Eurasian Academy of Sciences
Eurasian Business & Economics Journal
2023 Volume:32 S: 47- 56
Published Online April 2023 (http://busecon.eurasianacademy.org)
https://doi.org/10.17740/eas.econ.2023-V32-04

CUSTOMER SEGMENTATION USING RFM ANALYSIS AND CLUSTERING ALGORITHMS IN EXPRESS DELIVERY INDUSTRY

Helin Öykü DEMİRCİOĞLU* Berna ŞİMŞEK Serdar DİLMEN*** Tuncay ÖZCAN******

* Corresponding Author, DHL Worldwide Express Transport and Trade Inc.,Arnavutkoy, İstanbul, helinoyku.demircioglu@dhl.com
** DHL Worldwide Express Transport and Trade Inc., Arnavutköy, İstanbul, berna.simsek@dhl.com
*** DHL Worldwide Express Transport and Trade Inc., Arnavutköy, İstanbul, serdar.dilmen@dhl.com
**** Istanbul Technical University, Macka, İstanbul, tozcan@itu.edu.tr

Received Date: 15.02.2023 Accepted Date:16.04.2023

ABSTRACT

As in every sector, one of the most important issues in the express delivery industry is to get to know the customers. Understanding customer preferences and identifying loyal and churn customers provides significant competitive advantages for companies. At this point, customer segmentation based on RFM analysis is widely used. RFM analysis segments customers using recency (R), frequency (F) and monetary (M) attributes. In this study, customer segmentation is presented with a case study with the actual data taken from an express delivery company. For this purpose, firstly, k-means and fuzzy c-means algorithms are used for customer segmentation based on RFM analysis. Then, Silhoutte and Dunn indices are calculated to evaluate the clustering quality and to determine the best number of clusters. Finally, on the basis of the obtained cluster centers, strategic analysis is carried out and churn and loyal customers are determined.

Keywords: Customer Segmentation; Clustering Algorithms; RFM Analysis; Express Delivery.
JEL Classification:

HIZLI TESLİMAT SEKTÖRÜNDE RFM ANALİZİ VE KÜMELEME ALGORİTMALARI KULLANARAK MÜŞTERİ SEGMENTASYONU

ÖZET

Her sektörde olduğu gibi ekspres kargo sektöründe de en önemli konulardan biri müşteriyi tanımaktır. Müşteri tercihlerini anlamak, sadık ve kaybedilen müşterileri belirlemek şirketler için önemli rekabet avantajları sağlar. Bu noktada, RFM analizine dayalı müşteri segmentasyonu yaygın olarak kullanılmaktadır. RFM analizi, müşterileri yenilik (R), sıklık (F) ve tutar (M) nitelikleri kullanarak gruplara ayırmaktadır. Bu çalışmada, müşteri segmentasyonu, bir hızlı teslimat şirketinden alınan gerçek verilerle bir vaka çalışması ile sunulmaktadır. Bu amaçla, ilk olarak, RFM analizine dayalı müşteri segmentasyonu için k-ortalamalar ve bulanık c-ortalamalar algoritmaları kullanılmıştır. Daha sonra, kümeleme kalitesini değerlendirmek ve en iyi küme sayısını belirlemek için Silhoutte ve Dunn indeksleri hesaplanmıştır. Son olarak, elde edilen küme merkezleri bazında stratejik analizler yapılarak kayıp ve sadık müşteriler belirlenmiştir.

Anahtar Kelimeler: Müşteri Segmentasyonu; Kümeleme Algoritmaları; RFM Analizi; Hızlı Teslimat
JEL Sınıflaması:

1. INTRODUCTION

During the last ten years, the parcel delivery, courier and express sector has been one of the fastest-growing businesses in the economy. Parcel delivery, courier services and express companies aim to deliver each and every package in the quickest and most efficient way possible. These companies are now beginning to realize just how important customer retention is. Growth in the global express markets has been driven by e-commerce business and also operations in public and private sectors, manufacturing companies, real estate companies, accounting firms, law firms, schools and universities, pharmaceutical industry. This case requires customer segmentation analysis across a wide range of customers. Especially, after the Covid-19 pandemic, with the rapid growth of e-commerce volume, customer satisfaction and the quality of the service provided to the customer have become very important. Although various customer segmentation analyzes have been carried out in many sectors before, customer segmentation analysis has become more important than ever for parcel delivery, courier service and the express sector. In the next years, the industry's biggest marketing challenge will be to control churn rates by identifying customers who are most likely to leave and then taking appropriate steps to retain them. Every time a customer leaves, it represents a significant investment loss. Both time and effort need to be channeled into replacing them. Being able to predict when a customer is likely to leave, and offer them incentives to stay, can offer huge savings to a business. Moreover, understanding what keeps customers engaged is extremely valuable knowledge, as it can help the express companies to develop their retention strategies, and to roll out customer journey practices aimed at keeping customers from walking out the door. For this reason, customer segmentation is very important in order to make this analysis and predictions. Customer segmentation analysis helps the business units identify pain points throughout the entire customer journey. Understanding those pain points then opens up avenues to improve new business models supportive products and services. Customer segmentation analysis directly affects the number of customers interacting with any given campaign, new business model and other related operation steps. Although there are many methodologies that analyze customer movements and, variables, it has been acted on that RFM analysis, which takes into account the variables of recency (R), frequency (F) and monetary (M) is a suitable and widely used model.

This study investigates determinants of customer segmentation in the parcel delivery, courier and express sectors. The recent literature on RFM analysis and clustering techniques in customer segmentation is reviewed and a discussion on the future research directions is offered. In this study, customer segmentation is illustrated with a case study with the real-life data using k-means and fuzzy c-means algorithms. Also, clustering indices such as Silhoutte and Dunn are used to to determine the best clustering method and number of clusters. According to the clustering results, Churn and loyal customer segments are determined and analyzed.

The remainder of this paper is as follows. In Section 2, a literature review is presented on customer segmentation and RFM analysis. In Section 3, clustering algorithms such as k-means and fuzzy c-means and Silhoutte and Dunn indices are described. In Section 4, an application

of the RFM analysis and clustering algorithms for the customer segmentation in express delivery industry is given. Finally, the results and conclusions are discussed.

2. LITERATURE REVIEW

Customer loyalty is provided by meeting customer expectations (Kursunluoglu, 2011). In order to ensure this loyalty, it is possible to analyze customer data well and to know the customer based on data (Handoyo, 2021). In order to ensure customer satisfaction, it is very important to know the customers and to determine the customer segments by measuring their purchasing behaviour. At this point, customer segmentation has a critical importance for companies. The integration of RFM analysis and clustering algorithms provide insights into customer behavior. There are many studies on RFM analysis and customer segmentation in the literature. These studies can be summarized as follows:

Ha (2007) proposed a hybrid RFM - decision tree method to define changes in the recency, frequency, monetary data of customers. In this study, classification rules are created to forecast the next customers' recency, frequency, monetary data. Birant (2011) applied a new three-step approach that uses RFM analysis in data mining tasks. This approach includes clustering task, classification rules and association rule mining. In the clustering task is used to find customer segments with similar RFM values. Ghoumrassi and Tigu (2018) investigated on the effect of logistics management on customer satisfaction and stated that manufacturers, marketers, suppliers, transporters, warehouses, albeit indirectly, have a great impact on meeting the needs of the customer.

There are also many studies in the literature in which the RFM model and clustering algorithms are integrated. Cheng and Chen (2009) proposed a new approach that integrates RFM features, K-means algorithm and rough set theory. In this study, the effectiveness of the developed approach demonstrated with real-life data from a company in Taiwan's electronic industry. Khajvand and Tarokh (2011) proposed a framework to predict customer lifetime value (CLV) based on weighted RFM analysis. In this study, K-means, two step and x-means algorithms are used for customer segmentation based on customer lifetime value (CLV) and the best algorithm and optimum number of clusters are determined using Dunn index. Chen et al. (2012) integrated RFM model with K-Means method and decision tree to define various meaningful groups for the customer segmentation for an online retailer. Jintana and Mori (2019) proposed a method of customer segmentation through the use of the RFM model and k-means method to cluster customers for the courier business. In this study, customers are organized into two different groups that help identify market segments more clearly and help decision-makers to develop a suitable marketing strategy. Anitha and Patil (2022) presented a model based on RFM analysis and K-Means algorithm for the e-commerce customer segmentation. In this study, Silhouette Coefficient is used to validate the proposed model. Bhupathiraju and Raghavendra (2022) applied K-Means algorithm to segment the customers of the business according to the RFM model, and the main characteristics of customers in each segment analyzed. Li et al. (2022) proposed the RFM weighted K-means method for customer

segmentation. In this method, the weight of each index is calculated with the Analytic hierarchy process (AHP)-entropy weight methods and the k-means method is used to group the customers. Tasabat et al. (2023) suggested the RFMS model for customer segmentation by adding the economic variable 'Sensitivity' to the traditional RFM model.

3. METHODOLOGY

In this section, k-means and fuzzy c-means algorithms and Dunn and Silhoutte indices used for customer segmentation are described.

3.1. K-Means

K-means is a widely used algorithm for cluster analysis and it has been used in various fields including customer segmentation and other business applications. The K-Means method divides the items into k clusters according to their properties. One of the best features of this method is that the inter-cluster distance is maximized while the inter-cluster distances are minimized. This ensures that items with similar characteristics are in a closer cluster. K centers are randomly selected, and then the remaining elements are assigned to the centers they are closest to according to the cluster centers. Then, the new cluster is calculated and the process is repeated until the cluster centers do not change (Han et al, 2022).

The calculation steps of the K-means algorithm are as follows:

Step 1: Initialize the number of clusters and randomly select k objects as center
Step 2: Calculate the distances of each object from the centers
Step 3: Assign each object to the closest cluster center
Step 4: Calculate the new centers with the average of objects in each cluster.
Step 5: Repeat Steps 2-4 until there is no change in cluster assignments.

3.2. Fuzzy C-Means

Fuzzy C-Means is an algorithm that allows different data to be included in more than one cluster. In the K-means method, a data should only be in one set, whereas in Fuzzy C-means, this situation is different and provides a better result (Christy et al, 2018).

In the fuzzy c-means algorithm the number of clusters is also known. The objective function of this algorithm is given in Equation (1) (Ansari and Riasi, 2016).

$$J = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} d_{ik}^{2} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} ||x_k - v_i||^2 \quad (1)$$

In this equation, the following notations are used:

m indicates a real number larger than 1

x_k indicates the kth object
v_i indicates the centroid of the ith cluster,
u_{ik} indicates the degree to which object k belongs to cluster i (between 0 and 1)
$||x_k - v_i||$ denotes the Euclidean distance between kth object and ith cluster center.

When all the elements of the U matrix are 0 or 1, the algorithm changes to the classical c algorithm. The sum of the degrees of belonging to the clusters of each object should be 1 as in Equation (2).

$$\sum_{i=1}^{c} u_{ik} = 1 \ , \forall k = 1, \dots, n \tag{2}$$

Using the Equation (2) and by minimizing the objective function, Equation (3)-(4) are obtained.

$$V_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m} \tag{3}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{2/(m-1)}} \tag{4}$$

3.2. Clustering Indices

In this study, Silhoutte and Dunn indices are used to measure clustering quality and determine the best clustering algorithm and number of clusters. These indices can be explained as follows.

The Silhouette index is one of the indexes to evaluate clustering performance and to determine the optimum number of clusters. In this index, the calculation is made based on the average distance within the cluster and the difference between the minimum distance between the clusters. Silhoutte index can be calculated using Equation (5) (Wang and Xu, 2019).

$$\bar{S} = \frac{1}{n} \sum_{i=0}^{n} \left(\frac{b(i) - a(i)}{max\{a(i), b(i)\}} \right) \tag{5}$$

In Equation (5), $a(i)$ denotes the average distance of object i to other objects in the cluster and $b(i)$ denotes the minimum distance of the object from the object i to the other clusters.

Dunn's index is another measure used to evaluate clustering quality. The formula, which focuses on the fact that the data sets are in well-separated sets and the Dunn index is maximum, is as in Equation (6) (Abdulhafedh, 2021).

$$D = \frac{min_{1 \le i < j \le n} d(i,j)}{max_{1 \le k \le n} d'(k)} \tag{6}$$

In Equation (6), i, j and k are indices for clusters. While d measures the inter-cluster distance, d' measures the intra-cluster difference.

4. APPLICATION

In this study, customer segmentation is performed with real-life data taken from an express delivery company. Clustering algorithms such as K-means and Fuzzy C-means algorithms were used for customer segmentation based on RFM analysis. RFM analysis includes three attributes such as Recency (R), Frequency (F) and Monetary (M) that reflect the purchasing behavior of customers. Recency denotes to the interval between the time, that the latest consuming behavior happens, and the present. Frequency is the number of transactions that a customer has made within a certain period. Monetary refers to the cumulative total of money spent by a particular customer (Birant, 2011).

The main steps of the proposed approach are as follows:

1.) Dataset is created from database tables for the Recency, Frequency, and Monetary attributes.
2.) Data transformation is applied to calculate frequency, monetary and recency.
3.) Outliers are detected with Box plot and Grubbs' test.
4.) Both linear normalization (min-max) and standardization (Z-score normalization) are used and tested.
5.) Best clustering method and best number of clusters are calculated by using internal measures such as Silhouette Index and Dunn Index. From 5 to 15 number of clusters were tested. Along with number of clusters, K-means and Fuzzy C-Means algorithms are tested.

In this study, customer transaction data for the period 01.01.2022-31.12.2022 is used for customer segmentation. Descriptive statistics of the dataset for the Recency, Frequency and Monetary attributes are presented in Table 1. Due to data privacy, all variables are shown with scaled data.

Table 1. Descriptive statistics for Recency, Frequency and Monetary in the dataset

	Recency	Frequency	Monetary
mean	0.4065	0.0038	0.0057
std	0.3950	0.0209	0.0238
min	0.0000	0.0000	0.0000
25%	0.0400	0.0001	0.0002
50%	0.2300	0.0003	0.0009
75%	0.9500	0.0015	0.0035
max	1.0000	1.0000	1.0000

The total number of customers is 24793. Density plots of recency, frequency and monetary attributes are as in Figure 1.

In this study, the clustering algorithms are coded in Python 3.9. Also, different Python libraries such as matplotlib, scikitlearn and skfuzzy are used for data analysis, data visualization and algorithm development.

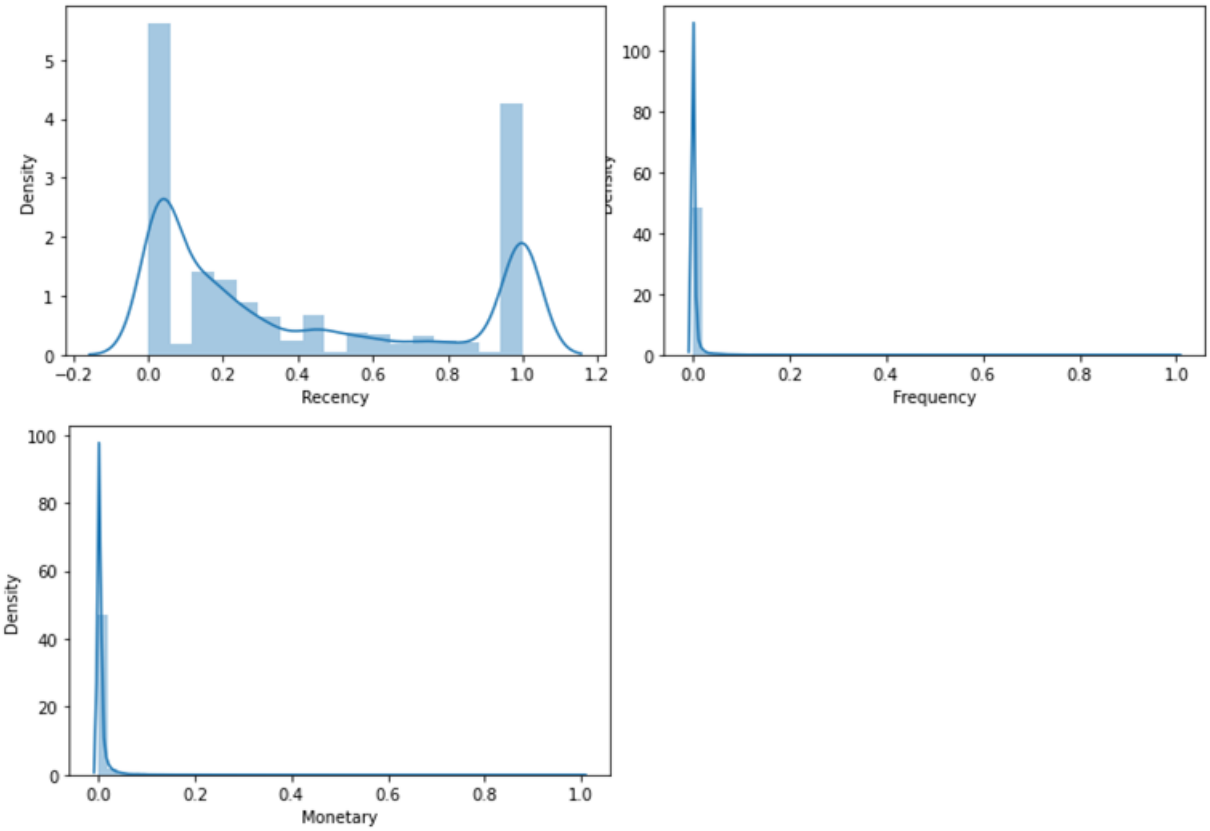

Figure 1. Density plots of recency, frequency and monetary attributes

In order to determine the best clustering algorithm and the number of clusters, k-means and fuzzy c-means algorithms are run for 5-15 cluster numbers. Clustering results are compared using Silhoutte and Dunn indices. The obtained values are presented in Table 2.

Table 2. Silhoutte and Dunn indices for different clustering algorithms

		5	6	7	8	9	10	11	12	13	14	15
K-Means	Sil.	0.2456	0.2532	0.2613	0.2622	**0.2644**	0.2534	0.2442	0.2311	0.2304	0.2289	0.2218
	Dunn	0.1116	0.1208	0.1267	0.1178	**0.1258**	0.1208	0.1187	0.1214	0.1234	0.1078	0.1065
Fuzzy C Means	Sil.	0.2446	0.2268	0.2382	0.2406	0.2431	0.2366	0.2248	0.2217	0.2188	0.2176	0.2106
	Dunn	0.0893	0.0919	0.0924	0.0927	0.0946	0.1056	0.1087	0.1073	0.1102	0.1086	0.1094

According to the values of the clustering indices, k-means is determined as the best clustering algorithm and 9 clusters is determined as the best number of clusters. Cluster centers obtained

according to these parameters are presented in Table 3. Also, cluster distributions as three dimensions are presented in Figure 2.

Table 3. Recency, frequency and monetary values for cluster centers

Cluster	Recency	Frequency	Monetary	Size
0	0.2487	0.0007	0.0020	1459
1	0.3382	0.0005	0.0014	1291
2	0.1807	0.0008	0.0022	1795
3	0.7000	0.0003	0.0010	1556
4	0.0167	0.1670	0.2043	175
5	0.1293	0.0011	0.0030	2121
6	0.9915	0.0003	0.0006	6565
7	0.4839	0.0005	0.0014	1650
8	0.0343	0.0070	0.0101	8181
Average	0.4065	0.0038	0.0057	24793

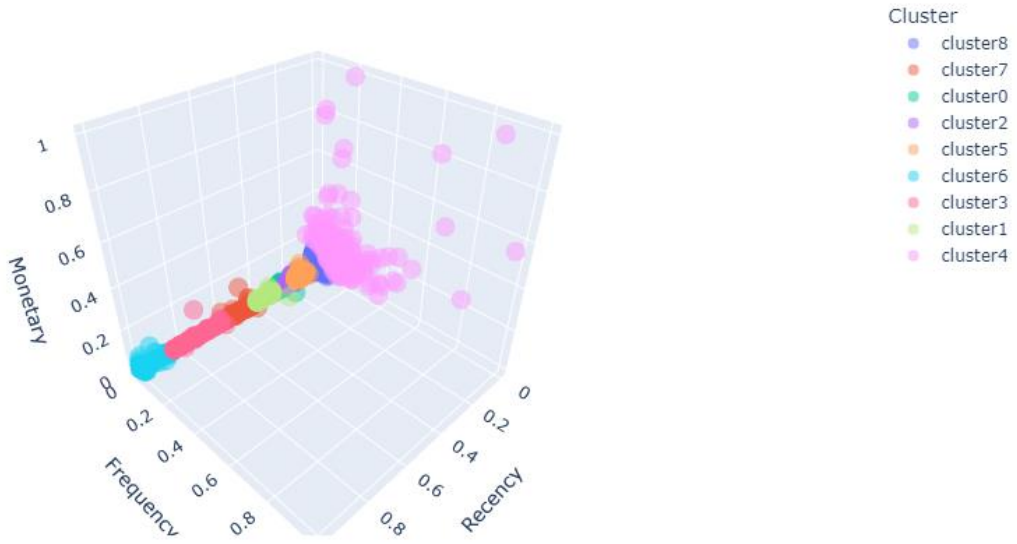

Figure 2. 3D plot of clustering results

Strategic analysis of clustering results can be performed according to Table 3 and Figure 2. For example, It can be stated that 175 customers in Cluster 4 are loyal and valuable customers. Similarly, customers in Cluster 3 and Cluster 6 are churn customers. In order to retain these customers, different opportunities and campaigns can be offered according to their purchasing preferences.

5. CONCLUSIONS

Customer segmentation and clustering are one of the keys to the success of a business. By satisfying customers, companies can increase revenue and get a wider market share. Although it has become a common topic to analyze and cluster customer data in sectors such as banking, retail, and production, this situation has just begun to develop in the express and parcel delivery sector. In express delivery companies, as in other industries, satisfied customers are loyal longer, buy more, are less price sensitive, and leave positive reviews about the company. With the improvement of service quality, it is expected that consumers will feel satisfied and consumer loyalty will be provided spontaneously.

Studies in recent years, the clustering algorithms is being integrated within the RFM concept for the definition of patterns and relationships hidden within the data. In this study, a RFM based customer segmentation model is proposed with the actual data taken from an express delivery company. At this point, k-means and fuzzy c-means algorithms are applied for customer segmentation based on RFM analysis. Silhoutte and Dunn indices are used to compare the clustering results and to determine the best number of clusters. Churn, loyal and valuable customer clusters are determined using the obtained cluster centers.

In future studies, self -organized map (SOM), hierarchic clustering and density-based cluster algorithms can be applied for customer segmentation. Also, new attributes can be added to the traditional RFM model. Tau and Davis-Bouldin indices and Elbow criteria can be used in the evaluation of clustering quality.

REFERENCES

- Abdulhafedh, A. (2021). Incorporating k-means, hierarchical clustering and pca in customer segmentation. *Journal of City and Development*, *3*(1), 12-30.
- Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 1785-1792.
- Ansari, A., & Riasi, A. (2016). Customer clustering using a combination of fuzzy c-means and genetic algorithms. *International Journal of Business and Management*, *11*(7), 59-66.
- Bhupathiraju, G. V., & Raghavendra, V. R. T. S. (2022) Data mining for the online retail industry: Customer segmentation and assessment of customers using RFM and k-means.
- Birant, D. (2011). Data mining using RFM analysis. In *Knowledge-oriented applications in data mining*. IntechOpen.
- Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*, 197-208.
- Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert systems with applications*, *36*(3), 4176-4184.

- Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking–An effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*, *33*(10), 1251-1257.
- Ghoumrassi, A., & Tigu, G. (2018). The impact of the logistics management in customer satisfaction. In *Proceedings of the International Conference on Business Excellence* (Vol. 12, No. 1, pp. 407-415).
- Ha, S. H. (2007). Applying knowledge engineering techniques to customer analysis in the service industry. *Advanced Engineering Informatics*, *21*(3), 293-301.
- Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.
- Handoyo, D. (2021). The Effect of Corporate Image, Location, Price and Service Quality on Customer Satisfaction of J&T Express Delivery Services. *Journal of Marketing and Business Research (MARK)*, *1*(1), 25-38.
- Jintana, J., & Mori, T. (2019). Customer clustering for a new method of marketing strategy support within the courier business. *Academia Book Chapter*, *31*(2), 1-19.
- Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. *Procedia Computer Science*, *3*, 1327-1332.
- Kursunluoglu, E. (2011). Customer service effects on customer satisfaction and customer loyalty: A field research in shopping centers in Izmir City-Turkey. *International Journal of business and social science*, *2*(17).
- Li, P., Wang, C., Wu, J., & Madleňák, R. (2022, January). An E-commerce Customer Segmentation Method based on RFM Weighted K-means. In *2022 International Conference on Management Engineering, Software Engineering and Service Sciences (ICMSS)* (pp. 61-68). IEEE.
- Taşabat, S. E., Özçay, T., Sertbaş, S., & Akca, E. (2023). A New RFM Model Approach: RFMS. In *Industry 4.0 and the Digital Transformation of International Business* (pp. 143-172). Singapore: Springer Nature Singapore.
- Wang, X., & Xu, Y. (2019, July). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. In *IOP Conference Series: Materials Science and Engineering* (Vol. 569, No. 5, p. 052024). IOP Publishing.