A Comparative Analysis of RFM-based Customer Segmentation with K-Means and BIRCH Clustering Techniques

Rajesh M V[1], S Rao Chintalapudi[2] and M H M Krishna Prasad[3]

[1] Pragati Engineering College, Surampalem, A.P., India
[2] CMR Technical Campus, Hyderabad, Telangana, India
[3] UCEK, JNTUK, Kakinada, A.P, India
magavenkatarajesh@gmail.com

Abstract. Marketing is an expensive activity in the realm of product sales. In today's world, most businesses have a lot of digital data that involves consumer transaction records. Segmenting clients into various prominent groupings and designing personalized activities for each cluster is a critical technique for determining such effective marketing tactics. Techniques for obtaining relevant insights from digital data have progressed significantly over time. Importantly, machine learning, a process that allows computers to gain insight and interpret data, has piqued the interest of researchers. This paper illustrates the implementation of the marketing technique called RFM model along with the k-means and BIRCH machine learning clustering algorithms on the e-commerce customers sales dataset resulting in fruitful customer segmentation. A comparative analysis is also performed which resulted in k-means outperforming the BIRCH.

Keywords: Customer Segmentation, E-Commerce, Machine Learning, K-Means, BIRCH, RFM Model.

1 Introduction

Within the world of e-commerce domains, customer segmentation is indispensable. The tactic is a prerequisite for firms since it enables them to obtain deeper insights into their consumer base and modify their marketing efforts accordingly. In the e-commerce industry, various customer segmentation methods have been extensively examined in research papers, providing valuable insights into this field.

The quick development of e-commerce in the contemporary real world has changed the way businesses work and lock in with their clients. Understanding client behavior has gotten to be basic for businesses to be competitive within the computerized commercial territories, with the growing assortment of online commerce services

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_94

and applications, thereby leading to the huge sum of information generated by online transactions.

Customer segmentation, the process of grouping customers according to similarities, has emerged as a powerful tool for organizations in order to gain a deeper insight into their customer base and customize their marketing efforts accordingly.

In the past, online purchaser division has been done physically, which is time-consuming, subjective, and has impediments when it comes to capturing the complicated linkages and designs seen in huge datasets. However, thanks to advancements in data mining and machine learning methods, organizations now have the opportunity to use artificial intelligence to augment and automate the client segmentation process.

Using the RFM model when coupled with K-means clustering is one methodology that has sparked interest in the research literature [1]. Recency, Frequency, and Monetary or RFM, is an acronym for a paradigm for segmenting clients owing to their purchase rhythms and tendencies. The RFM model and K-means clustering have been used in several research studies to segment customers in e-commerce. For instance, the RFM model is an invaluable instrument for data analytics and consumer segmentation since it is capable of analyzing crucial data like purchase amount, sales volume, and customer behavior [2].

This study attempts to explore how machine learning calculations can be utilized to fragment online customers using RFM model with K-means clustering techniques and also, RFM model with BIRCH clustering techniques. Businesses can pick up more exact and valuable insights by utilizing machine learning strategies to identify covered-up patterns and relationships in customer information.

2 Related Work

Numerous researchers provide segmentation strategies to boost revenue and maintain market leadership.

V. Asha, et. al [3], uses RFM (Recency, Frequency, and Monetary) ratings to suggest a machine learning-based prototype for customer segmentation that can assist organizations in determining consumer needs and enhancing customer satisfaction.

A. Agrawal, et. al [4], employ a hybrid technique for e-commerce buyer segmentation that fuses the Elbow method with the K-means clustering technique. They offer an effective means for companies to pinpoint unmet consumer desires and tweak their marketing plans accordingly. Their research made use of a Kaggle dataset, which increases the usefulness and efficacy of their discoveries in real-world scenarios.

Alamsyah, Alamsyah. et. al [5], attempt to carry out customer segmentation in retail businesses by combining the K-Means cluster algorithm with the RFM model, enhanced through the utilization of the Elbow method. This technique is used to enhance the K-Means algorithm performance by resolving its flaw in the initial decision of the cluster count and assisting in the selection of the optimal value for k indicating the number of clusters.

A. Patra, et. al [6], investigate the applicability of buyer segmentation in the domain of e-commerce, highlighting the value of effective data utilization and dividing customers into groups according to spending, watch time, gender, and location. The method used in this study comprises implementing one-dimensional clustering on each of the columns for recency, frequency, and monetary data independently. The next thing to do is to calculate an overall score and divide the customer pool into three groups.

Alzami F, et. al [7], implemented the RFM analysis and the clustering methodology called k-means, to categorize consumers based on dataset from Brazilian e-commerce enterprises and show the results presented through a dashboard employing the Streamlit framework. The results of this study can aid organizations in comprehending their clientele and informing choices about customer acquisition and retention methods, such as the identification of one-time buyers and the assessment of the K-Means algorithm's performance.

A. Solichin, et. al [8], employed k-means clustering utilizing the RFM model in combination with User Event Tracking (UET) parameter for the categorization of customers. A arduous method for data analysis is a necessity to discover an accurate consumer segmentation. The customer segmentation that resulted from testing on 1,447,984 records related to transactions and 932,021 data pertaining to user tracking is classified into three categories, namely Silver, Gold and Platinum.

M. Tavakoli, et. al. [9], suggested a R+FM model that adopts K-Means to cluster consumers in accordance with business developments. They utilized an approach on Digikala, the largest Middle Eastern e-commerce firm, and compared it to the company's earlier RFM model, which employed Customer Quantile Method. In addition, they developed SMS advertising tactics for each category and implemented them. The findings demonstrated that their segmentation model increased the number of purchases and baskets' average dollar value.

Israa Lewaa [10], churn prediction employing largely transactional data leveraging a blend of machine learning techniques and RFM methods. A dataset was obtained from a dataset search engine that specialized in online retail datasets. RFM ratings are produced for each consumer based on the information that is accessible. A churn measure that reflects whether or not a client has completed a transaction within a specific time frame is considered. Different approaches are compared in their research.

Aryuni, Mediana, et. al [11], developed clustering models using customer profile data from web-enabled banking of a bank. K-Means and K-Medoids clustering techniques utilize RFM scores derived from customers' Internet Banking transactions. Both approaches' performance was assessed and compared. Based on intra-cluster (A WC) distance, the K-Means technique beat the K-Medoids method. K-Means surpasses K-Medoids in terms of the Davies-Bouldin index.

S. H. Shihab, et. al [12], For RFM-based market segmentation, a comparison of agglomerative, k-means, and enhanced versions of k-means is performed. The experimental findings indicate that agglomerative clustering demands more processing time for large datasets when compared to both k-means and advanced clustering.

Alrawi, et. al [13], Random Forest classification, K-mean clustering, and Decision Tree algorithms were proposed as machine learning solutions for consumer segmentation and prediction. When compared to other categorization models, it performs better. After segmenting the dataset into three clusters, pre-processing it with data engineering approaches considerably enhances customer prediction accuracy. The suggested technique demonstrates that learning representations using K-mean for segmentation improves performance in customer groups.

Karacan Ismet, et. al [14], presented a unified CRM framework for managing both business-to-customer and business-to-business relationships. RFM Analysis categorizes customers into segments that delineate their position within the comprehensive approach. DBSCAN Cluster Analysis confirms RFM Analysis and identifies data abnormalities for further investigation, whilst the Decision Tree Classification Algorithm anticipates tailored queries specific to individual companies to help analysts reduce their operations.

3 Background Study

This theoretical terminology pertaining to this research work, customer segmentation, clustering, algorithms based on density, and so on, are briefly described here. Customer segmentation has been implemented by many researchers through the application of several techniques. Utilizing both the RFM analysis and the K-Means or BIRCH algorithms will aid in the process of grouping all kinds of customers. Data mining procedures may be used by businesses to segment and obtain critical information about customers. We may divide up customers based on their RFM scores.

3.1 RFM Model

The RFM model is one of the most well-known and commonly utilized consumer segmentation methods among data-driven marketers. It has been in use since the 1970s, notably in the field of catalogue marketing, for both assessing customer value and anticipating future consumer behavior. The temporal gap between the most recent transaction and the current transaction or the time at which the transactional data gets released, is referred to as Recency (R). Frequency refers to how frequently a client transacts at a given moment (F). Monetary refers to the amount of money actually owned by the client during the transaction; as the quantity of money increases, so does the numerical worth of M [15]. These three characteristics explain how recently a client purchased, how many orders they placed, and how much money they spent, and are a potent indicator of consumer value as well as a prediction of future behavior. RFM Analysis works on the premise of customer segmentation, categorizing clients based on their purchase habits. This powerful marketing method starts with data from your company's client database.

3.2 K-means Clustering

It is an established practice for vector quantization. The objective of this expedient is to fractionate n items into k sets or groups. A partitioning approach divides things into k partitions with respect to a dataset, D, which contains n objects, and a parameter of k ($k \leq n$), which specifies how many clusters to be formed, with each partition representing a cluster [16]. Clusters are generated with the goal to improve an ultimate partitioning criterion, which may include a dissimilarity function based on distance. Based on the features of the dataset, this procedure ensures that elements within a cluster are reminiscent of one other while remaining distinct to objects in other clusters.

The centroid of a cluster is defined by the k-means algorithm as the average numerical measure of the points inside the cluster. It goes as follows. First, it chooses k items at random from D, each of which indicates a cluster mean or center at first. The remaining objects are assigned to the group that best matches their features in accordance with the Euclidean proximity measure involving each item and the mean of the cluster.

The k-means algorithm gradually improves the homogeneity of each cluster. It accomplishes this by calculating again the mean of the cluster based on the items given to it in the previous iteration. Following that, all objects are redistributed to the new clusters using the recalculated means. This technique is repeated until the assignments stabilize, indicating that the clusters generated in the current iteration are identical to those established in the prior one.

The sum of squared errors (SSE), also referred to as scatter, can be employed as our target benchmark to assess the stature of a clustering, which is represented in the below equation (1).

$$\mathrm{SSE} = \sum_{i=1 \text{ to } n} \sum_{x \in Ci} \mathit{dist}(c_i, x)^2 \tag{1}$$

Where *dist* is the typical Euclidean spacing between a pair of objects.

3.3 BIRCH Clustering

BIRCH is a hierarchy-based clustering technique that generates a cluster tree with each node representing a cluster. It is intended to handle huge datasets and is especially efficient when the data cannot fit into memory due to the usage of a compact summary of the data in memory.

K-Means is susceptible to the initial positioning of centroids and is prone to convergence to local minima, necessitating repeated runs of the method with varied initializations for stable results. When compared to K-Means, BIRCH can effectively detect dense regions in data and is less susceptible to outliers. When clusters have uneven forms, it might struggle to perform better than K-Means.

4 Implementation

As previously stated, this work implements segmentation of e-commerce consumers utilizing the RFM analysis and K-means clustering, and implementation is carried out as follows.

4.1 Experimental Setup

Implementation is carried on Intel(R) Core(TM) i5 CPU with a 8.00 GB RAM. Jupyter Notebook 6.3.0 of Anaconda distribution is used for program development [17]. scikit-learn represents a Python programming language library dedicated to machine learning [18].

4.2 Dataset Description

An e-commerce customer sales dataset from Kaggle is harnessed in this paper for customer segmentation [19]. This dataset contains information about orders made by customers. There are 8164 rows, each row about an order for a single item by the customer, attributed with 18 columns, which include the date, order identification, customer identification, location of the customer, product name, quantity ordered, gross amount, and other characteristics of the customers.

Table 1. List of columns in the dataset

S.No.	Attribute Name / Column Name in the Dataset
1	Date
2	Order Id
3	Order Status
4	Customer Code
5	City
6	State
7	Product Name
8	Order quantity
9	Gross amount
10	Day
11	Month
12	Week
13	sku
14	Variant Name
15	Skin Tones
16	Category 2
17	Week Ending Date
18	Zone

4.3 Methodology

Customer segmentation is implemented as per the following steps.

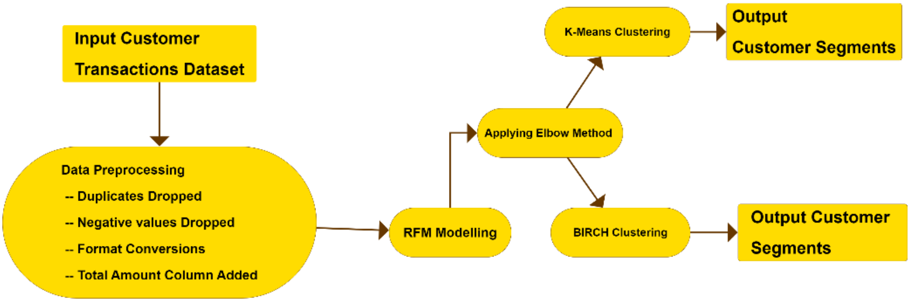

Fig. 1. Steps to perform customer segmentation

Figure 1 represents the proposed sequence of steps to perform customer segmentation with RFM Modelling and clustering techniques.

Data Preprocessing. As part of the preprocessing, the following steps are performed.

i. All duplicate entries are dropped
ii. Enumerate the Customer distribution by state
iii. Considered the data pertaining to ‘MH’ state only
iv. Validate for any negative values in Order quantity column
v. Validate for any negative values in Gross amount column
vi. Drop all the entries with negative values for the above two columns
vii. Convert the string date field to date time
viii. New column called total amount is added.

As a result of the preprocessing done, the shape of the dataset is (1553,19).

RFM Modelling. RFM Modelling scores for each customer are created as per the below-mentioned mapping.

{'Week Ending Date': 'Recency',
'Order Id': 'Frequency',
'Total Amount': 'Monetary'}

RFM scores resulted are as below

Table 2. Sample RFM Scores of few records

Customer Code	Recency	Frequency	Monetary
cust_code-102	67	5	485
cust_code-1027	46	10	1890
cust_code-1056	46	2	240
cust_code-108	67	1	80
cust_code-1081	46	2	150

The following steps are performed.

i. Using quantiles, the RFM scores are separated into four different categories.

ii. Calculate and include the R, F, and M score values as columns in the existing dataset.
iii. Compute and include an RFMGroup value column that displays the combined and concatenated RFM score.
iv. Calculate and introduce an RFMScore value column that represents the total sum of RFMGroup values.
The resulted RFM related values are as below.

Table 3. Sample RFMScores

Customer Code	Recency	Frequency	Monetary	R	F	M	RFM Group	RFMScore
cust_code-102	67	5	485	4	2	2	422	8
cust_code-1027	46	10	1890	4	1	1	411	6
cust_code-1056	46	2	240	4	4	3	443	11
cust_code-108	67	1	80	4	4	4	444	12
cust_code-1081	46	2	150	4	4	4	444	12

Table 4. Customer Loyalty Levels

Customer Code	RFMGroup	RFMScore	RFM_Loyalty_Level
cust_code-102	422	8	Gold
cust_code-1027	411	6	Platinum
cust_code-1056	443	11	Bronze
cust_code-108	444	12	Bronze
cust_code-1081	444	12	Bronze

The customer segmentation has been performed on the chosen dataset, based on RFM Model, and the customers are segmented as ‘Bronze’, ‘Silver’, ‘Gold’ and ‘Platinum’ clusters or segments.
To visualize the distribution of customers into these four segments based on the Recency and Frequency parameters, a scatter plot is plotted using ‘plotly’ library.

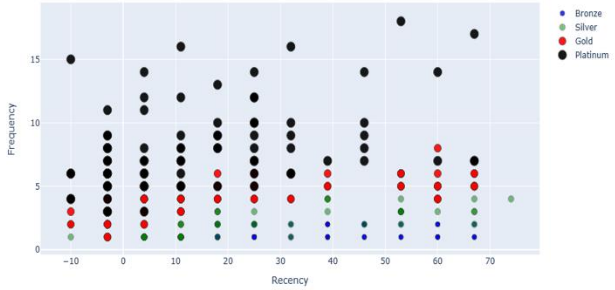

Fig. 2. Customer data clustered into four segments

Figure 2 represents the four cluster segments based on the RFM parameters indicated with blue, green, red and black dots.

Now, the clustering is performed based on the 'k-means' the machine learning technique implemented through 'scikit-learn' library.

As part of this, the optimal value for 'k' is determined to be 3 using the Elbow method.

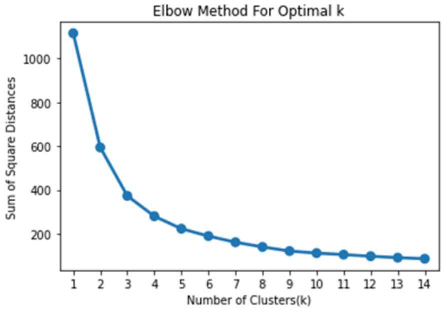

Fig. 3. Optimal value for k using Elbow Method

Figure 3 represents the correlation between sum of squared distances and the number of clusters, used to identify the optimal value for 'k' based on the Elbow method.

5 Results & Discussion

Finally based on the optimal value for k=3, while applying the K-means clustering on the given dataset of e-commerce customer sales, the final plotting between the Recency and Frequency Scores for the given data set after performing the k-means customer segmentation is depicted in the below figure 4.

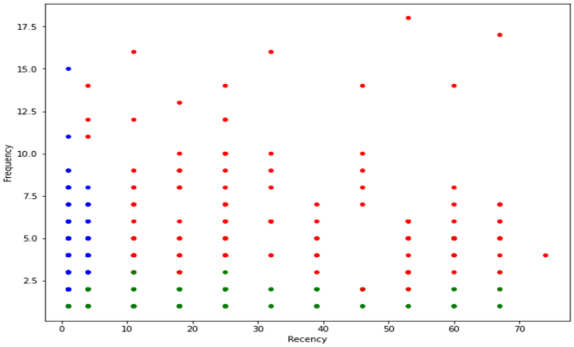

Fig. 4. K-Means Clustering resulted Customer Segments as Red, Green and Blue

Figure 4 represents the three segments resulted using the k-means clustering performed which are indicted with blue, green and red dots. It has been realized that only based on the RFM scores, the appropriate customer segmentation cannot be performed, rather in addition to the RFM model, an application of a machine learning technique like k-means can aid in proper customer segmentation. This approach can be adapted to various datasets, more specifically on the e-commerce datasets which is more temporal and also very diversified population that will be there as part of the customer base.

To measure the performance of the k-means clustering applied, the 'Silhouette Score' measure is used. Silhouette analysis is a way of interpreting and validating consistency within data clusters. The silhouette value compares the proximity a component is to its clustering (cohesion) to other clusterings (separation). The Silhouette score range is [-1, 1].

With the k-means technique for clustering, for the dataset used in this study, the Silhouette score is 0.4389.

As a comparative analysis, another clustering technique called BIRCH is also evaluated on the same dataset. So, based on the optimal value for k=3, while applying the BIRCH clustering algorithm (Balance Iterative Reducing and Clustering using Hierarchies), on the given dataset of e-commerce customer sales, the final plotting between the Recency and Frequency Scores for the given data set after performing the BIRCH clustering is depicted in the below figure 5.

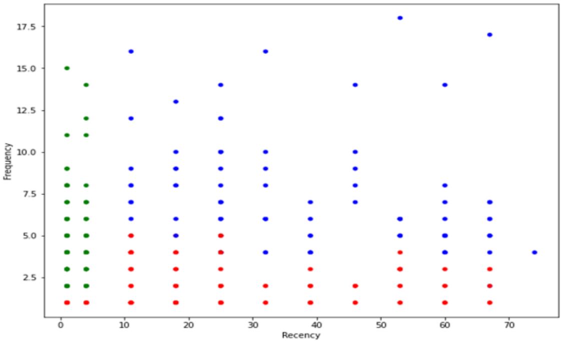

Fig. 5. BIRCH Clustering resulted Customer Segments as Red, Green and Blue

Figure 5 represents the three segments resulted using the BIRCH clustering performed which are indicated with blue, green and red dots. With the BIRCH technique for clustering, for the dataset used in this study, the Silhouette score is 0.3919. A comparison between the performance of K-means and BIRCH is shown in Fig 6.

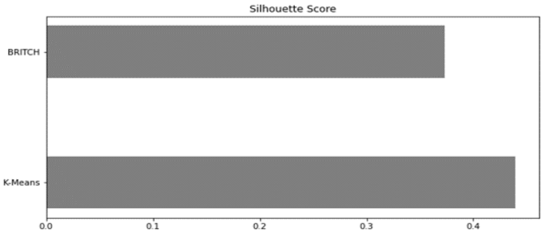

Fig. 6. Performance comparison between K-Means and BIRCH

Figure 6 represents the comparison between the cluster quality of the resulting clusters from K-means and BIRCH clustering techniques, measured using Silhouette Score.

6 Conclusion & Future Scope

For customer data clustering, RFM scores with k-means and BIRCH methods are used in this research paper. The e-commerce-related dataset has been experimented in this research work, by applying the marketing strategy of RFM model first and then the machine learning techniques like k-means and BIRCH. The implementation results illustrated that this kind of approach may be effectively utilized for client segmentation in the e-commerce domain. This research advocates that the k-means approach outperforms the BIRCH clustering approach based on the performance measure “Silhouette Score”. This research work can be extended to include more consumer behavior characteristics (e.g., clicks, comments, etc.) into the model to categorize consumers. Alternatively, other machine learning methods like k-medoids, neural network algorithms can be employed to conceive better results in terms of customer segments for large e-commerce datasets.

Hence, a research work has been carried out on the application of RFM model in conjunction with the k-means machine learning can be efficiently used rather than the application of RFM model in conjunction with BIRCH for customer segmentation in e-commerce domains.

References

1. Supangat S., Mulyani Y: Customer Loyalty Analysis Using Recency,Frequency, Monetary (RFM) And K-Means Cluster For Labuan Bajo Souvenirs in Online Store. Journal of Information Systems and Informatics (2023).
2. Lee, Z.-J., Lee, C.-Y., Chang, L.-Y., Sano, N.: Clustering and Classification Based on Distributed Automatic Feature Engineering for Customer Segmentation. Symmetry. 13, 1557 (2021).
3. Asha, V., Binju Saju, Singh Navnit Dhirendra, Yuvraj Kaswan, Prajwal G C and S. P. Sreeja: Machine Learning based prototype for Customer Segmentation using RFM. 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), IEEE (2023).

4. A. Agrawal, P. Kaur and M. Singh: Customer Segmentation Model using K-means Clustering on E-commerce. 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), IEEE, Erode, India (2023).
5. Alamsyah, Alamsyah., P., Eko, Prasetyo., Sunyoto, Sunyoto., Siti, Harnina, Bintari., Danang, Dwi, Saputro., Shohihatur, Rohman., Rizka, Nur, Pratama: Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm. Scientific Journal of Informatics (2022).
6. A. Patra, R. Khan and S. Vijayalakshmi: Customer Segmentation and Future Purchase Prediction using RFM measures. 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 753-759. IEEE, Greater Noida, India (2022)
7. Alzami, Farrikh, Fikri Diva Sambasri, Mira Nabila, Rama Aria Megantara, Ahmad Akrom, Ricardus Anggi Pramunendar, Dwi Puji Prabowo and Puri Sulistiyawati: Implementation of RFM Method and K-Means Algorithm for Customer Segmentation in E-Commerce with Streamlit. ILKOM Jurnal Ilmiah (2023).
8. A. Solichin and G. Wibowo: Customer Segmentation Based on Recency Frequency Monetary (RFM) and User Event Tracking (UET) Using K-Means Algorithm. 2022 IEEE 8th Information Technology International Seminar (ITIS), Surabaya, Indonesia, pp. 257-262. (2022).
9. M. Tavakoli, M. Molavi, V. Masoumi, M. Mobini, S. Etemad and R. Rahmani: Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: A Case Study. 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), pp. 119-126, Xi'an, China (2018).
10. Lewaa, Israa: Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis. Journal of Data Science and Intelligent Systems (2023)
11. Aryuni, Mediana, Evaristus Didik Madyatmadja, and Eka Miranda: Customer segmentation in XYZ bank using K-means and K-medoids clustering. 2018 International conference on information management and technology (ICIMTech), IEEE, Jakarta, Indonesia (2018).
12. Shihab, Sabbir Hossain, Shyla Afroge, and Sadia Zaman Mishu: RFM based market segmentation approach using advanced k-means and agglomerative clustering: a comparative study. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE (2019).
13. Alrawi, Alia Hamed, and Naim Ajlouni: Intelligent Machine Learning Customer Segmentations Algorithm. Manchester Journal of Artificial Intelligence and Applied Sciences 3.1 (2022).
14. Karacan Ismet, Inanc Erdogan, and Ufuk Cebeci: A Comprehensive Integration of RFM Analysis, Cluster Analysis, and Classification for B2B Customer Relationship Management. Proceedings of the 4th European International Conference on Industrial Engineering and Operation Management, Rome, Italy (2021).
15. Madhavi, K. Reddy, A. Vinaya Babu, A. Anand Rao, and S. V. N. Raju. "Identification of optimal cluster centroid of multi-variable functions for clustering concept-drift categorical data." In Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pp. 124-128. 2012.
16. Madhavi, K. Reddy, S. Viswanadha Raju, and J. Avanija. "Data Labeling and Concept Drift Detection using Rough Entropy For Clustering Categorical Attributes." HELIX 7, no. 5 (2017): 2077-2085.
17. Tamrakar, S., Choubey, S.B., & Choubey, A. (Eds.). (2023). Computational Intelligence in Medical Decision Making and Diagnosis: Techniques and Applications (1st ed.). CRC Press. https://doi.org/10.1201/9781003309451

18. Anaconda Distribution Homepage, https://www.anaconda.com/, last accessed 2023/09/17.
19. Scikit-learn Homepage, https://scikit-learn.org/stable/index.html, last accessed 2023/09/17.
20. Dataset Homepage, https://www.kaggle.com/datasets/roheetbatra/ecommerce-customersales-data, last accessed 2023/09/17.