Customer Segmentation Using the K-Means Clustering as a Strategy to Avoid Overstock in Online Shop Inventory

Anindhita Dewabharata

{anindhita.d@universitaspertamina.ac.id}

Universitas Pertamina, Jalan Teuku Nyak Arief, Simprug, Kebayoran Lama, Jakarta, Indonesia

Abstract. In recent years, the fast-growing internet technology has increased online activities, including online shopping. As a result, many companies doing business in thisarea successfully earn high profits. With the e-shopping success, many online shops compete to offer more variety products and let customers have more choices. However,if the seller does not know the prospective customers' purchase interests, offering morecollections can increase the risk of overstock in inventory. Therefore, this study aims to reduce overstock by focusing on customer segmentation to determine purchase interests.First, as a case study, we use an online shop in one popular online marketplace in Indonesia. Then, we use the k-means algorithm clustering method to determine the segmentation of prospective customers. Finally, we recommend the inventory strategy for online shops to prioritize the top three prospective customers segment as the target market.

Keywords: Customer Segmentation, Inventory, Overstock, Clustering, K-Means.

1 Introduction

In recent years, the number of internet users in Indonesia has increased. According to Katada[1], e-commerce users in Indonesia will increase to 212.2 million people in 2023, where usersconsist of sellers and buyers. Moreover, in Bank Indonesia's 2020 annual report, the nominalvalue of e-commerce transactions in 2020 increased by Rp. 266.3 Triliun [2].

With the increasing nominal of e-commerce, many online shops compete to offer more variety products and let customers have more choices. However, if the seller misunderstands customer purchase interests, offering more collections can increase the risk of overstock in inventory.

One online shop that has an overstock problem is the Noi.Project online shop. This shop sells hijab products through e-commerce and social media. Hijab products sold include the pashmina hijab, instant hijab, and rectangular hijab. Unfortunately, only 20% of hijabs were sold at the sale, so the remaining unsold hijabs are stored in the warehouse. According to Gattorna [3] this happens because the seller only focuses on the supplier but forgets that the end goal is the customers. Therefore, this study aims to reduce overstock by focusing on customer segmentation to determine purchase interests.

ICONIC-RS 2022, March 31-April 01, DKI Jakarta, Indonesia

DOI 10.4108/eai.31-3-2022.2320688

This study uses the k-means algorithm as a clustering method to categorize customers based on similar characteristics [4]. The k-means method uses the average value as the cluster's center and does not depend on the order in which the dataset is entered [5]. So, it can produce information in customer scoring and customer profiling more precisely.

Several variables are widely used in customer behavior research to get customer characteristics: recency, frequency, and monetary (RFM). However, according to Atyanto etal [6] and Angelie, [7], the three variables have not been able to represent the actual customercharacteristics and recommended to use of demographics, psychographics, geography, and behavior of prospective customers. Thus, this study uses these four variables.

2 Literature review

2.1 Customer segmentation

Customer segmentation is to group customers according to specific characteristics in common[6]. The purpose of segmenting is to improve the target market share and be able to design products that are more responsive to consumer needs. Table 1 [8] shows that customer segmentation is divided into four variables.

Table 1. Customer segmentation variables

Geography	Demographics	Psychographics	Behavior
Region	Age	Lifestyle	Purchase Status
City	Gender	Personality	Usage Rate
District	Work	Social Class	Purchase Interest
	Education		
	Income		
	Religion		

Table 2. Customer characteristics

Characteristics	Customer Class
Superstar	a) Customers who have high loyalty. b) Customers can afford to incur high costs in a single purchase. c) Have a high frequency of spending.
Golden customer	a) Customers can afford to spend less than the superstar class of customers. b) Have a high frequency of purchases or the same as the superstar customer class.
Typical customer	a) Customers can incur lower spending costs than the golden customer class. b) It has a lower spending frequency than the golden customer class.
Occasional customer	a) The costs incurred by the customer in one purchase are lower than the typicalcustomer class. b) The frequency of spending is lower than the typical customer class.
Dormant customer	a) Customers incur the lowest spending costs. b) The lowest frequency of customer spending.

Customer segmentation analysis can be categorized based on the characteristics of the customer class and divided into five classes: superstar, golden customer, typical customer, occasional customer, and dormant customer. These classes are distinguished based on customers' costs and frequency of product purchases. Table 2 shows a more detailed explanation of each class [9].

2.2 K-Means algorithm

The k-means method has been researched by many researchers from various fields of science [10]. The advantage of this method is the way the data is grouped without knowing the target class. However, the drawbacks of the k-means method are that it is difficult to determine the optimal number of clusters [11]. Thus elbow method and validation are needed to determinethe number of clusters. The clustering steps using the k-means method are as follows:

1) Select the number of clusters k as the number of clusters to be formed.
2) Determine the k data, which is the center point or the initial centroid of the cluster location.
3) Group the data into k clusters according to the nearest predetermined centroid point.
4) Update the value of the centroid point and repeat step 3 until the value of the centroid point does not change.

3 Methodology

The research methodology shows the research stages that are shown in **Figure 1**.

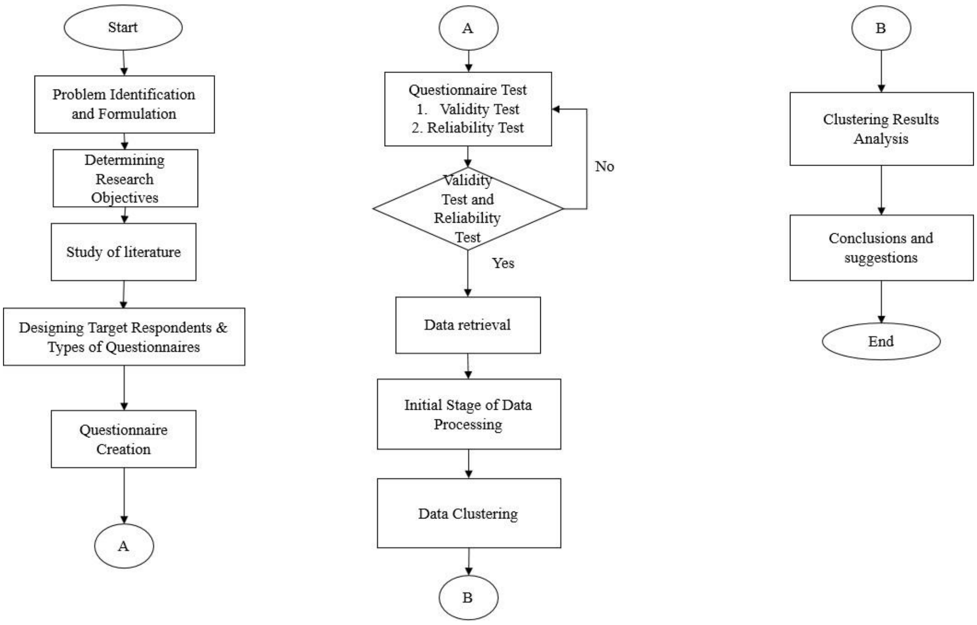
Start
Problem Identification and Formulation
Determining Research Objectives
Study of literature
Designing Target Respondents & Types of Questionnaires
Questionnaire Creation
A
A
Questionnaire Test
1. Validity Test
2. Reliability Test
Validity Test and Reliability Test
No
Yes
Data retrieval
Initial Stage of Data Processing
Data Clustering
B
B
Clustering Results Analysis
Conclusions and suggestions
End

Fig. 1. Research methodology diagrams

3.1 Data collection

This study carries out data collection by distributing questionnaires to the respondents, with the criteria selected in filling out this questionnaire as follows:

1) A woman who wears a hijab and is an end-user.
2) Never shop online.
3) A woman who is wearing a hijab every day.
4) Focus on buyers who use hijab products directly and not purchases for others or gifts.
5) Age in the range of 24 years to 65 years.
6) Domiciled in Surabaya, Jakarta, Medan, Bogor, Depok, Tangerang, Bekasi (Bodetabek) and Bandung areas.

The sampling technique in this study is divided into two ways, namely probability, and non-probability. The probability technique is a sampling technique for each member of the population who has the same opportunity to be selected [12]. This sample uses the cluster random sampling method by taking random samples from the general public in a particular area group: Surabaya, Jakarta, Medan, Bogor, Depok, Tangerang, Bekasi (Bodetabek), and Bandung.

The non-probability technique is a sampling technique that is carried out based on several researchers' considerations [12]. The method used in this technique is purposive sampling, which is deliberately based on the criteria set by the researcher. The researcher's sampling comes from Noi.Project customers. Later, researchers can observe the desire to buy hijab from prospective customers' perspectives, namely the general public and Noi.Project customers.

Further, the questionnaire is tested using validity and reliability tests. These tests are carried out with the help of SPSS software.

3.1.1 Questionnaire sample

Researchers distributed a questionnaires sample to 30 respondents. The sample of respondents consisted of 13 respondents from the Noi.Project customers and 17 respondents from the general public. After distributing the sample questionnaire and getting the result, the next stepis to test the questionnaire using validity and reliability tests. The purpose of this is to determine the feasibility of the questionnaire. These tests are carried out with the help of SPSSsoftware.

3.1.2 Validity test and reliability test

The validity test is conducted to determine the eligibility of each attribute. It is valid if the calculated r-value is greater than the r table value of 0.361 with a significance level of 0.05. Table 3 shows the validity test results with all questions for each indicator used in this studyare valid.

The reliability test is carried out to determine whether the questionnaire instrument can provide a constant measure if the measurement is repeated. It can be reliable if the CronbachAlpha value is > 0.6. Table 4 shows all indicators of each existing variable produce an Alphavalue greater than 0.6, which means reliable. So all the questions in this questionnaire can betrusted and used for this research.

Table 3. Validity test result

Variable	Indicator	Corrected Item Total Correlation / r count	r table	Criteria
Geography	City	0.991	0.361	Valid
Psychographics	Cost	0.991	0.361	Valid
	Shopping Style	0.991	0.361	Valid
Behavior	Frequency	0.991	0.361	Valid
	Model	0.991	0.361	Valid
	Types of Hijab	0.991	0.361	Valid
	Hijab Design	0.991	0.361	Valid

Table 4. Realibility test result

Variable	Indicator	Cronbach Alpha	Information
Geography	City	0.978	Reliable
Psychographics	Cost	0.978	Reliable
	Shopping Style	0.978	Reliable
Behavior	Frequency	0.978	Reliable
	Model	0.978	Reliable
	Types of Hijab	0.978	Reliable
	Hijab Design	0.978	Reliable

According to validity and reliable test, the questionnaire is valid and reliable. Furthermore, the questionnaire is distributed to 200 respondents. The questionnaire is filled out for ten days,from June 21, 2021, to July 1, 2021.

3.1.3 Questionnaire result

After collecting the results for 200 respondents, we conduct the descriptive analysis. The result is shown in Table 5. The average age of the respondents who filled out the questionnaireis 30 years, with the minimum and maximum ages of respondents between 24 and 60 years. The average frequency of buying hijab online is 0.99 or rounded up to one time a month.

Table 5. Descriptive statistics

Variable	Indicator	N	Min. Value	Max. Value	Mean	Std. Dev
Geography	City	200	0	4	1.57	1.18
Psychographics	Cost	200	0	4	1.69	1.11
	Shopping Style	200	0	2	0.71	0.74
Behavior	Frequency	200	0	4	0.99	1.13
	Model	200	0	2	0.82	0.59
	Type	200	0	9	4.72	2.36
	Color Design	200	0	3	0.92	0.92
Demographics	Age	200	24	60	30	7.22
Customer Profile Suitability	Online shopping	200	0	0	0	0
	Age range 24-65 years old	200	0	0	0	0
	Live in Surabaya, Jakarta, Medan, Bodetabek, and Bandung	200	0	0	0	0

In addition, the customer profile suitability variable, it is known that the average value is 0. This value is not the actual value but is a labeling of the answer choices contained in the questionnaire. Where the value 0 is the answer choice yes. So it can be concluded that the variable suitability of the customer profile who fills out this questionnaire follows the criteria determined. The criteria are yes, all respondents have shopped online, are in the age range of24 to 65 years, and live in the areas of Surabaya, Jakarta, Medan, Bodetabek, and Bandung.

3.2 Data preprocessing

Data preprocessing is conducted using data collected from the results of distributing questionnaires. The steps in data preprocessing include data cleaning, data transformation, and data normalization.

3.2.1 Data cleaning

This stage checks the data related to the number of attributes, attribute names, and attribute types and whether the data has missing values. As shown in table 6, there are two data types:integer and object types. In addition, there is a non-null word that shows the meaning that there is no empty data or missing value from all existing data.

Table 6. Data Description

Column	Non-Null Count	Dtype
City	200 non-null	Object
Cost	200 non-null	Int64
Shopping Style	200 non-null	Int64
Frequency	200 non-null	Object
Model	200 non-null	Int64
Types of Hijab	200 non-null	Object
Hijab Design	200 non-null	Object

3.2.2 Data transformation

As explained in the data cleaning step, the dataset for this study has different types, integer and object types. The integer data can be used directly for clustering using K-means. On the other hand, the object data type consists of text or categorical values (basically non-numericalvalues). Since most algorithms expect numerical values to achieve state-of-the-art results, these values should be numeric. Therefore, a data transformation stage is needed to change the original object type data into a numerical type.

3.2.3 Data normalization

The data normalization stage is a process that equates the range of values for each variable. Each data variable to be processed has a different range of values. Therefore, normalization is needed to equalize the range of values and, at the same time, simplify data processing. In this study, the normalization is done by scaling each feature to a given range between zero and one.

3.3 Clustering

The clustering process with the k-means method aims to group the data into questionnaire results that potential customers have filled out. They are then grouped based on the similarityof data in the same cluster. The implementation of this clustering uses the functions containedin the scikit-learn [13] package, namely sklearn.cluster.Kmeans.

In implementing the k-means algorithm, it is necessary to determine the number of clusters. The methods to determine the number of clusters are the Elbow Method, Davies-Bouldini Index, Silhouette Score, and Calinski-Harabasz Score.

3.3.1 The elbow method

The elbow method finds the number of clusters based on the SSE value or the inertia value. In this method, it can be stated that the optimal number of clusters is indicated by the SSE value, which has decreased significantly and forms a right angle. A significant decrease that forms like a right angle is shown when one of the SSE values and the next SSE value does not experience a significant difference in value, or it can be seen if the graph decline is morestable. But if the value is compared with the previous value, there is a significant difference.

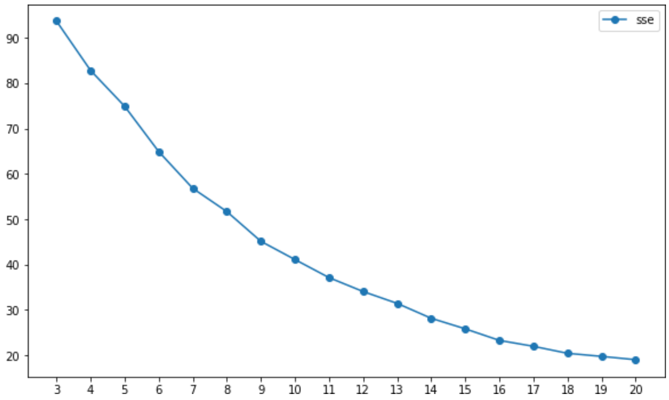

Fig. 2. The elbow method

The graph in **Figure 2** shows the number of clusters 17 experienced an increase in the graph,so the number of clusters 18 experienced a drastic and slightly steeper graph decline. Meanwhile, the number of clusters 19 and so on tends to have no significant difference, or the graph is more stable. Therefore, it can be concluded that the optimal number of clusters in this method is the number of clusters 18. Moreover, the results of this method need to be compared with the results from the other three methods.

3.3.2 Davies-Bouldini index method

One of the methods used in determining the optimal number of clusters is the davies-bouldini index (DBI) method. The smaller the value of DBI generated, it will indicate that the numberof clusters is optimal. The representation of the DBI value results can be seen in **Figure 3**. **Figure 3** indicates the smallest DBI value is cluster number 18.

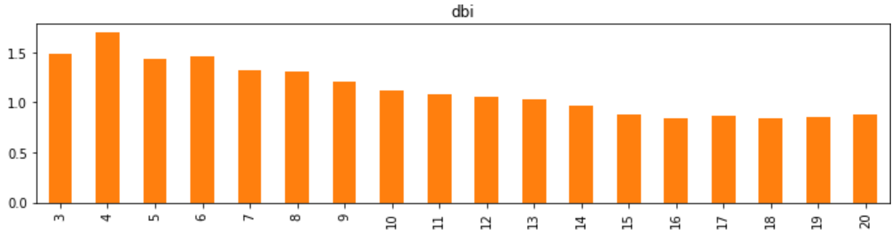

Fig. 3. Davies-Bouldin Index Value Chart

3.3.3 Silhouette score

The Silhouette score method is also used to find the optimal cluster value. The best value of silhouette score is 1, and the worst value is -1. **Figure 4** represents the Silhouette Coefficient score graph. **Figure 4** shows that the most significant Silhouette score value is the number of clusters 19, with slight friction from cluster number 18.

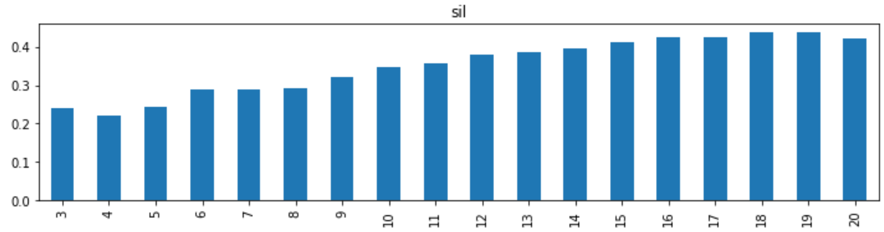

Fig. 4. Silhouette Coefficient Value Chart

3.3.4 Calinski-Harabasz score

The following method that evaluates the number of cluster values used in this study is the Calinski-Harabasz (CH) method. Where a higher Calinski-Harabasz score relates to a model with better-defined clusters, the representation of the results of the CH values can be seen in **Figure 5**, where it shows the highest estimated graph is in clusters number 18.

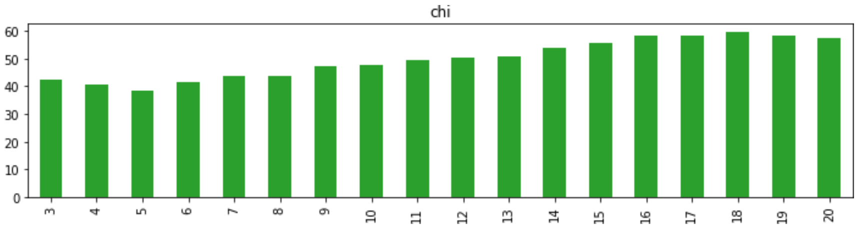

Fig. 5. Calinski-Harabaz Value Chart

Based on the Elbow Method, Davies-Bouldini Index, Silhouette Score, and Calinski- Harabasz Score, it can be concluded that the optimal number of clusters in this study is 18. Moreover, **Figures 6** and **7** illustrate the clustering result with k-values is 18. As we can see, the object values tend to accumulate a lot in the middle of each cluster. It means the relationship between one cluster with another cluster has the same level of similarity.

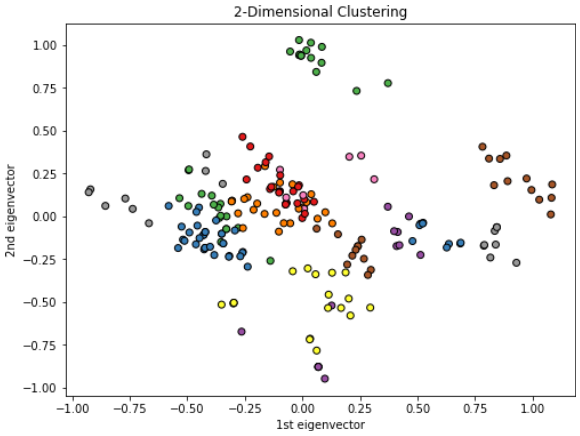

Fig. 6. Two Dimensional Plot Scatter Graph

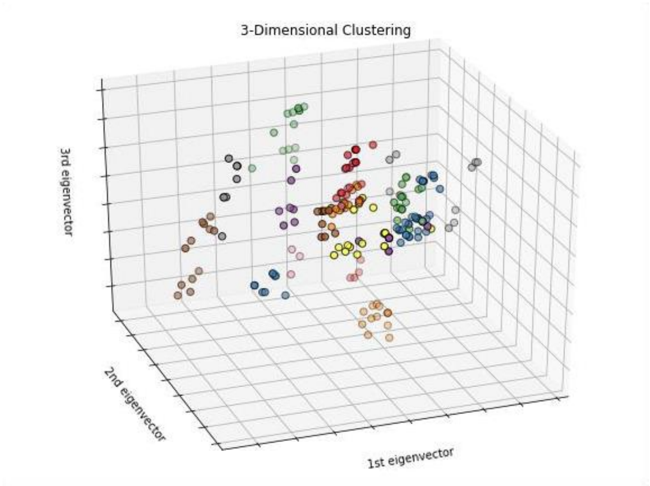

Fig. 7. Three Dimensional Plot Scatter Graph

4 Result analysis

After getting the customer cluster, each cluster is re-categorized into its customer class: superstar, golden customer, typical customer, occasional customer, and dormant customer. This categorization is based on the similarity of the frequency and cost of purchasing hijab products by each prospective customer. Table 7 recapitulates the categorization of 18 clustersinto five customer classes.

Table 7. Customer class category

Customer Class Category	Information
Superstar	cluster 17 dan cluster 7
Golden customer	cluster 0, cluster 10, cluster 14, cluster 12, cluster 5, cluster 3
Typical customer	cluster 6, cluster 9, cluster 4, cluster 11
Occasional customer	cluster 1, cluster 2, cluster 8, cluster 13
Dormant customer	cluster 16, cluster 15

This study provides recommendations for which classes have the potential for the advancement of the online shop. The strategy recommendation is based on Sumayang's [14] research on inventory strategy, where the best rule is 80-20. These rules are based on customercategories that have the potential for the company's progress. The 80% strategy is used for customer categories with high potential and provides significant profits for a company. At thesame time, the 20% inventory strategy is a category of customers with less potential and profitfor the company.

Table 8. Hijab stock strategy priority

Types of Hijab	Color Design
Hijab Segi Empat Bella Square	Earth tone, pastel dan light
Hijab Segi Empat Saudia	Earth tone, pastel dan light
Hijab Segi Empat Voal *Laser Cut*	Earth tone, pastel dan light
Hijab Segi Empat Motif bahan satin	Earth tone, pastel dan light
Hijab Pashmina Ceruty Babydoll	Earth tone, dark dan pastel
Hijab Pashmina Tali Instant	Earth tone, dark dan pastel
Hijab Pashmina Diamond	Earth tone, dark dan pastel
Hijab pashmina plisket	Earth tone, dark dan pastel
Hijab khimar	Pastel
Hijab syar'i wolfis	Pastel

Thus, from the five categories of customer class, this study recommends an 80% strategy to the top three classes: the superstar class, golden customer, and typical customer. The results show that the average respondents from the top three classes can be prioritized because of their high-frequency purchases and big-spending costs. These top three classes are interested in the category of hijab types and color designs described in Table 8.

While the other two customer classes, namely occasional customer dan dormant customer, are still maintained with an inventory strategy of only 20% of the total inventory. The occasional customer dan dormant customer like the hijab designs shown in Table 9.

Table 9. Hijab stock strategy low priority

Types of Hijab	**Color Design**
Hijab pashmina plisket	Dark
Hijab Segi Empat Bella Square	Dark
Hijab Segi Empat Saudia	Dark
Hijab Segi Empat Voal *Laser Cut*	Dark
Hijab Segi Empat Motif bahan satin	Dark
Hijab khimar	Dark dan earth tone
Hijab syar'i wolfis	Dark dan earth tone

5 Conclusion

This study aims to determine the segmentation of potential customers and the character of each group to minimize the potential of overstock in the inventory of online shops. First, we utilize the K-Means clustering algorithm to categorize customers into superstar, golden customer, typical customer, occasional customer, and dormant customer. Then, recommend the online shop to prioritize the top three customer classes: superstars, golden customers, and typical customers as the primary target and set the inventory for these top three classes at 80% of inventory level.

References

[1] Jayani DH. Tren pengguna e-commerce terus tumbuh [Internet]. Katadata; 2019 Oktober 10 [cited 2021 Februari 6]. Available from: https://databoks.katadata.co.id/datapublish/2019/10/10/tren-pengguna-e-commerce-2017-2023.

[2] Bank Sentral Republik Indonesia. Laporan tahunan 2020 [Internet]. Bank Sentral Republik Indonesia; 2021 [cited 2021 March 2021]. Available from: https://www.bi.go.id/id/publikasi/laporan/Documents/Laporan-Akuntabilitas-Bank-Indonesia-2020.pdf.

[3] Gattorna J. Dynamic supply chains: Delivering value through people. London: Pearson Financial Times; 2010.

[4] Adiana BE, Soesanti I, Permanasari AE. Analisis segmentasi pelanggan menggunakan kombinasi RFM model dan teknik clustering. Jurnal Terapan Teknologi Informasi. 2018;2(1):23-32

[5] Pramesti DF, Furqon MT, Dewi C. Implementasi metode K-Medoids Clustering untuk pengelompokan data potensi kebakaran hutan/lahan berdasarkan persebaran titik panas (hotspot). Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. 2017;1(9):723-732. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/204.

[6] Atyanto DB, ER M, Soelaiman R. Customer profiling dengan menggunakan K-Means Clustering untuk mendukung pengambilan keputusan strategis di PT. Pelabuhan Indonesia III (Persero). Prosiding Seminar Nasional Manajemen Teknologi XIV. 2011.

[7] Angelie AV. Segmentasi pelanggan menggunakan Clustering K-Means dan model RFM (Studi kasus: PT Bina Adidaya Surabaya) [Thesis]. Surabaya: Institut Teknologi Sepuluh Nopember; 2017.

[8] Kassali R. Membidik pasar Indonesia: Segmentasi, targeting, positioning. Jakarta: PT Gramedia Pustaka Utama; 1998.

[9] Tsiptsis KK, Chorianopoulos A. Data mining techniques in CRM: Inside costumer segmentation. New Jersey: John Wiley & Sons; 2011.

[10] Suyanto D. Data mining untuk klasifikasi dan klasterisasi data. Bandung: Informatika Bandung; 2017.

[11] Aghabozorgi S, Shirkhorshidi AS, Wah TY. Time-series clustering–a decade review. Information Systems. 2015;53:16-38.

[12] Umar DH. Desain Penelitian. Jakarta: PT Rajagrafindo Persada; 2010.

[13] Pedregosa F, et al. Scikit-learn: Machine learning in Phyton. Journal of machine Learning Research. 2011;12:2825-2830.

[14] Lalu S. Dasar-dasar manajemen produksi dan operasi. Jakarta: Penerbit Salemba Empat; 2003.