Available online at https://journal.rescollacomm.com/index.php/ijqrm/index

International Journal of Quantitative Research and Modeling

e-ISSN 2721-477X
p-ISSN 2722-5046

Vol. 2, No. 4, pp. 201-208, 2021

IDX30 Stocks Clustering with K-Means Algorithm based on Expected Return and Value at Risk

Ahmad Fawaid Ridwan[1*], Subiyanto[2], Sudradjat Supian[3]

[1]*Master of Mathematics Program, Faculty of Mathematics and Sciences, Universitas Padjadjaran,*
Jl. Raya Bandung-Sumedang KM 21, Jatinangor, Sumedang, West Java, 45363, Indonesia
[2]*Department of Marine Science, Faculty of Fishery and Marine Science, Universitas Padjadjaran,*
Jl. Raya Bandung-Sumedang KM 21, Jatinangor, Sumedang, West Jav, 45363, Indonesia
[3] *Department of Mathematics, Faculty of Mathematics and Sciences, Universitas Padjadjaran,*
Jl. Raya Bandung-Sumedang KM 21, Jatinangor, Sumedang, West Java, 45363, Indonesia

**Corresponding author email: ahmad20034@mai.unpad.ac.id*

Abstract

Stocks are one of the investment instruments available in the capital market. Several indices show the characteristics of stocks listed on the Indonesia Stock Exchange. IDX30 is one of several indications that show the combined stocks are stocks with large market capitalization, high liquidity, and good fundamentals. The selection of assets to be allocated in the portfolio is an important factor in investing where the purpose of investing is to maximize returns and minimize risk. This study aims to classify stocks that have certain characteristics based on the expected return and value at risk of the stocks incorporated in IDX30 with a clustering algorithm. The clustering algorithm used is the K-Means algorithm. K-Means is a non-hierarchical clustering algorithm by groups each object based on its proximity to the cluster center. The method used in this research is a clustering simulation study using the K-Means algorithm on IDX30 stock data. By identifying the characteristics of the stock based on the characteristics of the cluster formed, it is hoped that it can be considered in choosing the assets to be used in the formation of an optimal portfolio.

Keywords: expected return, IDX30, K-Means, stock, value at risk

1. Introduction

Investment is an activity of placing funds in one or more types of assets in the hope of obtaining profits or increasing the value of investments in the future during a certain period (Safitri et al., 2020; Hamka et al., 2020; Pradesyah and Triandhini, 2021). Investment arrangements, in general, can be done through the capital market and money market in the form of securities (Ghosh and Mahanti, 2014). The level of profits obtained in the capital market in the form of securities, especially stocks, is greater than the level of profits in the money market investments in the form of deposits (Strassberger, 2006). In general, risks and benefits have a positive relationship, namely the greater the risk, the greater the benefits (Gambrah and Pirvu, 2014).

Stocks are securities as proof of participation or ownership of a person or legal entity in a company (Putri and Hasibuan, 2020). The Indonesia Stock Exchange (IDX) has several indicators that show the characteristics of the listed stocks. IDX30 is one of several indicators that measure the price performance of 30 stocks that are supported by good company fundamentals and have high liquidity and large market capitalization (IDX, 2021). The purpose of investing is to maximize and face challenges so that the assets to be allocated in the portfolio are an important factor in investing (León et al., 2017).

The problem in the portfolio is how to determine the right composition for each asset in it so that the investor's goals are achieved. In terms of the selection of combined assets, the more diverse the assets in the portfolio, the more diversified it can be so that it is considered able to reduce the risk that occurs in the portfolio (Subekti et al., 2017). The diversification approaches that exist in portfolios seem quite diverse, but this is still an interesting research topic. Stocks are selected to compose a portfolio to achieve optimal diversification.

The formation of a portfolio with an asset selection strategy has been widely carried out by utilizing clustering techniques to group assets that have certain characteristics. One method that is often used to group objects based on their attributes is the K-Means algorithm. K-Means is a non-hierarchical clustering algorithm where the grouping begins by determining the number of clusters to be formed and then dividing the data as many as clusters based on the proximity of the object to the center of the cluster (Cebeci and Yildiz, 2015). Therefore, the purpose of this study is to determine the grouping of IDX30 stocks using the K-Means algorithm based on the expected return and value at risk of the stock. The main objective of this study is to classify stocks that have certain characteristics based on the expected return and value at risk of the stocks incorporated in IDX30 with a clustering algorithm.

2. Literature Review

Clustering analysis is a multivariate technique that has the main objective of grouping objects based on their characteristics. The clustering technique is one of the unsupervised data mining methods, meaning that this method is applied without training and teaching and does not require an output target (Lao and Shao, 2004). There are two types of clustering analysis, namely hierarchical and non-hierarchical. In non-hierarchical methods such as K-Means which is a grouping of objects by determining the number of groups that will be formed first, that is as many as k groups. All objects to be analyzed are then grouped based on the similarity of characteristics of these objects (Feng and Zhang, 200).

The K-Means clustering technique is a method that tries to partition the data into several groups using the average value as the center of the cluster. The purpose of this grouping is to reduce diversity within a group and maximize diversity between groups (Cebeci and Yildiz, 2015). This grouping aims to provide an objective function. The advantages of K-Means are that it can handle large data, and the cluster members can be customized, but it has drawbacks where the K-Means algorithm is sensitive to outliers, sensitive to data scale, and ineffective for various clusters (Feng and Zhang, 2020).

Many studies are related to improving the clustering accuracy of the K-Means algorithm. In terms of distance metrics based on outlier detection, Deb and Dey (2017) have identified outliers based on the distance between the data point and the nearest centroid. In terms of preprocessing techniques, K-Means++ was used as an additional filtering step in Im et al. (2020) to eliminate data points as outliers before applying conventional K-Means. Although the results of the clustering of these techniques are encouraging, the clustering process is only carried out on the remaining data without outliers. Outlier data is completely removed and not classified to known clusters as originally collected. In addition to the influence of outliers, the similarity of different metrics causes various forms of clustering which can increase or decrease the accuracy of k-means clustering (Gupta and Chandra, 2020). Among the distance of metrics, Euclidean distance is commonly used with K-Means to group data. Grouping the same dataset based on these metrics can result in a different way of grouping which is highly dependent on the distance model that fits the data domain (Bekhet and Ahmed, 2020).

Several studies on portfolio and asset selection strategies have been carried out by utilizing clustering techniques in the formation of an optimal asset portfolio. Jiang et al. (2014) in their research discuss the use of clustering techniques on the expected value and risk value of the stock to select the assets to be used in building a stock portfolio using the cardinality-constrained mean-variance method. Furthermore, Putra et al. (2021) in their research use clustering techniques to select assets to be used in building a portfolio from the KOMPAS-100 Stock Index using B-Spline-based clustering.

In this paper, to improve the clustering accuracy of the K-Means algorithm, preprocessing data is carried out by transforming the historical data into attribute data that represents the characteristics of each object. In selecting assets to build an optimal portfolio, the most important factors to consider are the rate of return provided and the risk of loss that may be experienced in a certain period. In this case, the attribute used is expected return, which represents the level of possibility that an asset has in generating profits in a certain period (Sukono et al., 2017). In addition, to represent the risk of loss of an asset, the value at risk attribute is used where this value is more focused on the downside risk of an asset (Ismanto, 2016). Furthermore, the removal of outliers is not carried out in the algorithm process because all objects are considered important and remain grouped in certain clusters.

3. Materials and Methods

3.1. Materials

This study uses stock data that are incorporated in IDX30 which consists of 30 stocks. The data used is historical daily stock closing price data. The data period starts from November 1, 2020, to October 29, 2021, which is downloaded from https://finance.yahoo.com.

3.2. Methods

The method used in this research is an experimental study by simulating the stock clustering of the K-Means algorithm. The clustering process is carried out using the Python programming language. The steps taken in this research are as follows.

1. Collecting data.

   The data used directly uses Python coding. To simplify the calculation process, the downloaded data is then formed into a data frame format.

2. Calculating the expected return and value at risk of stocks

   The expected return is the level of profit expected from stock within a certain period. The expected return can be calculated using the following equation (Sukono et al., 2017).

$$\mu = \frac{\sum_{t=1}^{n} \left(\frac{S_t - S_{t-1}}{S_{t-1}} \right)}{N} \tag{1}$$

   where μ is expected return, S_t is the stock price in period t, S_t is the stock price in period $t-1$, and N is the number of observations. Meanwhile, value at risk is the level of loss that can occur in stock within a certain period. Value at risk can be calculated using the following equation (Sukono et al., 2017).

$$VaR = -(\mu + \sigma z_\alpha) \tag{2}$$

   where VaR is the value at risk, μ is expected return, σ is the standard deviation of daily returns, and z_α is the percentile of the standard normal distribution with a significance level of$(1-\alpha)\%$.

3. Determine the number of clusters

   The number of clusters in the K-Means algorithm can be determined using the Elbow method which produces a variance plot that is checked after the increase in the number of clusters is plotted against the number of clusters (Naeem and Wumaier, 2018). Each number of clusters is evaluated using Sum Squared Error (SSE) with the following equation (Hassan et al., 2021).

$$SSE = \sum_{i=1}^{k} \sum_{xj \in C_i} \left\| x_j - \bar{x}_i \right\|^2 \tag{3}$$

   where x_j is an object in each cluster and C_i is the centroid of the cluster. If the line diagram looks like an arm, then the "elbow" on the arm is the value of k which is the corresponding k (Naeem and Wumaier, 2018).

4. Clustering with the K-Means algorithm

   For k clusters, K-Means is based on an iterative algorithm minimizing the sum of distances from each object to its cluster centroid. The objects are moved between clusters until the sum cannot be decreased anymore. K-Means algorithm involves the following steps (Cebeci and Yildiz, 2015):

   (1) Centroids of k clusters are chosen $\boldsymbol{X}$ randomly.

   (2) Distances between data points and cluster centroids are calculated. The distance is measured with the Euclidean norm by the following equation.

$$D_{ij}^2 = \left\| x_j - v_i \right\| \tag{4}$$

   where x_{ij} represents the number of data points in cluster i.

   (3) Each data point is assigned to the cluster whose centroid is closest to it.

   (4) Cluster centroids are updated by using the following formula.

$$v_i = \frac{\sum_{i=1}^{n_i} x_{ij}}{n_i} \tag{5}$$

   where v_i is the centroid of cluster i, x_{ij} is the object in cluster i, and n_i is the number of objects in cluster i.

(5) Distances from the updated cluster centroids are recalculated.
(6) If no data point is assigned to a new cluster the run of the algorithm is stopped. Otherwise, the steps from (2) to (5) are repeated for probable movements of data points between the clusters

5. Describing the clusters

The clusters that have been formed are then described according to the characteristics of the objects in these clusters. This characteristic can be seen by referring to the attributes used as the basis for clustering.

4. Results and Discussion

4.1. Daily Closing Price, Expected Return, and Value at Risk

The data used to determine the expected return and value at risk are daily closing prices for 30 stocks indexed on IDX30 for the period August 2021 to January 2022 starting from November 1, 2020, to October 29, 2021 can be seen in Figure 1.

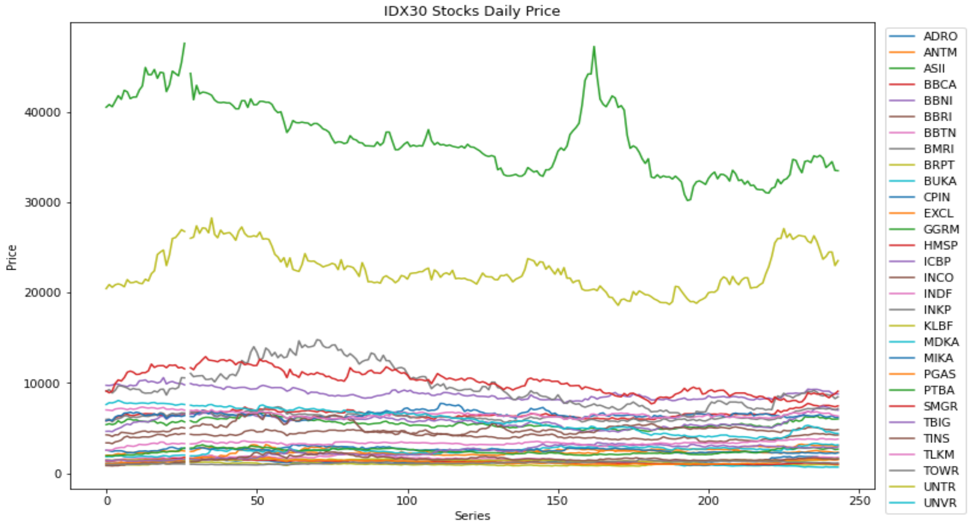

Figure 1. IDX30 daily closing price data

Based on this data, the expected return and value at risk of each IDX30 stock for 1 year can be seen in Table 1.

Table 1. Expected return and value at risk IDX30

No	Stock Code	Expected Return	Value at Risk	No	Stock Code	Expected Return	Value at Risk
1	ADRO	0.00202	0.04339	16	INCO	0.00099	0.04978
2	ANTM	0.00367	0.06479	17	INDF	-0.00028	0.02618
3	ASII	0.00058	0.03380	18	INKP	0.00011	0.05755
4	BBCA	0.00110	0.02443	19	KLBF	0.00052	0.03701
5	BBNI	0.00196	0.03408	20	MDKA	0.00304	0.05659
6	BBRI	0.00126	0.03505	21	MIKA	-0.00010	0.04040
7	BBTN	0.00135	0.03982	22	PGAS	0.00197	0.04843
8	BMRI	0.00087	0.03112	23	PTBA	0.00166	0.03812
9	BRPT	0.00079	0.05520	24	SMGR	0.00028	0.04317
10	BUKA	-0.00669	0.06829	25	TBIG	0.00338	0.05270
11	CPIN	0.00060	0.03865	26	TINS	0.00335	0.06799
12	EXCL	0.00219	0.03978	27	TLKM	0.00190	0.03225
13	GGRM	-0.00030	0.03236	28	TOWR	0.00105	0.03777
14	HMSP	-0.00075	0.03515	29	UNTR	0.00104	0.04195
15	ICBP	-0.00036	0.02677	30	UNVR	-0.00201	0.03580

The expected return and value at risk data are then used as attribute values for each stock to be grouped. The expected return is used to represent the expected return in the future, while the value at risk is used to represent the risk of loss that can be experienced in investing in a stock.

4.2. Determining the Number of Clusters

The first step of K-Means clustering is to determine the number of clusters. In this study, the number of clusters was determined by the Elbow method where for each number of clusters the SSE value was evaluated. The SSE value obtained for each number of clusters is then plotted as shown in Figure 2.

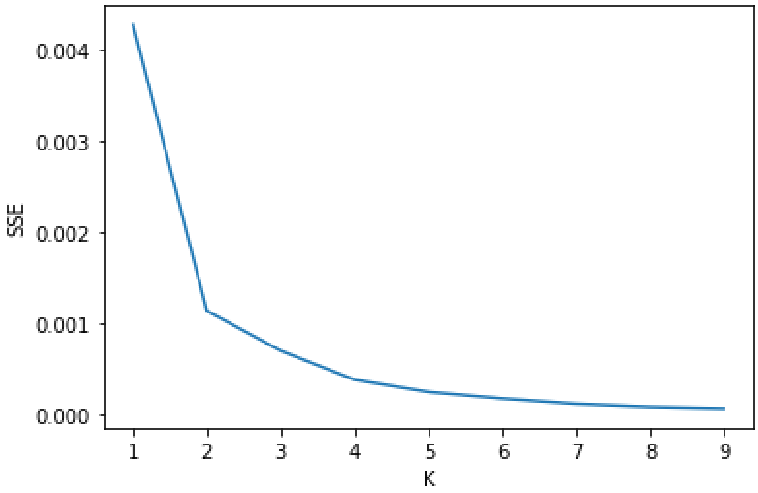

Figure 2. SSE Plot for each number of clusters

Based on Figure 2, it can be seen that the SSE value decreases as the number of clusters increases. The plot also forms like an arm where curved at the elbow. The node that forms the elbow is the number of clusters that will be used for the next step. It can be seen that the elbow node lies at $K = 2$ with an SSE value of 0.0011, which means that the number of clusters to be used is 2 clusters.

4.3. Clustering with K-Means Algorithm

Clustering begins by randomly determining 2 cluster centroids, then each object is clustered according to the closest distance from the centroid. Next, the centroid is updated and the grouping process by the closest distance is repeated until the objects that are clustered do not move cluster towards the centroid that is continuously updated. The results of clustering can be seen in Figure 3.

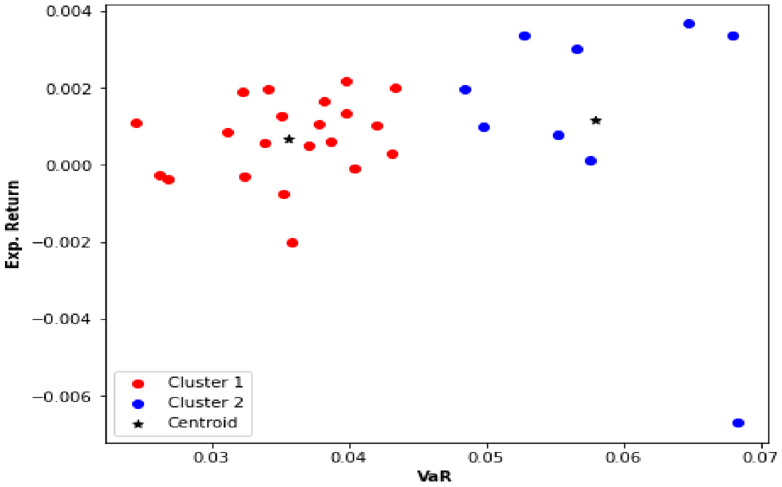

Figure 3. Scatter Plot of IDX30 Clustering with K-Means

In Figure 3, it can be seen that the scatter plot shows that the clustered objects correspond to their closest distance to one of the centroids. It is also seen that there are outlier objects that are located too far from the centroids but are still grouped in certain clusters. This is one of the weaknesses of the K-Means method where the results of the clustering are sensitive to outliers. However, the outlier object is considered important and is still included in the clustering process since the purpose of the clustering is to determine the grouping of each stock listed on IDX30.

4.4. Cluster Description

The results of grouping using the K-Means algorithm obtained 2 clusters where each cluster has different characteristics according to the characteristics of the objects contained in it. The members of each cluster can be seen in Table 2 and Table 3.

Table 2. List of IDX30 stocks in cluster 1

No	Stock Code	Expected Return	Value at Risk
1	ADRO	0.00202	0.04339
2	ASII	0.00058	0.03380
3	BBCA	0.00110	0.02443
4	BBNI	0.00196	0.03408
5	BBRI	0.00126	0.03505
6	BBTN	0.00135	0.03982
7	BMRI	0.00087	0.03112
8	CPIN	0.00060	0.03865
9	EXCL	0.00219	0.03978
10	GGRM	-0.00030	0.03236
11	HMSP	-0.00075	0.03515
12	ICBP	-0.00036	0.02677
13	INDF	-0.00028	0.02618
14	KLBF	0.00052	0.03701
15	MIKA	-0.00010	0.04040
16	PTBA	0.00166	0.03812
17	SMGR	0.00028	0.04317
18	TLKM	0.00190	0.03225
19	TOWR	0.00105	0.03777
20	UNTR	0.00104	0.04195
21	UNVR	-0.00201	0.03580

Based on Table 2, it can be seen that cluster 1 consists of 21 stocks. These stocks have expected returns that vary from -0.00201 to 0.00219 with an average expected return of 0.00069 whereas 6 stocks have negative expected returns, namely GGRM, HMSP, ICBP, INDF, MIKA, and UNVR. It is also seen that stocks in cluster 1 have a lower value at risk than in cluster 2 with a value at risk range of 0.02443 to 0.04339 and an average value at risk of 0.03557.

Table 3. List of IDX30 stocks in cluster 2

No	Stock Code	Expected Return	Value at Risk
1	ANTM	0.00367	0.06479
2	BRPT	0.00079	0.05520
3	BUKA	-0.00669	0.06829
4	INCO	0.00099	0.04978
5	INKP	0.00011	0.05755
6	MDKA	0.00304	0.05659
7	PGAS	0.00197	0.04843
8	TBIG	0.00338	0.05270
9	TINS	0.00335	0.06799

Based on Table 3, it can be seen that cluster 1 consists of 9 stocks. These stocks have an expected return that varies from -0.00669 to 0.00367 with an average expected return of 0.00118 where there is 1 stock that has a negative expected return, namely BUKA. It is also seen that stocks in cluster 1 have a higher value at risk than in cluster 2 with a value at risk range of 0.04843 to 0.06829 and an average value at risk of 0.05792.

5. Conclusion

Based on the grouping process using the K-Means algorithm on IDX30 stock data, the conclusions are as follows:

1. The grouping results form 2 stock clusters with the Sum of Squared Error value of 0.0011.
2. Cluster 1 which consists of 21 stocks has varying expected returns and lower value at risk, while Cluster 2 which consists of 9 stocks have varying expected returns and higher value at risk.

References

Bekhet, S., & Ahmed, A. (2020). Evaluation of Similarity Measures for Video Retrieval. *Multimedia Tools and Applications*, *79*, 6265–6278.

Cebeci, Z., & Yildiz, F. (2015). Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures. *Journal of Agricultural Informatics*, *6*(3), 13–23.

Deb, A. B., & Dey, L. (2017). Outlier Detection and Removal Algorithm in K-Means and Hierarchical Clustering. *World Journal of Computer Application and Technology*, *5*(2), 24–29.

Feng, Z., & Zhang, J. (2020). Nonparametric K-means algorithm with applications in economic and functional data. *Communications in Statistics-Theory and Methods*, 1-15.

Gambrah, P., & Pirvu, T. (2014). Risk Measures and Portfolio Optimization. *Journal of Risk and Financial Management*, *7*(3), 113–129.

Ghosh, A., & Mahanti, A. (2014). Investment Portfolio Management : A Review from 2009 to 2014. *Proceedings of 10th Global Business and Social Science Research Conference*, *1*(1), 1–21.

Gupta, M. K., & Chandra, P. (2020). An Empirical Evaluation of K-Means Clustering Algorithm Using Different Distance/Similarity Metrics. *Proceedings of ICETIT 2019*, *605*, 884–892.

Hamka, H., Jupri, M., & Budiono, R. (2020). The Influence of Financial Literacy on Interest in Investing for the Academic Community of Akademi Keuangan & Bisnis Indonesia Internasional (AKBII), Bandung, Indonesia. *International Journal of Business, Economics, and Social Development*, *1*(1), 1-12.

Hassan, B. A., Rashid, T. A., & Hamarashid, H. K. (2021). A Novel Cluster Detection of COVID-19 Patients and Medical Disease Conditions using Improved Evolutionary Clustering Algorithm Star. *Computers in Biology and Medicine*, *138*(1), 104866.

Im, S., Moseley, B., Sun, X., & Zhou, R. (2020). Fast Noise Removal for K-Means Clustering. *ArXiv*, 1–10.

Ismanto, H. (2016). Analisis Value at Risk dalam Pembentukan Portofolio Optimal (Studi Empiris pada Saham-saham). *University Research Colloquium 2016*, *3*(1), 243–255.

Jiang, K., Li, D., Gao, J., & Yu, J. X. (2014). Factor Model Based Clustering Approach for Cardinality Constrained Portfolio Selection. *IFAC Proceedings*, *19*(3), 10713–10718.

León, D., Aragón, A., Sandoval, J., Hernández, G., Arévalo, A., & Niño, J. (2017). Clustering Algorithms for Risk-Adjusted Portfolio Construction. *Procedia Computer Science*, *108*, 1334–1343.

Lao, L. J., & Shao, Y. M. (2004). Dynamic Clustering Analysis of Return Series of Industrial Indexes in Chinese Stock Market [J]. *The Study of Finance and Economics*, *11*.

Naeem, S., & Wumaier, A. (2018). Study and Implementing K-mean Clustering Algorithm on English Text and Techniques to Find the Optimal Value of K. *International Journal of Computer Applications*, *182*(31), 7–14.

Putra, Y. E., Saepudin, D., & Aditsania, A. (2021). Portfolio Selection of KOMPAS-100 Stocks Index Using B-Spline Based Clustering. *Procedia Computer Science*, *179*(2020), 375–382.

Putri, D. M., & Hasibuan, L. H. (2020). Penerapan Gerak Brown Geometrik pada Data Saham PT. ANTM. *Mathematics & Applications Journal*, *1*(1), 1–10.

Safitri, I. N. N., Sudradjat, S., & Lesmana, E. (2020). Stock portfolio analysis using Markowitz model. *International Journal of Quantitative Research and Modeling*, *1*(1), 47-58.

Strassberger, M. (2006). Capital Requirement, Portfolio Risk Insurance, and Dynamic Risk Budgeting. *Investment Management and Financial Innovations*, *3*(1), 78–88.

Subekti, R., Kusumawati, R., & Sari, E. R. (2017). K-Means Clustering dan Average Linkage dalam Pembentukan Portfolio Saham. *Seminar Matematika Dan Pendidikan Matematika UNY 2017*, 219–224.

Sukono, Sidi, P., Bon, A. T. Bin, & Supian, S. (2017). Modeling of Mean-VaR Portfolio Optimization by Risk Tolerance when the Utility Function is Quadratic. *AIP Conference Proceedings*, *1827*(020035).

Pradesyah, R., & Triandhini, Y. (2021). The Effect Of Third Party Funds (Dpk), Non Performing Financing (NPF), And Indonesian Sharia Bank Certificates (SBIS) On Sharia Banking Financing Distribution In Indonesia. *International Journal of Business, Economics, and Social Development*, *2*(2), 72-77.