis, hopefully more work can be done towards a more equitable and effective education system in Jakarta, ultimately benefiting the students and the future of the nation.

REFERENCES

[1] S. Zahroh and R. S. Pontoh, “Education as an important aspect to determine human development index by province in Indonesia,” *Journal of Physics: Conference Series*, vol. 1722, no. 1, p. 012106, 2021.

[2] D. Solichah, “Hubungan Antara Fasilitas Sekolah, Minat Belajar, Dan Kebiasaan Belajar Siswa MI Al-Huda,”Tech. Rep., 2018.

[3] Sholihah, A. K., & Mufidah, N. (2021). Pengaruh Lingkungan dan Fasilitas Belajar terhadap Prestasi Belajar Siswa pada Mata Pelajaran IPS. JIIPSI: Jurnal Ilmiah Ilmu Pengetahuan Sosial Indonesia, 1(2), 164-173.

[4] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, “A short review on different clustering techniques and their applications,” *Advances in Intelligent Systems and Computing*, pp. 69–83, 2019.

[5] P. Marrone, P. Gori, F. Asdrubali, L. Evangelisti, L. Calcagnini, and G. Grazieschi, “Energy benchmarking in educational buildings through cluster analysis of energy retrofitting,” *Energies*, vol. 11, no. 3, p. 649, 2018.

[6] Y. Irawan, “Implementation of data mining for determining majors using K-means algorithm in students of SMA Negeri 1 Pangkalan Kerinci,” *Journal of Applied Engineering and Technological Science (JAETS)*, vol. 1, no. 1, pp. 17–29, 2019.

[7] A. F. Mohamed Nafuri, N. S. Sani, N. F. Zainudin, A. H. Rahman, and M. Aliff, “Clustering analysis for classifying student academic performance in Higher Education,” *Applied Sciences*, vol. 12, no. 19, p. 9467, 2022.

[8] S. Križanić, “Educational data mining using cluster analysis and decision tree technique: A case study,” *International Journal of Engineering Business Management*, vol. 12, p. 184797902090867, 2020.

[9] P. R. V. Terlapu, B. G. Lakshmi, H. Swapna Rekha, and M. DhanaLakshmi, “Student Performance Analysis with using statistical and Cluster Studies,” *Soft Computing in Data Analytics*, pp. 743–757, 2018.

[10] Govindasamy, K.; Velmurugan, T. Analysis of Student Academic Performance Using Clustering Techniques. Int. J. Pure Appl. Math. 2018, 119, 309–323

[11] Navarro, Á.A.M.; Ger, P.M. Comparison of Clustering Algorithms for Learning Analytics with Educational Datasets. IJIMAI 2018, 5, 9–16.

[12] Sandeep Trivedi and Nikhil Patel, “Clustering Students Based on Virtual Learning Engagement, Digital Skills, and E-learning Infrastructure: Applications of K-means, DBSCAN, Hierarchical, and Affinity Propagation Clustering,” Sage Science Review of Educational Technology, vol. 3, no. 1, pp. 1–13, Jan. 2020.

[13] K. P. Sinaga and M.-S. Yang, “Unsupervised K-means clustering algorithm,” *IEEE Access*, vol. 8, pp. 80716–80727, 2020. doi:10.1109/access.2020.2988796

[14] D. M. SAPUTRA, D. SAPUTRA, and L. D. OSWARI, “Effect of distance metrics in determining K-value in k-means clustering using elbow and silhouette method,” Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019), 2020. doi:10.2991/aisr.k.200424.051

[15] A. Kristianto, “IMPLEMENTASI DBSCAN dalam clustering data Minat mahasiswa setelah pandemi covid19,” *KONSTELASI: Konvergensi Teknologi dan Sistem Informasi*, vol. 2, no. 2, 2022. doi:10.24002/konstelasi.v2i2.5638

[16] E. Muningsih, I. Maryani, and V. R. Handayani, “Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa,” EVOLUSI : Jurnal Sains dan Manajemen, vol. 9, no. 1, Mar. 2021, doi: 10.31294/evolusi.v9i1.10428.

[17] K. R. Shahapure and C. Nicholas, “Cluster quality analysis using Silhouette score,” *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020. doi:10.1109/dsaa49011.2020.00096