Jurnal Teknik Informatika C.I.T Medicom, 16 (6) (2025) 342-352

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Outlier detection in the clustired data

Efori Bu'ulolo[1], Rian Syahputra[2], Elsya Sabrina Asmita Simorangkir[3]
[1] Teknik Informatika, Universitas Budi Darma, Medan, Indonesia
[2,3] Teknik Komputer/Manajemen Informatika, Politeknik Negeri Medan, Medan, Indonesia

Article Info

Article history

Received : 28 Dec, 2024
Revised : Jan 28, 2025
Accepted : Jan 31, 2025

Kata Kunci:

Data;
Detection;
Cluster;
Outlier

Abstract

The purpose of this study is to detect outliers in data clusters. Outliers in data cluster datasets often occur in the data clustering process, especially in the K-Means algorithm. Outliers in cluster data are members/cluster items that are far from the centroid value and are not found in the dominant cluster. Outliers in cluster data are caused by various factors such as inaccurate K values, inaccurate centroid point values, poor data quality and others. To detect outliers in cluster data using the blox plot method, Z-Score and relative size factor (RSF). The input value is the sum of squared error (SSE), calculated by summing the squares of the distance of each data point from the cluster centroid. The dataset used consists of 3 (three) variances, namely high data variance, medium data variance and low data variance. The method used for outlier detection in this study can detect outliers in all data variances used, only not all outlier detection methods are optimal for all data variances. The plox plot method is optimal for high data variance and medium data variance, the RSF method is optimal for medium data variance and the Z-Score method is not optimal for high data variance.

Corresponding Author:

Efori Bu'ulolo,
Teknik Informatika ,
Politeknik Negeri Medan
Universitas Sumatera Utara, Jl. Almamater No.1, Padang Bulan, Medan City, North Sumatra 20155, Indonesia
E-mail : buuloloefori21@gmail.coml

1. Introduction

Data mining refers to the extraction, exploration, or search for information or knowledge from large datasets[1][2]. One of the purposes of data mining is to organize data into groups[3]. Data clustering is the process of grouping data based on distance or characteristics, where data within the same cluster have short distances or similar characteristics, while data between clusters have large distances or different characteristics[4]. In clustered datasets, outliers or anomalies are often found, especially in clustering datasets using the K-Means algorithm. Outliers in clustered datasets refer to cluster members that are far from the centroid value and are not part of the dominant cluster[5][6][7]. These outliers are caused by various factors, such as an inaccurate K value, imprecise centroid values, poor data quality, and errors in the calculation process[8].

One outlier point can be seen as it is far from the center and significantly separated from its data group. To determine the presence or absence of outlier points in the clustered dataset, an outlier detection method is used. In this study, the outlier detection methods applied are box plot, Z-Score, and Relative Size Factor (RSF)[9][10][11]. These three outlier detection methods are applied to detect outliers

in clustered datasets with three types of variances. Dataset variance refers to a measure that represents the extent of data deviation from its central point[12]. The three variances include high data variance, medium data variance, and low data variance.

Research related to outlier detection includes the study conducted by Yezheng Liu, Zhe Li, Chong Zhou, and colleagues in 2019, which discusses how detecting outliers can be approached as a binary classification problem. Outlier samples originate from discrete data, where all data points have equal probability. The lack of information in high-dimensional data space leads to failures in classifying potential outliers. To tackle this problem, a novel approach named Single-Objective Generative Adversarial Active Learning (SO-GAAL) was introduced to manage potential outliers effectively. SO-GAAL is capable of identifying potential outliers in an insightful way. For high-dimensional datasets, Multiple-Objective Generative Adversarial Active Learning (MO-GAAL) is recommended for outlier detection. When compared to other methods, MO-GAAL demonstrates superior performance in detecting outliers and is well-suited for managing various types of data clusters.[13].

Research conducted by Guojun Gan and Michael Kwok-Po Ng in 2017 Clustering was performed using the K-Means algorithm while addressing the issue of outliers. One notable drawback of the K-Means algorithm is its sensitivity to fluctuations and a high presence of outliers. To overcome this limitation, the KMOR algorithm (K-Means with Outlier Removal) was introduced, improving K-Means for outlier detection. The KMOR algorithm incorporates an additional cluster, represented by the formula K+1, specifically to handle outliers. The results indicate that KMOR can effectively perform clustering and outlier detection simultaneously, while also offering higher accuracy and faster processing compared to other algorithms[14]. Research by K. Senthamarai Kannan and K. Manoj in 2015 focused on outlier detection in multivariate data. This approach uses distance-based outlier detection with robust regression diagnostic techniques. Sometimes, a single outlier can hide other outliers, necessitating a method to identify outliers concealed by others. Multivariate outlier detection uses a multiple-outlier detection procedure with a multivariate linear regression model[15].

The aim is not only to detect outliers in the clustered dataset but also to identify the most optimal outlier detection method for different data variances, as not all outlier detection methods perform optimally across various types of data variances. The input value utilized for the outlier detection methods is the Sum of Squared Error (SSE), which is calculated by summing the squared distances between each data point and its respective cluster centroid [16]. The SSE value is also employed in the elbow method to determine the optimal number of clusters for a given dataset [17].

Research references mentioned, there has not been any study related to outlier detection in clustered data using multiple outlier detection methods with input values derived from the Sum of Square Error (SSE) based on data variance. Thus, the novelty of this research lies in the use of SSE values as input for various outlier detection methods and identifying the optimal outlier detection method for specific data variances.

2. Research Methodology

The stages carried out in this research, from the beginning to the completion of the study, are as follows:

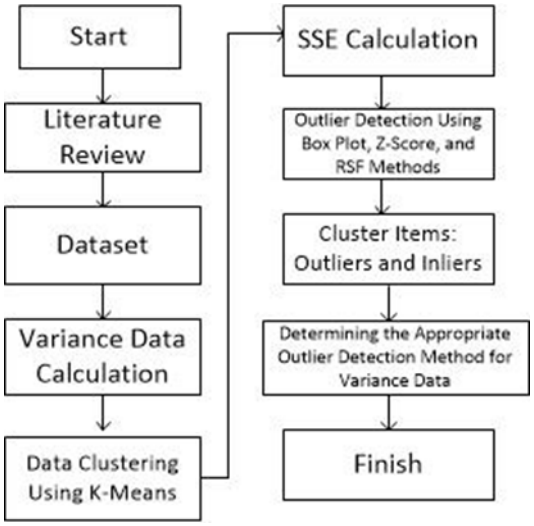


Figure. 1. Research Flowchart

1. Literature Study
   The first step undertaken by the author is reading and understanding theories related to the research topic, with the aim of expanding the author's insights and knowledge. The author has studied and comprehended theories related to outliers and their detection techniques, the K-Means algorithm, and the elbow method, drawing from various journals, proceedings, and books.
2. Dataset
   The dataset serves as a sample of data utilized in the outlier detection process. In this research, the dataset includes three types of data variances, arranged randomly, with three variables: X, Y, and Z values. The data variances included are high variance, medium variance, and low variance[18].
3. Calculation of Data Variance
   The arranged dataset is analyzed to calculate the variance values to determine which data has high variance, medium variance, and low variance. The formula for calculating data variance is[19]:
   $$\sigma^2 = \frac{1}{n}\sum_{n=1}^{n}(x_i - \bar{x})^2 \quad (1)$$
4. Data Clustering with K-Means
   Next, the existing dataset, which consists of three dataset variances, undergoes the clustering process. In this study, the value of K=3 is used for all datasets. Then, the centroid values are determined randomly, and the distance calculation is performed using the following formula[20]:
   $$d_{ij} = \sqrt{\sum_{i=1}^{n}(x_i - y_{i)}{}^2} \quad (2)$$
5. Sum of Square Error (SSE)
   The clustered data, derived from the three dataset variances, is processed to calculate the Sum of Square Error (SSE), the formula for calculating SSE is [21]:
   $$SSE = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (3)$$
6. Outlier Detection with Box Plot, Z-Score, and RSF
   To detect outliers from the formed clusters, the previously calculated SSE is used as input for the outlier detection methods: Box Plot, Z-Score, and RSF.
   a. Box Plot is a method of detecting outliers based on data distribution, measures of central tendency, and measures of dispersion. The definition of outlier detection using Box Plot is as follows[22]:

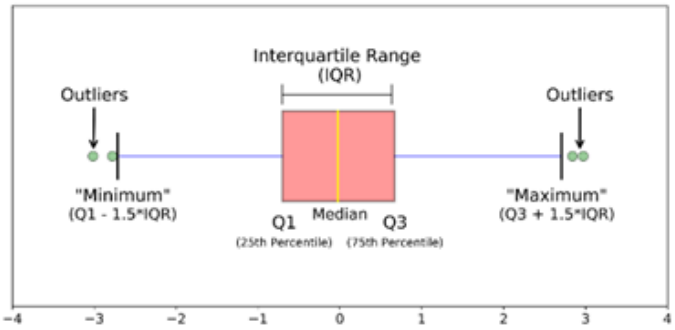


Figure. 2. Outlier Detection Using Box Plot[23]

b. Z-Score is an outlier detection method that calculates the difference between a data value and the mean of the data, divided by the standard deviation. The upper and lower thresholds are set at +3 and -3, respectively. Any value beyond these thresholds is regarded as an outlier. The formula for calculating the Z-Score is as follows[24]:

$$Z = \frac{x-\bar{x}}{s} \quad (4)$$

c. Relative Size Factor (RSF) is an outlier detection method that assesses the significance of differences within a data sequence. The significance value is determined by comparing the highest value in a subset with the second-highest value. The formula for the RSF method is as follows[25]:

$$RSF = \frac{The\ Highest\ Value\ in\ a\ Category}{The\ Second\ Highest\ Value\ in\ a\ Category} \quad (5)$$

To determine the upper and lower bounds for datasets with ≤11 records, the following formula is used:

$$\frac{n-1}{\sqrt{n}}, \quad (6)$$

For datasets with > 11 records, the upper and lower bounds are set at +3 and -3, respectively. These thresholds in the RSF method are used to determine whether a data cluster member is classified as an outlier or an inlier.

7. Outlier and Inlier Cluster Items
   The next step is to determine whether a cluster member or data item is an outlier or an inlier basedon the upper and lower threshold values defined by each outlier detection method.
8. Determining the Outlier Detection Method Suitable for Data Variance
   The final stage is determining the outlier detection method that is most suitable for each data variance. Not all outlier detection methods are optimal for all types of data variance. Thus, an analysis is performed to determine the most suitable outlier detection method for different data variances. In this study, the data variances include three types: high variance, medium variance, and low variance.

3. Result and Disscusion

The dataset used in this study includes three types of data variance: high variance, medium variance, and low variance, along with three variables: X, Y, and Z.

Tabel 1. High Variance Data

No	Subjek	X	Y	Z
1	MM	10	15	20
2	EE	24	34	45
3	EK	60	75	76
⋮				
20	NT	40	44	38

21	WT	85	78	53
22	CC	95	70	76
Varians Data				683,875

Tabel 2. Medium Variance Data

No	Subjek	X	Y	Z
1	ML	75	42	35
2	EC	70	68	25
3	EO	40	47	85
⋮				
20	NA	51	47	75
21	WY	85	78	53
22	CI	95	70	76
Varians Data				304,15

Tabel 3.
Low Variance Data

No	Subjek	X	Y	Z
1	MR	75	72	76
2	CC	70	78	87
3	EL	92	97	85
⋮				
20	NL	81	87	75
21	WY	85	98	93
22	IC	95	90	96
Varians Data				79,378

The variance values of the data, consisting of high variance, medium variance, and low variance, are presented in graphical form in Figure 4 below. The difference between the data variances lies in their variance values, where high variance data has a variance value of 683.875, medium variance data has a value of 304.15, and low variance data has a value of 79.378.

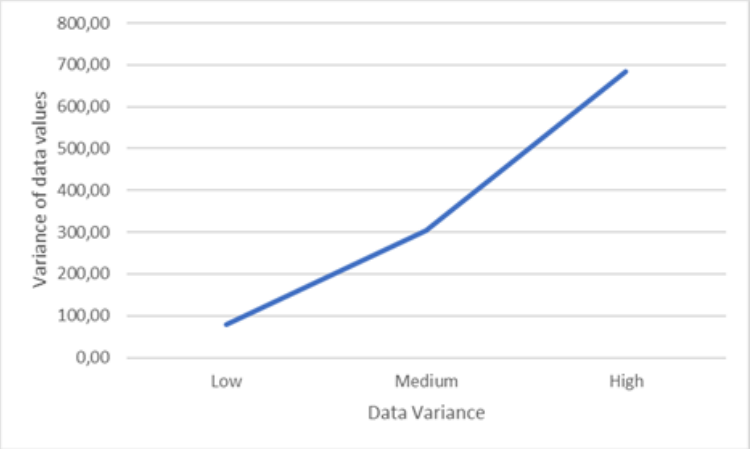


Figure. 3. Data Variance Graph

The three data variances above were clustered using the K-Means algorithm in Python with K=3, and the results are shown in Figures 5, 6, and 7. The slightly larger red dots represent the centroids. Additionally, each figure contains points in three different colors, which represent the data clusters for each dataset with K=3.

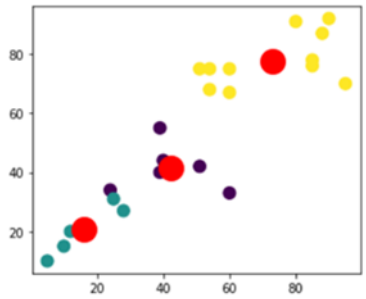


Figure. 4. Cluster from Table 1

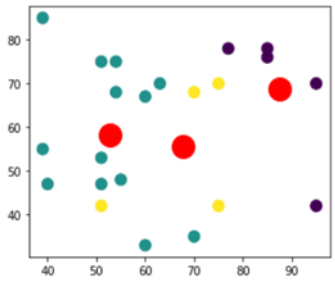


Figure. 5. Cluster from Table 1

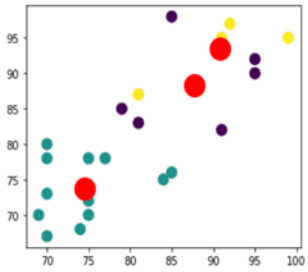


Figure. 6. Cluster from Table 1

After clustering the dataset, the next steps involve calculating the SSE (Sum of Squared Errors) and identifying outliers for each data point within the clusters of each dataset.

3.1 Outlier Detection for High Variance Data (Table 1)

Figure 7 illustrates the results of outlier detection using the box plot method for high variance data (Table 1) based on SSE values. The orange line indicates the upper inner fence, the gray line represents the lower inner fence, and the blue line shows the SSE values. Each cluster has distinct inner fence values:

Cluster I: Q3 + 1.5 * IQR = 488.8, Q1 - 1.5 * IQR = -183.2
Cluster II: Q3 + 1.5 * IQR = 841.708, Q1 - 1.5 * IQR = -75.958
Cluster III: Q3 + 1.5 * IQR = 521.851, Q1 - 1.5 * IQR = -97.126

For Cluster I and Cluster II, there are no outliers as all blue lines (SSE values) remain within the orange and gray lines. However, for Cluster III, there are three points that exceed the blue line, indicating three outliers in Cluster III: points TG, WT, and CC.

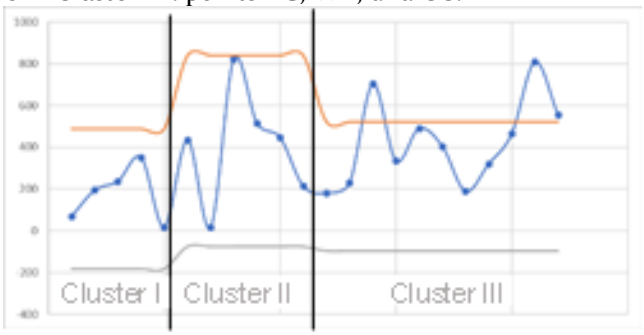


Figure 7. Outlier Detection Using Box Plot on Variance Data

Figure 8 demonstrates outlier detection using the Z-Score method. In this figure, no cluster members are flagged as outliers, as the blue line (SSE) remains within the upper and lower threshold values. The upper and lower threshold values for each cluster are as follows: Cluster I = 1,789 / -1,789, Cluster II = 2,041 / -2,041, Cluster III = 3 / -3.

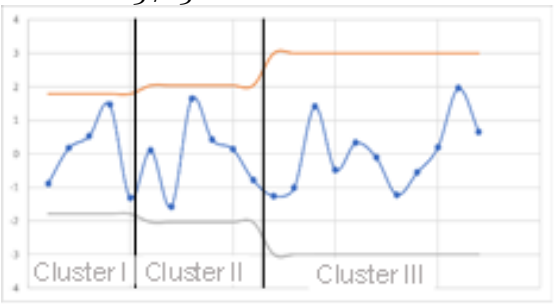


Figure 8. Outlier Detection with the Z-Score Technique

Figure 9 shows outlier detection using the Relative Size Factor (RSF) method. In this figure, only one point is significantly distant or not densely packed with the dominant points. Therefore, there is only one outlier, which is the point NT.

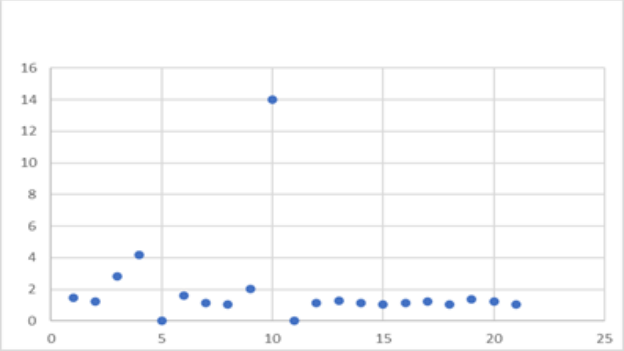


Figure 9. Outlier Detection Using the RSF Method

3.2 Medium Variance Data (Table 2)

Figure 10 presents the results of outlier detection using the box plot method for medium variance data (Table 2) based on SSE values. The orange line denotes the upper inner fence, the gray line indicates the lower inner fence, and the blue line represents the SSE values. Each cluster has different inner fence values.:

Cluster I: Q3 + 1.5 * IQR = 429.875, Q1 - 1.5 * IQR = 120.875

Cluster II: Q3 + 1.5 * IQR = 1091.095, Q1 - 1.5 * IQR = -388.904

Cluster III: Q3 + 1.5 * IQR = 964.974, Q1 - 1.5 * IQR = -281.596

In Cluster I, there is one outlier, which is point PN. In Cluster II, there are two outliers, points CA and AG, as the blue line (SSE) exceeds the orange line. In Cluster III, there are no outliers, meaning all items are inliers.

Figure 10. Outlier Detection Using the Box Plot Method

Figure 11 displays the results of outlier detection using the Z-Score method. In this figure, one outlier is identified in Cluster I, specifically point PN, since the blue line (SSE) surpasses the orange line, which indicates the upper inner fence. The upper and lower threshold values for each cluster using the Z-Score method are as follows:

Cluster I: 1.5/-1.5 , Cluster II: 3/-3 , Cluster III:2.267/-2.267

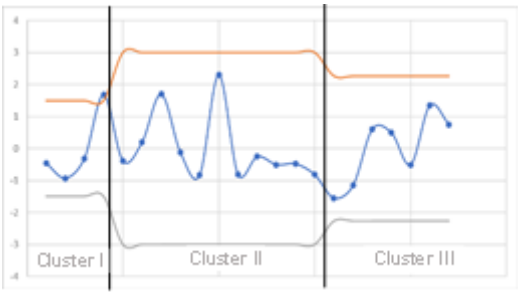


Figure 11. Identifying Outliers with the Z-Score Method

Figure 12 shows the results of outlier detection using the Relative Size Factor (RSF) method. In this figure, there are three points that are significantly distant or not densely packed with the dominant points. These three outliers are CA, HA, and DY.

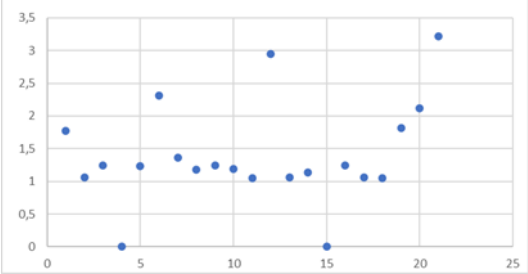


Figure 12. Outlier Detection Using the RSF Method

3.3 Low Variance Data (Table 3)

Figure 13 illustrates the results of outlier detection using the box plot method for low variance data (Table 3) based on SSE values. The orange line indicates the upper inner fence, the gray line represents the lower inner fence, and the blue line shows the SSE values. Each cluster has distinct inner fence values:

Cluster I: Q3 + 1.5 * IQR = 149.593, Q1 - 1.5 * IQR = -51.906
Cluster II: Q3 + 1.5 * IQR = 120.551, Q1 - 1.5 * IQR = 21.693
Cluster III: Q3 + 1.5 * IQR = 280.168, Q1 - 1.5 * IQR = -76.688

In Cluster I, there are no outliers, meaning all points are inliers as the blue line (SSE) does not exceed the orange or gray lines. In Cluster II, there is one outlier, point NL, and in Cluster III, there is one outlier, point NS, because the blue line exceeds the orange line.

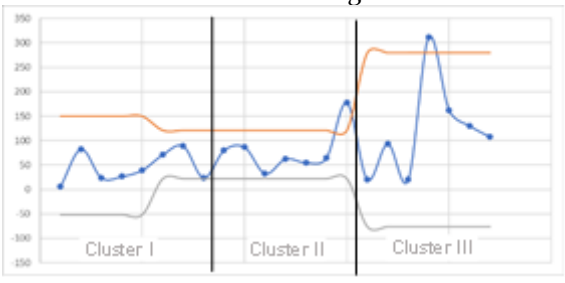
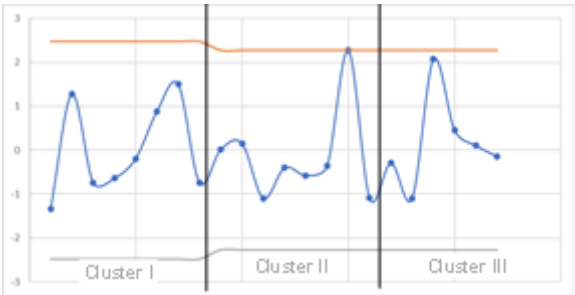


Figure 13. Outlier Detection Using the Box Plot Method

Figure 14 presents the results of outlier detection using the Z-Score method. In this figure, one outlier is identified in Cluster II, specifically point NL, as the blue line (SSE) surpasses the orange line, which represents the upper inner fence. The upper and lower threshold values for each cluster using the Z-Score method are as follows:

Cluster I: 2.474/–2.474, Cluster II: 2.267/–2.267, Cluster III: 2.267/–2.2672.267 / -2.267



Gambar 14. Deteksi Outlier dengan Metode Z-Score

Figure 15 displays the results of outlier detection using the Relative Size Factor (RSF) method. In this figure, two points, VI and BB, are identified as outliers due to their significant distance or sparse proximity to the dominant points.

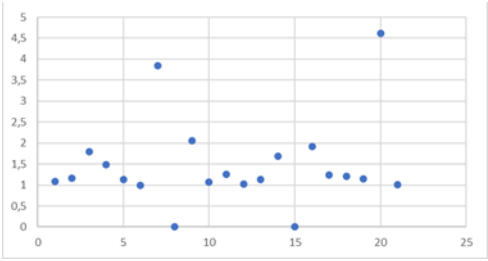


Figure 15. Outlier Detection Using the RSF Method

3.4 Comparison of Outlier Detection Results

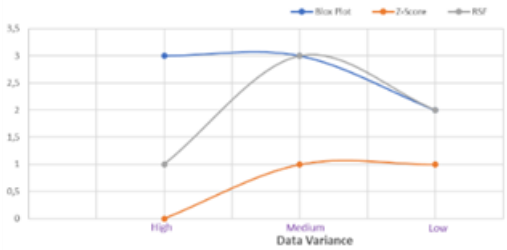


Figure 16. Comparison of Results from Outlier Detection Methods

After obtaining the outlier detection results using the box plot, Z-Score, and RSF methods for all three types of variance—high, medium, and low—a comparison is made based on the number of outlier points detected. In Figure 16, the blue line shows the number of outlier points identified by the box plot method, with outliers detected in all data variances. The gray line represents the number of outliers detected by the RSF method, where outliers are found in all variance types, though the number is smaller compared to the box plot method. The orange line illustrates the number of outliers identified by the Z-Score method, with outliers found only in the medium and low variance data, not in the high variance data.

4. Conclusion

Outliers in clustered data can be detected using the box plot, Z-Score, and RSF methods for all three types of variance—high, medium, and low—using the SSE value as input. The SSE value, which indicates the distance of a data point from its cluster centroid, is the basis for calculating the outlier detection method. Not all outlier detection methods are suitable for every type of variance. The box plot method works well for high and medium variance data, the RSF method is most effective for medium variance data but less effective for high variance data, while the Z-Score method is less effective for high and medium variance data. Outliers in clustered data are partly caused by suboptimal K selection and

inaccurate centroid values. As a result, a new model for determining K and centroids in data clustering is needed, which provides a basis for future research directions. The limitations of this study are the dependence on the K-Medoids algorithm and SSE (Sum of Squared Error), as well as the lack of parameter sensitivity analysis. For further research, namely the development of a more robust clustering method and parameter sensitivity analysis.

References

[1] Buulolo and Efori, *Data Mining Untuk Perguruan Tinggi*. Yogyakarta: deepublish, 2020. [Online]. Available: https://www.google.co.id/books/edition/Data_Mining_Untuk_Perguruan_Tinggi/-K_SDwAAQBAJ?hl=id&gbpv=1&dq=Data+Mining+Konsep+dan+Aplikasi+Menggunakan+Matlab&printsec=frontcover

[2] M. Kantardzic, *Data Mining*. 2011. doi: 10.1002/9781118029145.

[3] R. Muliono and Z. Sembiring, "Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen," *J. Comput. Eng. Syst. Sci.*, vol. 4, no. 2, pp. 2502–714, 2019.

[4] E. Bu'ulolo and B. Purba, "Algoritma Clustering Untuk Membentuk Cluster Zona Penyebaran Covid-19," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 12, no. 1, pp. 59–67, 2021, doi: 10.31849/digitalzone.v12i1.6572.

[5] P. A. Ariawan, "Optimasi Pengelompokan Data Pada Metode K-means dengan Analisis Outlier," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 2, pp. 88–95, 2019, doi: 10.25077/teknosi.v5i2.2019.88-95.

[6] E. Wahyuni and S. Suparman, "A Comparison of Outlier Detection Techniques in Data Mining," in *Science, Technology, Engineering, Economics, Education, and Mathematics*, 2019, vol. 1, no. 1, pp. 139–147.

[7] B. Tang and H. He, "A local density-based approach for outlier detection," *Neurocomputing*, vol. 241, pp. 171–180, 2017, doi: 10.1016/j.neucom.2017.02.039.

[8] E. T. K. Dewi, A. Agoestanto, and Sunarmi, "Metode Least Trimmed Square (Lts) Dan Mm-Estimation Untuk Mengestimasi Parameter Regresi Ketika Terdapat Outlier," *J. Math.*, vol. 5, no. 1, pp. 47–54, 2016, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/ujm/article/view/13104

[9] Ebrary.net, "RELATIVE SIZE FACTOR TEST," *ebrary.net*.

[10] E. Sabrina, A. Simorangkir, A. Putera, U. Siahaan, L. Marlina, and D. Nasution, "Deteksi Outlier Hasil Clustering Algoritma K-Medoids Menggunakan Metode Boxplot Pada Data KIP Kuliah," *J. Comput. Syst. Informatics*, vol. 5, no. 4, pp. 893–902, 2024, doi: 10.47065/josyc.v5i4.5479.

[11] A. S. Yaro, F. Maly, P. Prazak, and K. Maly, "Outlier Detection Performance of a Modified Z-Score Method in Time-Series RSS Observation With Hybrid Scale Estimators," *IEEE Access*, vol. 12, no. January, pp. 12785–12796, 2024, doi: 10.1109/ACCESS.2024.3356731.

[12] L. Ruwah Ibnatur Husnul, Nisak; Rima Prasetya, Eka;Sadewa, Prima;Ajimat; Ike Purnomo, *STATISTIK DESKRIPTIF*, no. 1. Pamulang: UNPAM PRESS, 2020. doi: 10.1007/978-3-662-48986-4_2900.

[13] Y. Liu *et al.*, "Generative Adversarial Active Learning for Unsupervised Outlier Detection," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1517–1528, 2020, doi: 10.1109/TKDE.2019.2905606.

[14] G. Gan and M. K. P. Ng, "K-Means Clustering With Outlier Removal," *Pattern Recognit. Lett.*, vol. 90, pp. 8–14, 2017, doi: 10.1016/j.patrec.2017.03.008.

[15] K. Senthamarai Kannan and K. Manoj, "Outlier detection in multivariate data," *Appl. Math. Sci.*, vol. 9, no. 45–48, pp. 2317–2324, 2015, doi: 10.12988/ams.2015.53213.

[16] N. T. Hartanti, "Jurnal Nasional Teknologi dan Sistem Informasi Metode Elbow dan K-Means Guna Mengukur Kesiapan Siswa SMK Dalam Ujian Nasional," vol. 02, pp. 82–89, 2020.

[17] A. Winarta and W. J. Kurniawan, "Optimasi Cluster K-Means Menggunakan Metode Elbow Pada Data Pengguna Narkoba Dengan Pemrograman Python," *JTIK (Jurnal Tek. Inform. Kaputama)*, vol. 5, no. 1, pp. 113–119, 2021, doi: 10.59697/jtik.v5i1.593.

[18] S. Bu'ulolo;Efori, Mesran, Hasibuan;Nelly Astuti, Utomo;Aripin;Soeb, Putro Utomo, *Big Data Analysis dengan Phyton untuk Perguruan Tinggi*, I. Yogyakarta, 2023.

[19] M. Bachmaier, "The striking criterion whether variance calculation requires dividing the sum of squares by the number of summands or by that number less one by," *Int. J. Educ. Res.*, vol. 1, no. 6, pp. 1–14, 2013.

[20] 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.

[21] R. Al Muiz, "Comparison of K-Means and Fuzzy C-Means for Optimizing Tuberculosis Management and Healthcare Service Allocation in Bojonegoro," *J. Stat. Dan Komputasi*, vol. 3, no. 2, pp. 80–91, 2024.

[22] N. Rokhman, H. Nugroho, and D. Saputri, "Peningkatan Efisiensi Penugasan Guru di Provinsi Daerah

Istimewa Yogyakarta Melalui Penghapusan Outlier," in *Konferensi Nasional Sistem Informasi 2018*, 2018, pp. 8–9.

[23] V. Agarwal, "Outlier detection with Boxplots," *medium.com*, 2019. https://medium.com/@agarwal.vishal819/outlier-detection-with-boxplots-1b6757fafa21

[24] M. KENAMON, Y. D. WINAWUNG, and H. HANINUN, "Prediksi Kebangkrutan Dengan Model Altman Z-Score Pada Perusahaan Farmasi Yang Terdaftar Di Bursa Efek Indonesia Periode 2012-2016," *J. Akunt. dan Keuang.*, vol. 9, no. 1, p. 10, 2018, doi: 10.36448/jak.v9i1.999.

[25] A. Nugroho, "Data Interrogation & Analysis: Teknik Mendeteksi Outlier - Metode Relative Size Factor," *Badan Pendidikan dan Pelatihan Keuangan*, 2020. https://klc2.kemenkeu.go.id/kms/knowledge/klc1-pknstan-data-interrogation-analysis-teknik-mendeteksi-outlier-metode-relative-size-factor/detail/