Jurnal Sisfotek Global
ISSN (Online): 2721 - 3161, ISSN (Print): 2088 – 1762
DOI: http://dx.doi.org/10.38101/sisfotek.v14i1.10860
Vol. 14, No. 1, March 2024, pp. 49-56

Implementation of K-means Clustering Algorithm for the Indonesian Stock Exchange

Bakti Siregar[1] & Yosia[2]

[1,2] Matana University, Tangerang, Indonesia, 15810
E-mail: [1]siregar.bakti@matanauniversity.ac.id, [2]yosia.yosia@student.matanauniversity.ac.id

ARTICLE HISTORY

Received : February 1st, 2024
Revised : February 27th, 2024
Accepted : March 26th, 2024

KEYWORDS

Stock Market
Machine Learning
K-Means Clustering

ABSTRACT

In the dynamic field of financial markets, effective analysis and understanding of stock market behavior are very crucial for investors, analysts, and policymakers. This study investigates the implementation of the K-means algorithm for clustering stocks listed on the Indonesian Stock Exchange (IDX). The main objectives of this research include exploring IDX's clustering patterns, identifying groups based on their trading characteristics, and evaluating algorithm performance. Some challenging parts have been addressed, such as data quality, feature selection, determining the optimal number of clusters, scalability, interpretability, and evaluation. Precise data preprocessing, feature engineering, and algorithm optimization provide insight into the clustering structure of the Indonesian stock market, helping investors in portfolio diversification, risk management, and strategic decision-making. The results show the potential of the K-means algorithm in thoroughly uncovering important patterns on the IDX, thereby contributing to the advancement of market analysis methodologies adapted to the Indonesian financial environment.

1. Introduction

The Indonesian Stock Exchange (BEI) presents a challenging environment characterized by a large amount of heterogeneous data, including stock prices, trading volumes and company fundamentals [1]. Navigating this complexity to gain actionable insights is a major challenge for investors and analysts [2]. Navigating this complexity to gain actionable insights is a major challenge for investors and analysts [3]. However, manually identifying these segments is laborious and prone to subjectivity.

Although clustering algorithms such as K-means have been widely used in other financial markets, this research fills a gap in the literature by exploring the effectiveness of clustering techniques specifically adapted to BEI [4]. Researchers use advanced data preprocessing techniques to handle missing values, normalize data, and select relevant features for clustering[5]. This ensures that the input data fed into the K-means algorithm is clean, standardized, and conducive to meaningful clustering [6]. To maximize the effectiveness of the K-means algorithm, this research explores various parameter settings and optimization strategies [7]. This includes experimenting with different numbers of clusters, initialization methods, and convergence criteria to identify configurations that produce the most meaningful clustering results [8]. Apart from simply classifying shares, this research also focuses on interpreting the resulting clusters in the context of the Indonesian market [9]. By analyzing the characteristics and dynamics of each cluster, this research aims to provide actionable insights for investors, such as identifying emerging sectors, detecting market anomalies, or predicting future market trends [10].

2. Literature Review

2.1 Stock Market

The data stock market tends to be volatile, unpredictable, and non-linear[11]. Therefore, to build a stock price prediction model depends on various factors including the pandemic, political conditions, the global economy, financial reports, company performance, etc[12]. Other than that, in the process of forming an optimal stock investment portfolio, a stock value prediction model is needed to estimate returns and volatility to analyze trends over the last few years[13].

2.2 Machine Learning

Machine learning is a collection of mathematical and statistical algorithms inserted into a computer system, which are adopted from learning data so that they can be used to produce predictions in the future [14]. The learning process was validated through two stages, including training and testing [15]. Various literature reveals that machine learning is divided into three categories: Supervised Learning, Unsupervised Learning, and Reinforcement Learning [16].

Supervised Learning is a classification method that uses dependent variables as initial labels to classify unknown classes. Meanwhile, the Unsupervised Learning technique is often called a cluster formation or search model because there are no initial labels that can be used to identify groups [17]. Simultaneously, Reinforcement Learning is a technique that combines Supervised Learning and Unsupervised Learning [18], usually used on dynamic data to complete classification without explicit initial labels [19].

Several real-world machine learning cases require time and money to label a wide variety of possibilities. Meantime, data that does not yet have a label can be obtained easily and even for free. One of the unsupervised learning models that will be used in the financial sector is popular algorithms such as k-means [18], [20].

2.3 K-means algorithm

The K-means algorithm can be used if the number of clusters that correspond to all the features used in the stock grouping process has been searched. According to [21], comparing 3 algorithms such as: The Elbow, Silhouette, and Statistical Gap Methods are highly recommended steps for determining the correct number of clusters. The main reference in this research issuing more variables, it was found that Jakarta and East Java were the clusters with the highest level of severity [22]. An interesting thing was also discovered by [23]who proved that the use of the K-means algorithm can be used in larger and more practical quantities of data without having to pay attention to the order of objects.

3. Research Methods

In the context of the Indonesian Stock Exchange (IDX), the motivation for using the K-means algorithm lies in its potential to provide valuable insights and solutions to various challenges and objectives pertinent to stock market analysis [24]. Whether for market segmentation, portfolio optimization, risk management, anomaly detection, sectoral analysis, predictive modeling, or quantitative research, K-means clustering offers a powerful tool to improve decision making and understand market dynamics on IDX [25].

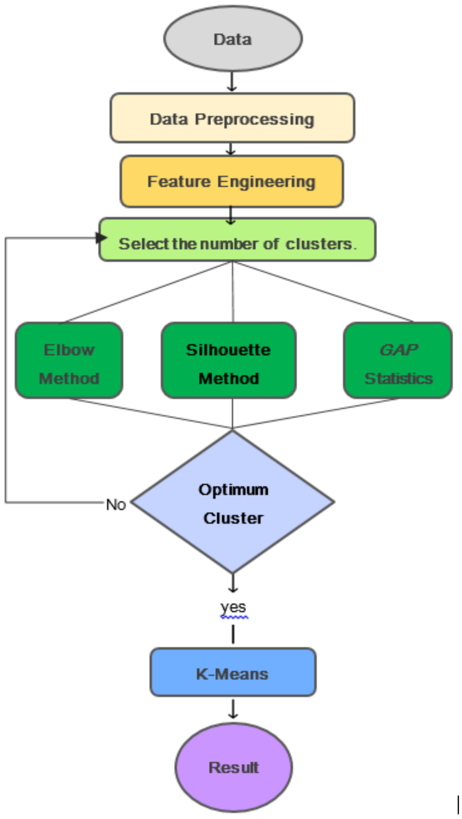


Figure 1. Research Stages

The stages of this research are shown in Figure 1. Researchers collected historical data from around the last five to six years on stock data listed on the IDX, which can be accessed on the site https://finance.yahoo.com/. Some of the features in this dataset include date, open price, highest price, lowest price, closing price, trading volume, and adjusted price. The open and close features represent the starting and ending prices at each stock is traded on a particular day. The high and low features represent the maximum and minimum share prices for that day. Trading volume is the total number of stocks or contracts traded for a particular security. The Adjusted Price represents the trading price of ordinary stocks on a national stock exchange, the market price of ordinary stocks on the applicable date (calculated based on the average closing price during the previous 20-day trading period).

In this research, the extraction process will also be carried out for several features needed in the process of implementing the K-means Clustering algorithm, such as volatility, liquidity, and capacity, etc. In general, this clustering process has 5 stages, namely:

a. Collecting data.

b. Extracting features/information from data.

c. Search for anomalous data and delete unnecessary data.

d. Grouping data based on its characteristics.

e. Interpretation of the characteristics of each cluster that has been formed.

4. Results

In the introductory section, several approaches used in forming clustering were described, especially partitional methods, density methods, hierarchical methods, etc. In this research, specifically use and compare two algorithms derived from partitional unsupervised machine learning methods to group stock data based on similar characteristics.

4.1 Data Collection

The data used in this panel comes from two sources, namely company profile data and daily share price movement data. Company profile data can be obtained from the official website of the Indonesian Stock Exchange (IDX) (https://www.idx.co.id/data-pasar/data-saham/register-saham/). There are around 858 stocks as of December 14, 2023, that have been listed on the Indonesian stock exchange.

4.2 Extraction Features

The feature extraction process is the main thing that is important to do to get the best results from the machine learning model used. In this case, 3 necessary features are extracted, volatility, liquidity, and capacity of each share to be collected. Volatility is an indicator of the level of change in share prices every day which is revealed. So, when stock prices rise or fall with a volatility coefficient percentage greater than 70 announce to be continuous, these conditions difficult to make price forecasts because it carries high risk [26]. Liquidity is an indicator that can be used to measure how easy it is for stocks to be sold and bought without affecting asset prices[27]. In other words, the liquidity of a stock would be considered from how large the volume. The parameter used to measure this liquidity is using the standard deviation value which must be lower than the average value of stock liquidity. The market capitalization value in Figure 3 is the multiplication of the total stocks and the share price[28]. When the market capitalization value gets bigger, it becomes more difficult for naughty investors to manipulate share prices.

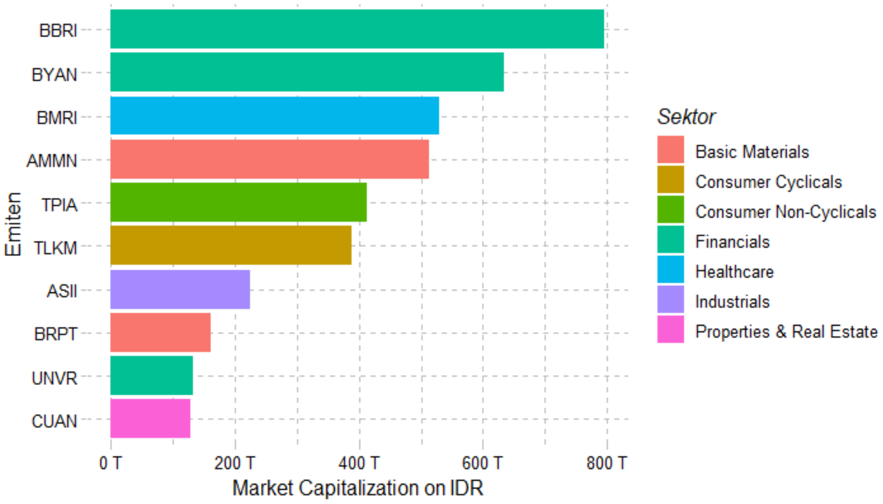
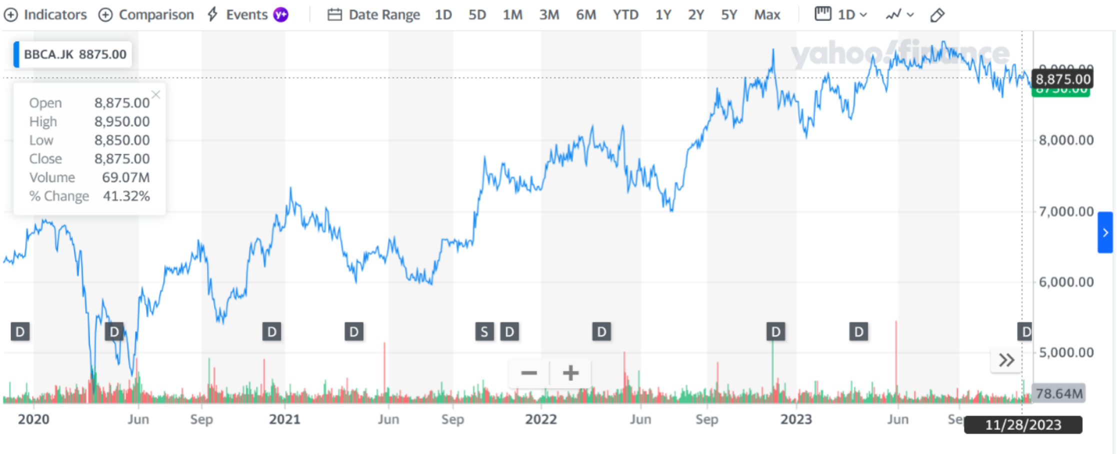


Figure 2. Volatiliy of Stocks

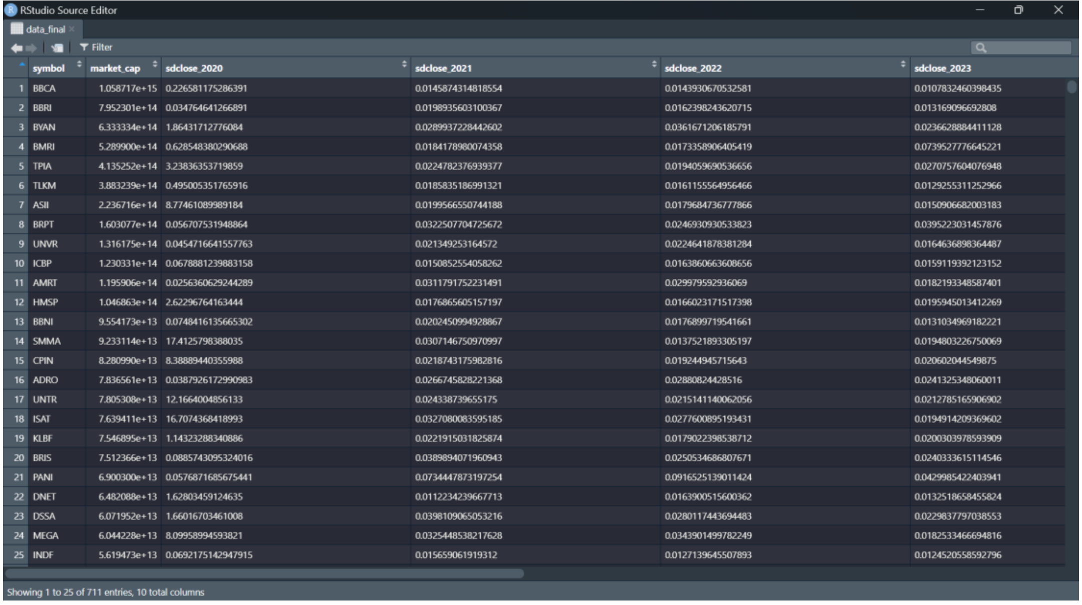


Open 8,875.00
High 8,950.00
Low 8,850.00
Close 8,875.00
Volume 69.07M
% Change 41.32%

Figure 3. Top 10 Market Capitalization

4.3 Data Frames

After data on the volatility, liquidity, and capacity of a stock is collected, then all these features are combined into the data frame shown in Figure 4.



	symbol	market_cap	sdclose_2020	sdclose_2021	sdclose_2022	sdclose_2023
1	BBCA	1.058717e+15	0.226581175286391	0.0145874314818554	0.0143930670532581	0.0107832460398435
2	BBRI	7.952301e+14	0.034764641266891	0.0198935603100367	0.0162398243620715	0.013169096692808
3	BYAN	6.333334e+14	1.86431712776084	0.0289937228442602	0.0361671206185791	0.0236628884411128
4	BMRI	5.289900e+14	0.628548380290688	0.0184178980074358	0.0173358906405419	0.0739527776645221
5	TPIA	4.135252e+14	3.23836353719859	0.0224782376939377	0.0194059690536656	0.0270757604076948
6	TLKM	3.883239e+14	0.495005351765916	0.0185835186991321	0.0161155564956466	0.0129255311252966
7	ASII	2.236716e+14	8.77461089989184	0.0199566550744188	0.0179684736777866	0.0150906682003183
8	BRPT	1.603077e+14	0.056707531948864	0.0322507704725672	0.0246930930533823	0.0395223031457876
9	UNVR	1.316175e+14	0.0454716641557763	0.021349253164572	0.0224641878381284	0.0164636898364487
10	ICBP	1.230331e+14	0.0678881239883158	0.0150852554058262	0.0163860663608656	0.0159119392123152
11	AMRT	1.195906e+14	0.0256360629244289	0.0311791752231491	0.029979592936069	0.0182193348587401
12	HMSP	1.046863e+14	2.62296764163444	0.0176865605157197	0.0166023171517398	0.0195945013412269
13	BBNI	9.554173e+13	0.0748416135665302	0.0202450994928867	0.0176899719541661	0.0131034969182221
14	SMMA	9.233114e+13	17.4125798388035	0.0307146750970997	0.0137521893305197	0.0194803226750069
15	CPIN	8.280990e+13	8.38889440355988	0.0218743175982816	0.019244945715643	0.020602044549875
16	ADRO	7.836561e+13	0.0387926172990983	0.0266745828221368	0.02880824428516	0.0241325348060011
17	UNTR	7.805308e+13	12.1664004856133	0.024338739655175	0.0215141140062056	0.0212785165906902
18	ISAT	7.639411e+13	16.7074368418993	0.0327080083595185	0.0277600895193431	0.0194914209369602
19	KLBF	7.546895e+13	1.14323288340886	0.0221915031825874	0.0179022398538712	0.0200303978593909
20	BRIS	7.512366e+13	0.0885743095324016	0.0389894071960943	0.0250534686807671	0.0240333615114546
21	PANI	6.900300e+13	0.0576871685675441	0.0734447873197254	0.0916525139011424	0.0429985422403941
22	DNET	6.482088e+13	1.62803459124635	0.0112234239667713	0.0163900515600362	0.0132518658455824
23	DSSA	6.071952e+13	1.66016703461008	0.0398109065053216	0.0280117443694483	0.0229837797038553
24	MEGA	6.044228e+13	8.09958994593821	0.0325448538217628	0.0343901499782249	0.0182533466694816
25	INDF	5.619473e+13	0.0692175142947915	0.015659061919312	0.0127139645507893	0.0124520558592796

Showing 1 to 25 of 711 entries, 10 total columns

Figure 4. Dataframe for Clustering

4.4 Checking for Outliers

Many journals discuss the process of checking outlier or anomalous data. one of them is using the DBSCAN algorithm. This algorithm is a method that uses a density approach between data (density method), the results are shown in Figure 5. From the search results, an EPS value of 2.5 was obtained, where the optimal value in the clustering process had reached minPts 8. So, from the results of examining outlier data that had been carried out in this research using the DBSCAN method, 10 outlier data were obtained. In this case, issuers that have outlier extreme values are shown in a biplot using Principal Component Analysis (PCA) in Figure 6. Outliers can be indicated as the red vertex. The outlier issuers are spread quite far from the black data distribution. So, in this research, issuers indicated as outliers were not used so that the clustering formation process would be representative. Other than that, a data scaling process is carried out from distance calculation between data using the K-Means method has the same range between data. The scaling process is carried out using the z-score method, where the scaling process is to change the scale of the data without changing the initial distribution.

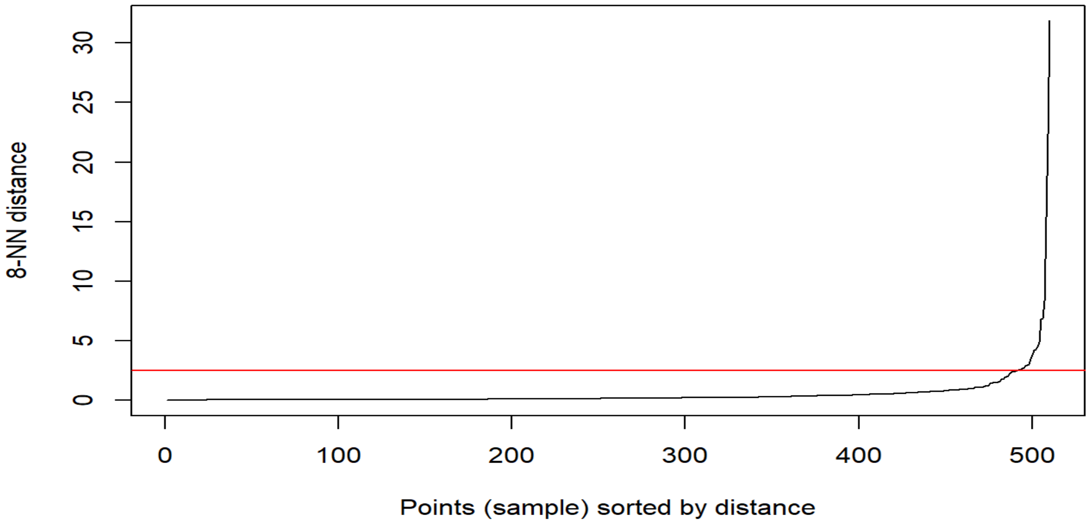


Figure 5. Examination of Outlier Data using the Density Method

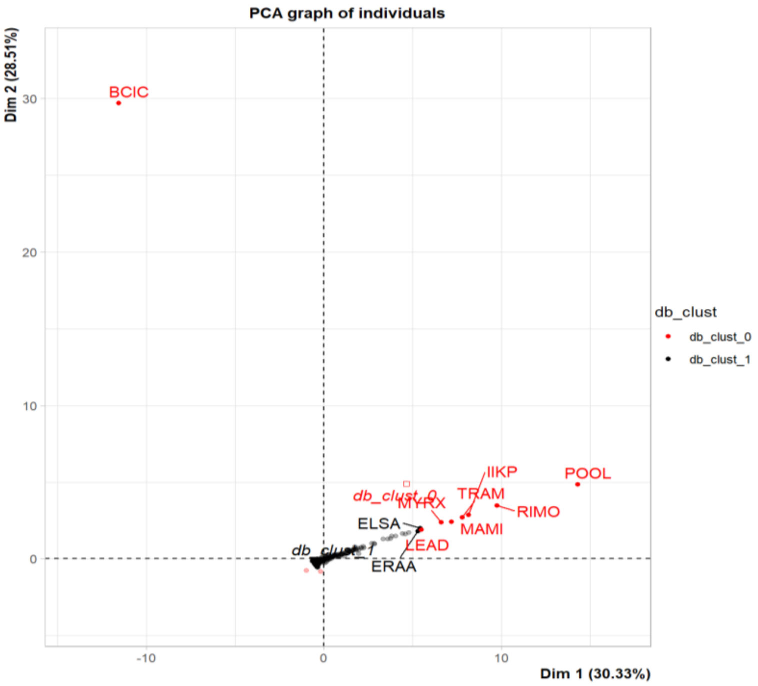


Figure 6. Principal Component Analysis

4.5 Implementation of the K-Means Algorithm

The application of the K-Means algorithm will produce a cluster center called a centroid. The centroid is a point that can represent the average value of each cluster-forming variable. In determining the optimum k value, the Elbow Method is used because it can optimize the distance between data and the centroid. Referring to the journal[29], the k value is said to be optimal when the number of clusters added continuously but the value of Within Sum of Squares (WWS) no longer changes drastically, as shown in figure 7. Assume to apply 7 clusters then increase become 7 clusters, the result does not significant decrease in the total centroid value. In other words, decrease in the centroid value has converged (no further changes), so it can be concluded that the optimal number of clusters selected is 7 clusters. Next, the optimal number of clusters is implemented in the cluster formation process using the K-Means algorithm. From the visualization in Figure 8, it can be concluded that the area coverage and number of cluster members are different. The number of members in cluster 4, only two stocks (MAPY and ELTY), is the smallest when compared to the number of members in other clusters, while cluster 2 has the largest number of members, 269 stocks.

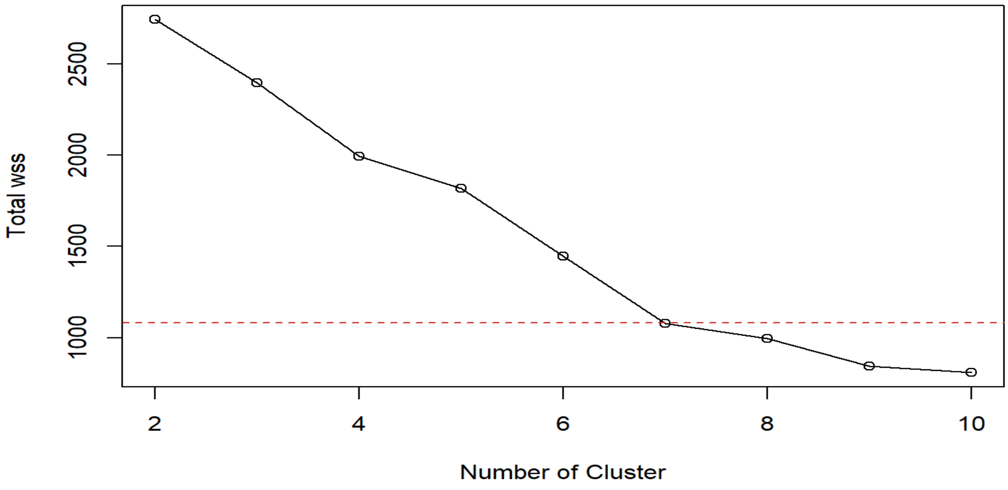


Figure 7. Elbow Method

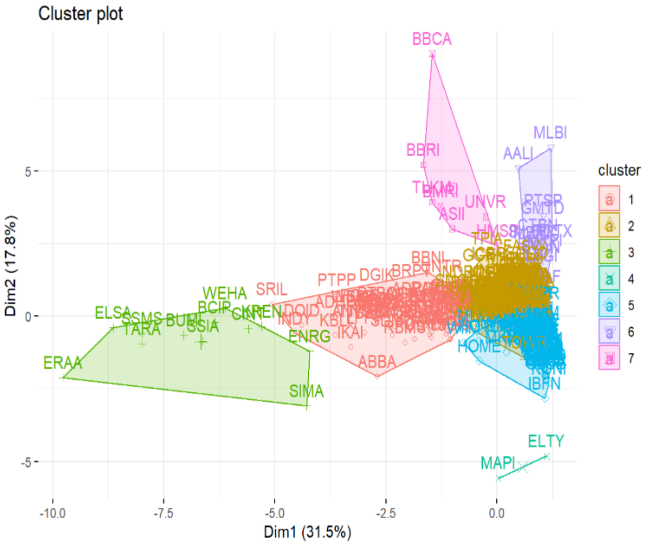


Figure 8. Results of Implementation of the K-Means Algorithm

5. Conclusions and Recommendations

This research has considered three important technical features, such as volatility, liquidity, and capacity in carrying out the stock grouping process. Of the 853 stocks listed on the IDX in 2020-2023, only 510 stocks can be used. There are 10 lists of stocks identified as outliers using the DBSCAN method. These stocks were not included in the cluster creation process using the K-Means algorithm and 7 clusters were produced with a total wss of 1080. The results of this grouping have different numbers of members and characteristics. So, stock market investors need to consider selecting issuers based on the characteristics of this grouping. Here are some conclusions to consider:

a. Cluster 1 is a level of liquidity with high consistency.

b. Cluster 2 has the largest number of members, 53% of the total stocks.

c. Cluster 3 has 12 issuers with high levels of liquidity.

d. Cluster 4 has high volatility, and its members are ELTY and MAPI stocks.

e. Cluster 5 is the lowest market capital, with an average of 2.6 trillion, with a total of 134 stocks as members.

f. Cluster 6 is the most volatile cluster, very different from the other clusters in that the minimum value of this cluster is also higher than the maximum value of the other clusters.

g. Cluster 7 has 7 stocks as members with the largest average market capital compared to other clusters.

As a note, for further research, you can add cluster search methods other than the elbow method, observe cluster shifts over time, consider simulating partitioning of training data and various testing data to be analyzed, and get the best proportions. Implement portfolio formation based on the clusters that have been found and compare the results with portfolio formation without clustering. Furthermore, similar research can be carried out by adding cluster-forming features that have a significant impact on determining stock portfolios.

6. Acknowledgement

Thank you to all levels of leadership, lecturers, staff, and all parties who have contributed to supporting us in realizing this research. We hope that Matana University will continue to be a center of excellence in education, producing a generation that is creative and contributes positively to society and the nation.

References

[1] A. Satar, M. Al Musadieq, B. Hutahayan, and others, “Enhancing Sustainable Competitive Advantage: The Role of Dynamic Capability and Organizational Agility in Technology and Knowledge Management: Indonesian Stock Exchange Evidence,” *International Journal of Operations and Quantitative Management*, vol. 29, no. 2, 2023.

[2] H. Padmanaban, “Navigating the Role of Reference Data in Financial Data Analysis: Addressing Challenges and Seizing Opportunities,” *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, vol. 2, no. 1, pp. 69–78, 2024.

[3] M. Disli, R. Nagayev, K. Salim, S. K. Rizkiah, and A. F. Aysan, “In search of safe haven assets during COVID-19 pandemic: An empirical analysis of different investor types,” *Res Int Bus Finance*, vol. 58, p. 101461, 2021.

[4] R. Raja, “Time-Series Clustering for Improving Predictions on Smart Home Appliances,” 2023.

[5] S. Wang *et al.*, “Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects,” *Information Fusion*, vol. 76, pp. 376–421, 2021.

[6] H. Rafiee, M. Aminizadeh, E. M. Hosseini, H. Aghasafari, and A. Mohammadi, “A cluster analysis on the energy use indicators and carbon footprint of irrigated wheat cropping systems,” *Sustainability*, vol. 14, no. 7, p. 4014, 2022.

[7] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,” *Inf Sci (N Y)*, vol. 622, pp. 178–210, 2023.

[8] A. E. Ezugwu *et al.*, “A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects,” *Eng Appl Artif Intell*, vol. 110, p. 104743, 2022.

[9] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, “Machine learning techniques and data for stock market forecasting: A literature review,” *Expert Syst Appl*, vol. 197, p. 116659, 2022.

[10] T. O. Kehinde, F. T. S. Chan, and S. H. Chung, “Scientometric review and analysis of recent approaches to stock market forecasting: Two decades survey,” *Expert Syst Appl*, vol. 213, p. 119299, 2023.

[11] R. H. Pasaribu, “KAJIAN TINGKAT EFISIENSI PASAR MODAL BENTUK LEMAH DI BURSA EFEK INDONESIA PADA PERIODE SEBELUM DAN SELAMA PANDEMIC COVID-19,” *Jurnal Ekonomi dan Manajemen*, vol. 1, no. 2, pp. 90–101, 2022.

[12] B. Siregar, F. A. Pangruruk, and P. A. Widjaja, “Perbandingan Berbagai Model Peramalan Indeks Harga Saham Gabungan (IHSG) di Masa Pandemi Covid-19,” *Jurnal Multidisiplin Madani*, vol. 2, no. 2, pp. 1035–1046, 2022.

[13] R. Antika, N. Satyahadewi, and H. Perdana, “ANALISIS PEMBENTUKAN PORTOFOLIO OPTIMAL MENGGUNAKAN MODEL BLACK LITTERMAN DENGAN PENDEKATAN ARCH/GARCH,” *Equator: Journal of Mathematical and Statistical Sciences*, vol. 1, no. 1, pp. 31–39, 2022.

[14] I. H. Sarker, “Machine learning: Algorithms, real-world applications and research directions,” *SN Comput Sci*, vol. 2, no. 3, p. 160, 2021.

[15] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, Dec. 2006, doi: 10.1016/j.neucom.2005.12.126.

[16] M. Somvanshi, P. Chavan, S. Tambade, and S. V. Shinde, “A review of machine learning techniques using decision tree and support vector machine,” in *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*, IEEE, Aug. 2016, pp. 1–7. doi: 10.1109/ICCUBEA.2016.7860040.

[17] R. Thupae, B. Isong, N. Gasela, and A. M. Abu-Mahfouz, “Machine Learning Techniques for Traffic Identification and Classifiacation in SDWSN: A Survey,” in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, IEEE, Oct. 2018, pp. 4645–4650. doi: 10.1109/IECON.2018.8591178.

[18] F. S. B. F. S. Board, *Artificial intelligence and machine learning in financial services: Market developments and financial stability implications*. Financial Stability Board, 2017.

[19] S. Das and M. J. Nene, “A survey on types of machine learning techniques in intrusion prevention systems,” in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, IEEE, Mar. 2017, pp. 2296–2299. doi: 10.1109/WiSPNET.2017.8300169.

[20] L. Li, J. Wang, and X. Li, “Efficiency analysis of machine learning intelligent investment based on K-means algorithm,” *Ieee Access*, vol. 8, pp. 147463–147470, 2020.

[21] S. Nanjundan, S. Sankaran, C. R. Arjun, and G. P. Anand, “Identifying the number of clusters for K-Means: A hypersphere density based approach,” *arXiv preprint arXiv:1912.00643*, 2019.

[22] D. N. Sari and I. Yunita, “TINGKAT KEPARAHAN DAN RISIKO PENYEBARAN COVID-19 DI INDONESIA DENGAN MENGGUNAKAN K-MEANS CLUSTERING,” *Seminar Nasional Official Statistics*, vol. 2020, no. 1, pp. 210–216, Jan. 2021, doi: 10.34123/semnasoffstat.v2020i1.706.

[23] N. Ulinnuha and S. A. Sholihah, “Analisis Cluster Untuk Pemetaan Data Kasus Covid-19 Di Indonesia Menggunakan K-Means,” *Jurnal MSA (Matematika dan Statistika serta Aplikasinya)*, vol. 9, no. 2, pp. 27–31, 2021.

[24] A. F. Riyadhi and R. M. Atok, “Impact of COVID-19 on Indonesia stock portfolio allocation based on a technical & fundamental approach using a machine learning algorithm,” *F1000Res*, vol. 12, p. 1475, 2023.

[25] S. Wiersma, T. Just, and M. Heinrich, "Segmenting German housing markets using principal component and cluster analyses," *International Journal of Housing Markets and Analysis*, vol. 15, no. 3, pp. 548–578, 2022.

[26] J. Höhler and A. O. Lansink, "Measuring the impact of COVID-19 on stock prices and profits in the food supply chain," *Agribusiness*, vol. 37, no. 1, pp. 171–186, 2021.

[27] C. Iman, F. N. Sari, and N. Pujiati, "Pengaruh likuiditas dan profitabilitas terhadap nilai perusahaan," *Perspektif: Jurnal Ekonomi dan Manajemen Akademi Bina Sarana Informatika*, vol. 19, no. 2, pp. 191–198, 2021.

[28] R. Handayani, S. Suhendro, and E. Masitoh, "Pengaruh profitabilitas, debt to equity ratio, price to eraning ratio dan kapitalisasi pasar terhadap return saham," *INOVASI*, vol. 18, no. 1, pp. 127–138, 2022.

[29] M. I. N. Rais, "PENENTUAN SEGMENTASI KONSUMEN PADA MARKETING DATA IFOOD MENGGUNAKAN METODE K–MEANS CLUSTERING," Fakultas Teknik Unpas, 2022.