Volume x No x xxxxx (20xx)

e-ISSN: **2615-3270**
p-ISSN: **2615-3599**

Eigen Mathematics Journal

Homepage jurnal: http://eigen.unram.ac.id

Measuring the Similarity of Stock Financial Statistics: A Cluster Analysis on West Nusa Tenggara Stocks and the Indonesia Stock Exchange

Sufria Dimi Permadi[1], Ratna Hidayati[2] , Selina Febiyanti3, Hajijatul Qubroh[4].

Faculty of Mathematics and Natural Sciences, University of Mataram, Jl. Majapahit No.62 Gomong, Kec.
Email: sufriadimii@gmail.com[1], ratnahidayat35@gmail.com[2], febiyantiselina6@gmail.com[3], hajijatulqubroh@gmail.com[4]

A B S T R A C T

This research aims to analyze the similarity of financial statistics of stocks in the West Nusa Tenggara region and stocks listed on the Indonesia Stock Exchange IDX using the hierarchical cluster analysis method. The data used includes the Return on Equity (ROE), Return on Assets (ROA), Profit Margin, and Cost to Income Ratio (CIR) ratios from the period 2020 to 2023. Data standardization is performed using z-score, and the Euclidean distance measure is used to measure the similarity between objects. The complete linkage method is applied to group stocks based on the similarity of financial characteristics. The clustering results were visualized with dendrograms and principal component plots. The analysis shows that most stocks have strong similarities in financial characteristics, regardless of the number of clusters assigned. Certain stocks consistently fall into the same cluster, indicating a significant similarity in financial statistics. These findings provide important information for investors and market analysts in grouping stocks based on similar characteristics for portfolio diversification strategies and stock performance prediction.

Keywords : Stocks, Investment, Cluster Analysis, Complete Linkage

Submitted: 18-12-2023; Received: 18-12-2023; Doi: https://doi.org/10.29303/emj.xxx.x

1. Introduction

Data mining is the process of searching for patterns and interesting information on selected data using certain methods or techniques. One of the data mining techniques for grouping data is the hierarchy technique [].

Shares are ownership rights owned by individuals (shareholders) in a company based on capital participation, thus giving them ownership and control of a certain part of the company [1]. People who invest in the capital market as shareholders are called investors. Based on data from the Indonesia Stock Exchange (IDX) and Yfinance, the number of investors in the Indonesian capital market consisting of stock, bond, and mutual fund investors increased by 56% during 2020 and reached a value of 3.87 million. Single Investor Identification (SID) until December 2020 [2]. This figure shows significant public interest in stock ownership participation.

An increase or decrease in stock price can be caused by the performance of the company itself, which can be seen from its financial statements. Financial ratios are now commonly used in analyzing financial statements. Financial

* Corresponding author.
Alamat e-mail: sufriadimii@gmail.com[1], ratnahidayat35@gmail.com[2], febiyantiselina6@gmail.com[3], hajijatulqubroh@gmail.com[4]

ratios consist of liquidity ratios, profitability, activity, solvency ratios, and market ratios. In this study, the liquidity financial ratio and the profitability financial ratio were used.

Applying Corporate Fundamentals Investors usually have specific criteria when evaluating a company. For example, the amount of dividends that are routinely paid to a company (dividend yield), debt ratio (DER), and the size of a company. A company's profit is expressed in the form of gross profit and return on equity (ROE). To make it easier to find and determine corporate issuers that have the potential to generate good investment returns, a system is needed that allows investors to see issuers according to the criteria they want, one of which is managed by issuers through the application of data mining.

2. Method

2.1 Data Standardization

Data standardization is carried out to avoid problems that will result from the use of different scale values between objects. The most common data standardization is the conversion of each object's value to the standard value or z-score by subtracting the middle value and dividing it by the standard deviation of each object. Standardization formula for each object (Walpole & Myers, 1995):

$$Z = \frac{x_i - \bar{x}}{s}$$

2.2 Distance Measurements

In cluster analysis, there are 3 measures that can be used to measure similarity between objects, one of which is the measure of proximity. The proximity measure is the most commonly used measure for metric scale data (interval or ratio). In fact, this measure is a measure of inequality. Long distances indicate small similarities, and short distances indicate large similarities. This measure focuses on the size of a value or object so that it has the same value even though the pattern is different. One measure of this proximity is the Euclidean distance. Euclidean distance is the distance measured in a straight line between the center of one facility and the center of another. The advantage of Euclidean compared to other methods is that the degree of similarity determination is higher. To find the Euclidean central distance of a facility from another facility the following formula is used:

$$d_{ij} = \left[(x_i - x_j)^2 + (y_i - y_j)^2\right]^{\frac{1}{2}}$$

2.3 Complete Linkage

Complete linkage is a method that utilizes the principle of minimum distance, which is to find the farthest distance between two clusters first, then the two clusters form a new cluster. The complete linkage method is also called the complete linkage method, determined based on the furthest distance between two objects in different clusters (furthest neighbor). This method is ideal for situations where existing objects come from very different groups. The general agglomeration algorithm first finds the smallest entry in D = {dik} and combines the corresponding objects (e.g. you and V) to get the cluster(UV). The distance between (UV) and each W cluster is calculated using the following equation

$$d_{(UV)W} = max\{d_{UW}, d_{VW}\}$$

where DUW and DVW are the distances between the farthest members of groups U and W and groups V and W respectively.

3. Research Methods

3.1 Data Collection

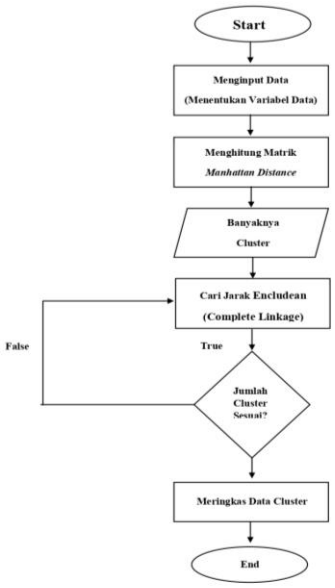
In this article, the data used is sourced from the Yahoo Finance and IDX websites. The data used is finencial statistical data, namely *Return on Equity* (ROE), *Return on Assets* (ROA), *Profit Margin* and *Cost to Income Ratio* (CIR) from the period 2020 to the 2023 period which focuses on stocks in West Nusa Tenggara and the Stock Exchange in Indonesia.

3.2 Data Processing and Analysis

The stages of clustering patterns of finencial or financial statistical similarity are carried out with the following steps:

1. Analyze descriptive statistics and inferences to prove the feasibility of both instruments as multivariate (bivariate) processed panel data
2. Calculation of the encludean distance of each object (monthly finance)
3. The process of clustering each object hierarchically using Complete Linkage.
4. Making dendograms as a result of the clustering process and their number.
5. Calculation of the number of members of each stock price cluster, analysis of the characteristics of each cluster, recurrence of the cluster in each observation period, and the length of the cluster that can survive in each period, as well as the average length of the cluster lasting.
6. Interpretation and conclusion.

3.3 Flowchart Complete Linkage



Gambar 3.1 Flowchart Complete Linkage

4. Results and Discussion

A total of 37 observations were made on stock data in West Nusa Tenggara and the Indonesia Stock Exchange during the research period. This data is assumed to be panel data and with the help of the Python programming language a *proximity* matrix with a size of 37×37 is obtained. Due to its considerable size, this matrix is not shown in the paper.

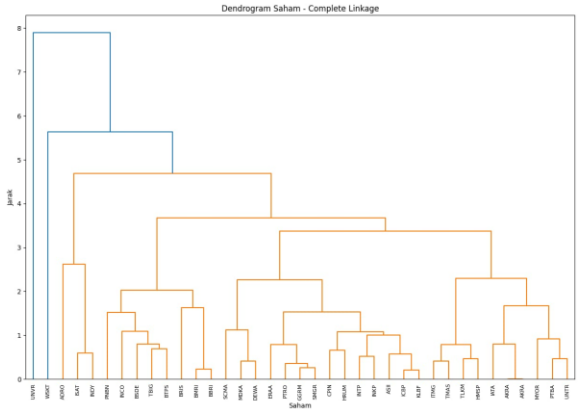
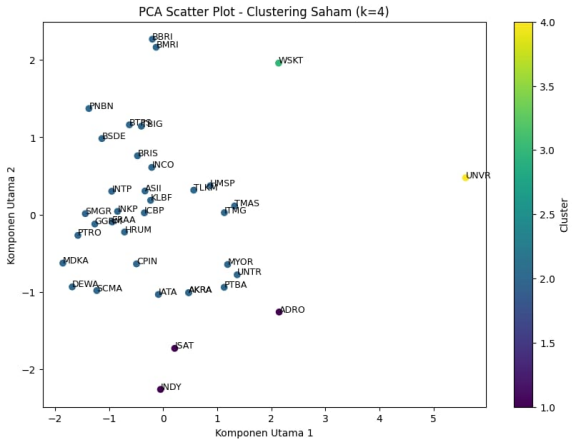


Figure 4.1 Stock Dendrogram Using Complete Linkage

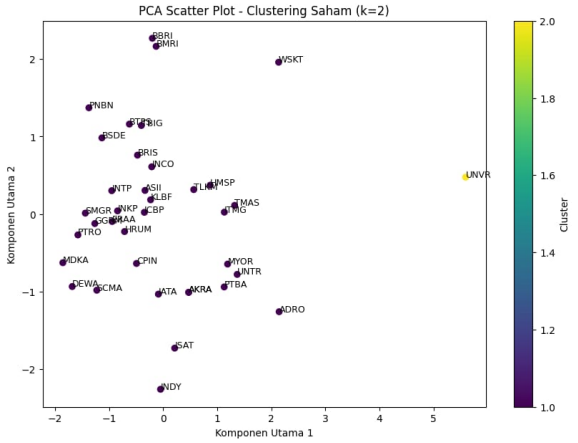
In Figure 4.1 Based on the grouping pattern, it is obtained that several stocks such as stocks in the mining and electronics sectors in NTB with the stock exchange in sub-clusters (ASII) and (HRUM), stocks in the economic sector in NTB such as (INKP), (ERAA) and stock exchanges (GGRM) and (INTP) are combined at a very short distance, showing high similarities. The degree of difference between the stocks is indicated by the depth or length of the branches on the dendogram; Longer branches show more differences, while shorter branches show more similarities.

NO	2 CLUSTER		3 CLUSTER			4 CLUSTER			
	1	2	1	2	3	1	2	3	4
1	BMRI	UNVR	BMRI	UNVR	WSKT	BMRI	UNVR	ANDRO	WSKT
2	BBRI		BBRI			BBRI		INDY	
3	ASII		ASII			ASII			
4	TLKM		TLKM			TLKM			
5	GGRM		GGRM			GGRM			
6	INTP		INTP			INTP			
7	ICBP		ICBP			ICBP			
8	TBIG		TBIG			TBIG			
9	KLBF		KLBF			KLBF			
10	SMGR		SMGR			SMGR			
11	BSDE		BSDE			BSDE			
12	PTBA		PTBA			PTBA			
13	CPIN		CPIN			CPIN			
14	MDKA		MDKA			MDKA			
15	ADRO		DRO			ITMG			
16	ITMG		ITMG			MYOR			
17	MYOR		MYOR			CMA			
18	SCMA		CMA			ERAA			
19	ERAA		ERAA			HMSP			
20	HMSP		HMSP			INKP			
21	INKP		INKP			AKRA			
22	AKRA		AKRA			BTPS			
23	BTPS		BTPS			BRIS			
24	ISAT		ISAT			IATA			
25	PNBN		PNBN			INCO			
26	WSKT		INDY			HRUM			
27	INDY		BRIS			TMAS			
28	BRIS		IATA			AKRA			
29	IATA		INCO			PTRO			
30	INCO		HRUM			UNTR			
31	HRUM		TMAS			DEWA			
32	TMAS		AKRA						
33	AKRA		PTRO						
34	PTRO		UNTR						
35	UNTR		DEWA						
36	DEWA								

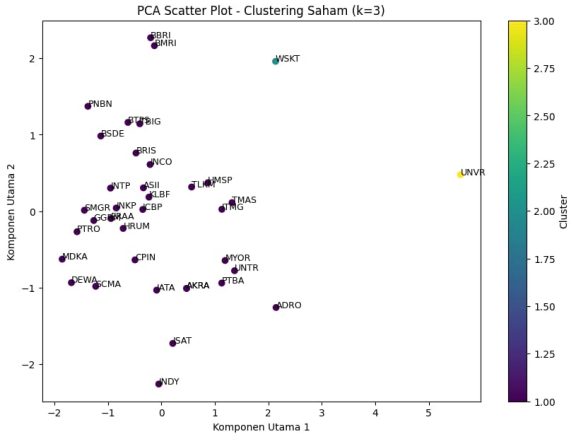
Table 4.1 Grouping of Cluster Members



Gambar 4.2 PCA Scatter Plot 2 Clustering



Gambar 4.3 PCA Scatter Plot 3 Clustering



Gambar 4.4 PCA Scatter Plot 4 Clustering

Table 4.1 obtained the results of grouping data for 37 cases based on the number of different clusters, namely 2, 3, and 4 clusters. Each row in the table shows the cluster membership of each case in each different cluster configuration.

For 2 clusters, all cases except one (UNVR) are included in the first cluster. UNVR was the only case to fall into the second cluster, suggesting that UNVR had significantly different characteristics from the other cases in this dataset. In this configuration, stocks such as BMRI, BBRI, ASII, TLKM, GGRM, INTP, ICBP, TBIG, KLBF, SMGR, BSDE, PTBA, CPIN, MDKA, ITMG, MYOR, SCMA, ERAA, HMSP, INKP, AKRA, BTPS, ISAT, PNBN, WSKT, INDY, BRIS, IATA, INCO, HRUM, TMAS, AKRA, PTRO, UNTR, and DEWA are all included in the first cluster. This suggests that these stocks have significant similarities in the variables analyzed, which distinguishes them from UNVR which falls into the second cluster.

When the number of clusters becomes 3, the grouping pattern remains relatively stable with most stocks still in the first cluster. However, there are some significant changes: ADRO, ISAT, and INDY cases move to the third cluster, and WSKT enters the third cluster. This suggests that there are smaller subgroups in the

first cluster that are beginning to be identified with an increase in the number of clusters. In this grouping, WSKT goes to the third cluster where the stocks that remain in the first cluster still maintain strong characteristic similarities.

With the grouping of 4 clusters, the breakdown becomes more specific. While most cases remained in the first cluster, ADRO and ISAT remained in the third cluster, and WSKT shifted to the fourth cluster, indicating that WSKT had significant enough differences from the others to form a separate cluster. In addition, the second cluster is only inhabited by UNVR, confirming its consistent differences from other cases.

Figure 4.3 and Figure 4.4 are visualizations using the Main Component Analysis (PCA) in the form of *scatter plot* In the data grouping for 37 stocks based on the number of different clusters, namely 2, 3, and 4 clusters are used to reduce the dimensions and visualization of clustering for each number of clusters above.

5. Conclusion and Advice

5.1 Conclusion

In this study, it can be concluded that the clustering analysis of a number of stocks, in Table 4.1 Cluster Member Grouping, shows that most stocks have consistent characteristics and are similar to each other, regardless of the number of clusters set (k=2, k=3, or k=4). Stocks that consistently fall into cluster "1" in all configurations show strong similarities in the variables analyzed. These results confirm that these stocks have similar finencial statistics, which can be an important indicator in investment decision-making and the results of this clustering are important for investors and market analysts, as they can help in grouping stocks with similar characteristics for portfolio diversification strategies. Stocks in the same cluster may exhibit similar behavior under different market conditions, so this grouping can be used to reduce risk and improve stock performance predictions.

5.2 Advice

These results provide a solid basis for further research:

- **Additional Variable Analysis**: Using more variables or different variables can be helpful in identifying additional characteristics that may affect the grouping of stocks.
- **Use of Other Clustering Methods**: Apply different clustering methods such as k-means++, or DBSCAN to see if consistent results can be obtained.

Bibliography

(2002) Big Indonesian Dictionary website. [online]. Available: https://www.kbbi.web.id/saham.

(2020) Indonesia Stock Exchange website. [online]. Available:https://www.idx.co.id/media/20221199/laporan-tahunan-2020.pdf

Irmayansyah, I., & Khaosaroh, S. (2019). Application of hierarchical agglomerative clustering method based on single linkage for grouping thesis titles. *Technolist : Scientific Journal of Information Technology and Science*, *9*(2), 53–64. https://doi.org/10.36350/jbs.v9i2.63

Prihartanti, W., Rasyid, D. A., Kunhadi, D., & Atmajawati, Y. (2023). Clustering of stock price similarity patterns using the hierarchical method. *G-Tech: Journal of Applied Technology*, *7*(4), 1760–1769. https://doi.org/10.33379/gtech.v7i4.3413

S, R. A., & Economics, F. (1945). *THE EFFECT OF ECONOMIC VALUE ADDED (EVA), NET PROFIT MARGIN (NPM), RETURN ON ASSET (ROA), RETURN ON EQUITY (ROE), EARNING PER SHARE (EPS), AND PRICE EARNING RATIO (PER) ON STOCK*

PRICES (A CASE STUDY OF PROPERTY AND REAL ESTATE COMPANIES LISTED ON BUR. 218–230.

Sophia Annisa Faisal, & Rifai, N. A. K. (2023). Application of the Hierarchical Clustering Multiscale Bootstrap Method for the Grouping of Human Development Index Indicators in 2021 in West Java. *Bandung Conference Series: Statistics*, *3*(1), 100–106. https://doi.org/10.29313/bcss.v3i1.6327

Srimurdianti, A., Sukamto, Setiawan, W., Esyudha, E., & Pratama. (2023). *JEPIN (Journal of Informatics Education and Research) Data Mining for Stock Grouping in the Energy Sector with the K-Means Method*. *9*(1), 76–81. https://jurnal.untan.ac.id/index.php/jepin/article/viewFile/62509/75676597024

Walpole, R. E., & Myers, R. H. (1995). Science of Opportunity and Statistics for Engineers and Scientists. Bandung: ITB

Wijaya, A., Ar, F., & Rusyana, A. (2021). Comparison of Complete Link Cluster Methods and Average Links for Regency/City Poverty Grouping in Indonesia. *Journal of Data Analysis*, *3*(1), 13–25.