Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 5, No. 3, September 2026, pp. 764–771
e-ISSN: 2964-2922, p-ISSN: 2963-6191

Open access

Original Research

Analysis of Shopee App User Behavior Patterns Using the K-Means Algorithm and RFM Model (Case Study: ZR Store)

Nafiah [a], Pamela Kareen [b*]

[a,b] Information System, Faculty of Computer Science, University of Pamulang, South Tangerang, Indonesia

*Corresponding Author: dosen03298@unpam.ac.id

ABSTRACT

The rapid growth of e-commerce in Indonesia has led to an increase in transaction volumes and user activity on online shopping platforms, such as the Shopee app. Large-scale transaction data can be leveraged to understand user behavior patterns, enabling companies to formulate more effective and targeted marketing strategies. This study aims to analyze the behavior patterns of Shopee app users using the Recency, Frequency, Monetary (RFM) method and the K-Means clustering algorithm. The RFM method is employed to assess customer characteristics based on the time of the last transaction (Recency), transaction frequency (Frequency), and total customer expenditure (Monetary). Subsequently, the K-Means algorithm is applied to segment users based on similarities in their transaction behavior. The dataset comprises 200 customers, with attributes including customer code, transaction date, age, gender, quantity of items purchased, shipping type, product category, and payment method. Using the K-Means algorithm, customers were grouped into three clusters: C1 (Loyal Customers), C2 (Less Loyal Customers), and C3 (New Customers). The results indicate that 90 customers fell into category C1, 39 into category C2, and 71 into category C3. Regarding rankings, the maximum rank ($R_{\text{Rank\_max}}$) was 50 with an obtained rank value of 12.0; for Frequency (F_{rank}), the value was 2 with a maximum rank ($F_{\text{rank\_max}}$) of 12.29; and an M_{rank} of 48.8 was obtained.

How to Cite:
Nafiah and P. Kareen, "Analysis of Shopee App User Behavior Patterns Using the K-Means Algorithm and RFM Model (Case Study: ZR Store)," *J. Kecerdasan Buatan dan Teknol. Inf.*, vol. 5, no. 3, pp. 764–771, 2026.
Doi: https://doi.org/10.69916/jkbti.v5i3.622

KEYWORDS

K-means
RFM
ZR Store
Loyal Customer
Behavioral Pattern

ARTICLE INFO

Received: August 3, 2026
Accepted: September 13, 2026
Available Online: September 18, 2026

1 INTRODUCTION

In the rapidly evolving digital era, the Shopee app has emerged as a primary platform for users to purchase products and services online. The swift growth of e-commerce app usage presents a significant opportunity for companies to gain deeper insights into user behavior, thereby enhancing the shopping experience and optimizing marketing strategies. However, managing and analyzing the data generated by Shopee app users poses a significant challenge. This data encompasses information regarding purchase transactions, user interactions with the app, product preferences, and various other aspects of user activity [1–3]. One approach to better understanding user behavior involves the application of the K-Means algorithm and the Recency, Frequency, Monetary (RFM) model within the context of data mining. The K-Means algorithm can assist in grouping users into segments based on their purchasing patterns and app interactions. Meanwhile, the RFM model provides a comprehensive overview of how often users transact (Frequency), the timing of their most recent transaction (Recency), and the value of their transactions (Monetary). To improve clarity and readability, authors are

encouraged to conclude the Introduction with a brief paragraph summarizing the research objectives, novelty, and contribution of the study [4, 5].

By analyzing user behavior patterns using the K-Means algorithm and the RFM model, the aim is to uncover valuable insights that can enhance user experience personalization through more relevant product recommendations, the development of targeted marketing strategies based on improved customer segmentation, and the optimization of pricing and promotional strategies to boost sales conversions [6–10]. Additionally, this analysis seeks to identify the factors influencing customer loyalty and facilitate the creation of effective loyalty programs. ZR Store is a fashion retailer that utilizes various social media platforms such as Facebook, Instagram, and YouTube to share information with its customers. Currently, ZR Store operates exclusively online. ZR Store was selected to provide in-depth insight into how a recommendation system can be implemented within a specific e-commerce context. By focusing on sports products, this research aims to make a practical contribution to enhancing services and the shopping experience within the e-commerce industry. In this context, recommendation systems play a crucial role in improving interactions between customers and e-commerce platforms. Traditional methods that rely solely on customer transaction history often fail to fully grasp customer preferences. Therefore, this study aims to develop a more advanced recommendation system by utilizing the K-Means clustering algorithm and the RFM model [11, 12].

Consequently, implementing a recommendation system using the K-Means clustering algorithm and the RFM model offers new advantages such as insights into behavioral and purchasing patterns, customer segmentation, personalization, inventory optimization, and comprehensive data analysis that cannot be achieved through conventional direct sales methods. The resulting recommendation system is expected to provide more personalized recommendations, boost customer retention, and give ZR Store a competitive edge in the increasingly crowded e-commerce market. Through this approach, the study aims to make a tangible contribution to the development of superior e-commerce recommendation systems.

Implementing a recommendation system using the Apriori algorithm and K-Means clustering introduces new capabilities absent in conventional direct sales methods specifically, a deeper understanding of behavioral patterns, which the RFM model. This method is used to extract association patterns from transaction data, thereby enabling an understanding of the relationships between products purchased simultaneously. For instance, it can reveal product combinations frequently bought together, allowing this information to be used to formulate more relevant recommendations [13]. By identifying items frequently purchased together, the Apriori algorithm can pinpoint product sets commonly found in a single transaction; this facilitates the offering of additional product recommendations to customers based on their past purchases. Furthermore, customer segmentation using K-Means Clustering allows for the grouping of customers based on their purchasing patterns. This enables the identification of customer groups with similar preferences and the delivery of recommendations tailored to those specific segments.

Consequently, this study titled “Analysis of Shopee App User Behavior Using the K-Means Algorithm and RFM Model (Case Study: ZR Store)” aims to bridge the gap between the data potential available to the Shopee application regarding ZR Store and its practical application in strategic decision-making to enhance business performance and user satisfaction.

2 RESEARCH METHODS

The flowchart below illustrates the research design stages for evaluating the performance of the K-Means algorithm in identifying the optimal clusters within the ZR Store sales dataset.

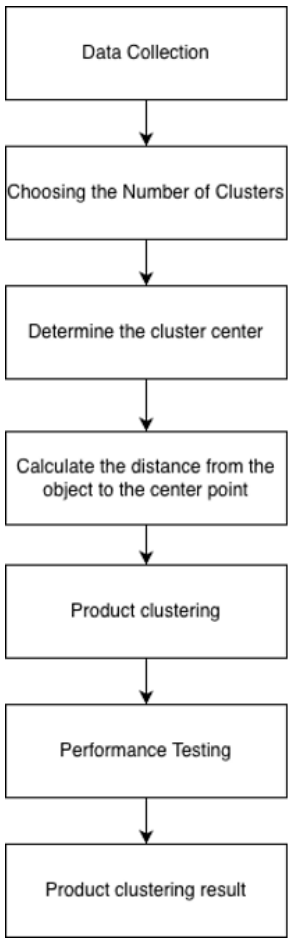

Figure 1. Research Stages

Data was collected by extracting records from the database covering a one-month period. The dataset includes codes for fashion products sold during that timeframe, along with figures for quantity, age, and unit price. The "quantity" metric represents the number of transactions made by consumers, while customer age is also recorded in the system. These metrics are used to measure the sales volume or demand for a product or service within an e-commerce context [14–16].

Selecting three clusters based on purchasing behavior is a method of grouping customers according to their sales volume. There are three general categories: loyal customers: high purchase volume, less loyal customers: moderate purchase volume (neither high nor low), and new customers: very low purchase volume, or no purchases at all.

A cluster center is a point within or near a data cluster that represents the central location of that dataset. It is frequently used in cluster analysis to identify the location or center of a data group. Determining the cluster center is typically achieved through various methods, such as calculating the arithmetic mean of the coordinates of all data points within the cluster or employing techniques like K-Means clustering to iteratively locate the center.

Calculating the distance from an object to a central point is the process of measuring the distance between an object and the center of a space or another object. In a more specific context, this can refer to measuring the distance from a particular object to a geometric center, such as the center of a circle, a sphere, or a plane. This is an important concept in mathematics, physics, and computer science, and is utilized in various applications such as mapping, navigation, and image processing.

Customer clustering is the process of grouping customers into categories based on shared characteristics or specific attributes. The objective is to understand patterns within product data and identify groups with common traits to facilitate decision-making, marketing efforts, or further analysis. By employing clustering methods, companies can identify potential market segments, gain insight into consumer preferences, and optimize their marketing and product sales strategies [16–20].

Table 1. Determination of new clusters

Determination of new clusters	Age	Quantity
New Centroid 1st	21.66667	22.8
New Centroid 2nd	46.3662	12.95775
New Centroid 3rd	27.36842	10.14035

The analyzed data yielded consistent results regarding customer purchasing volume across the following categories: loyal customers (89 records), less loyal customers (40 records), and new customers (71 records). Table 2 displays the product clustering results generated by the Orange application.

Table 2. Product Clustering Results

Row	Cluster	Silhouette	No	Kode	Age	Gender	Quantity	Shipping Type	Category	Payment
1	C3	537,711	0	0	6,042	Male	541	Express	Clothing	ShopeePay
2	C3	608,078	503	503	625	Male	1,081	Express	Clothing	COD
3	C3	589,039	1,005	1,005	1,875	Male	4,324	Free Shipping	Clothing	Transfer Bank
4	C3	595,979	1,508	1,508	1,042	Male	1,351	Next Day Air	Footwear	COD
5	C3	608,139	2,010	2,010	2,292	Male	1,081	Free Shipping	Clothing	ShopeePay
6	C2	505,232	2,513	2,513	6,250	Male	5,676	Standard	Footwear	COD
7	C2	547,618	3,015	3,015	9,792	Male	1,892	Free Shipping	Clothing	ShopeePay
8	C3	597,869	3,518	3,518	2,292	Male	2,703	Free Shipping	Clothing	Transfer Bank
9	C3	589,485	4,020	4,020	2,083	Male	1,081	Express	Outerwear	COD
10	C2	551,867	4,523	4,523	8,542	Male	1,892	2-Day Shipping	Accessories	COD

3 RESULTS AND DISCUSSION

This study focuses on analyzing the behavioral and purchasing patterns of Shopee app users at ZR Store, utilizing the K-Means algorithm and the RFM model. The research aims to provide recommendations for developing marketing strategies based on a deep understanding of these user patterns. The criteria used for the K-Means analysis to identify the optimal clusters at ZR Store are outlined in Table 3 below.

Table 3. Research Criteria

Variables	Y	X
Variables: Y, X (Age, Quantity) Context: Analyzing purchasing behavior patterns at ZR Store	C1 = Loyal Customers C2 = Less Loyal Customers C3 = New Customers	X1: Age X2: Quantity X3: UnitPrice

The dataset used consists of 200 product items. It includes attributes relevant to consumer analysis using the K-Means algorithm and the Recency, Frequency, and Monetary (RFM) model. Obtained from ZR Store, the dataset covers sales transactions over a specific period. It is essential for analysis as it provides the fundamental information needed to understand consumer behavior patterns, calculate RFM metrics, and perform clustering using the K-Means algorithm. By analyzing this data, distinct customer segments can be identified based on purchasing behavior, enabling the formulation of more effective marketing strategies.

To determine the “n” initial cluster centers, random numbers representing the input data sequence are generated. The initial cluster centers are derived from the data itself by randomly selecting points from the dataset rather than by defining new points. The values for the cluster centers are presented below.

Table 4. Cluster Centroid

Cluster	Age	Quantity	Category Description
C1	29	39	Loyal customer
C2	53	20	Less loyal customer
C3	20	2	New customer

Table 4 presents the cluster centroids: C1 (age 29, 39 shopping trips), C2 (age 53, 20 shopping trips), and C3 (age 20, 2 shopping trips) over a one-month period. Euclidean distance is used to measure the distance between data points and the cluster center. This involves an algorithm for calculating the distance between the data and the cluster center, with all calculations performed based on the Euclidean distance formula.

Testing and validation using the Orange application: The following describes the testing and validation process using the target dataset and the K-Means algorithm. The steps involved include data collection, data preparation, selection of test data, and the execution of testing and validation using the Orange application [21–24].

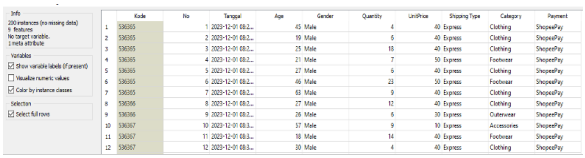

Figure 2. Reading the Sales Dataset

Figure 2 illustrates an Excel file imported into the Orange application, containing 200 instances, 9 features, and 1 attribute. The 9 features are Code, Date, Age, Gender, Quantity, UnitPrice, Shipping Type, Category, and Payment.

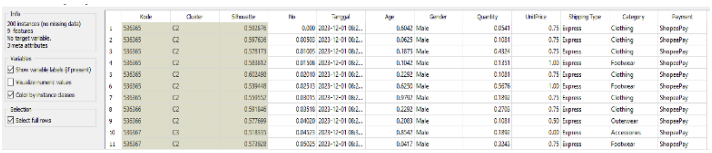

Figure 3. Clustering Results

Figure 3 illustrates the visualization of the customer clustering process based on purchase volume (Quantity). The data was grouped using the K-Means algorithm into predefined clusters: C1 (Loyal Customers) comprising 90 customers indicates a 45% tendency to shop regularly and in large quantities. C2 (Less Loyal Customers) comprising 39 customers indicates that 19.5% shop with relatively low frequency and volume. C3 (New Customers) comprising 71 customers indicates that 35.5% have likely just started shopping or make only small, infrequent purchases.

The Recency, Frequency, and Monetary (RFM) calculations in this study were performed using Python programming. The following steps describe how the RFM analysis was applied to the ZR Store sales dataset.

	Kode	Tanggal	recency	frequency	monetary
0	536365	2023-01-12 08:26:00	6	2	73
1	536366	2023-01-12 08:28:00	6	2	18
2	536367	2023-01-12 08:34:00	6	4	121
3	536368	2023-01-13 08:34:00	5	2	29
4	536369	2023-01-13 08:35:00	5	1	2
5	536370	2023-01-13 08:45:00	5	3	206
7	536372	2023-01-13 08:45:00	5	2	11
8	536373	2023-01-13 08:45:00	5	4	125
9	536374	2023-01-13 08:45:00	5	1	8
10	536375	2023-01-13 08:45:00	5	4	162

Figure 4. RFM Calculation Results

Figure 4 illustrates that, for instance, in row 0, customer code 536365 last transacted 6 days ago (Recency), has made 2 transactions (Frequency) within a one-month period, and has spent a total monetary value of 73 (Monetary).

	Kode	Tanggal	recency	frequency	monetary	r_rank_norm	f_rank_norm	m_rank_norm
0	536365	2023-01-12 08:26:00	6	2	73	12.0	48.837209	56.0
1	536366	2023-01-12 08:28:00	6	2	18	12.0	48.837209	28.0
2	536367	2023-01-12 08:34:00	6	4	121	12.0	100.000000	64.0
3	536368	2023-01-13 08:34:00	5	2	29	36.0	48.837209	44.0
4	536369	2023-01-13 08:35:00	5	1	2	36.0	18.604651	4.0

Figure 5. RFM Normalization

Figure 5 illustrates the normalization of RFM values, where each value is compared against the dataset's maximum value or calculated using a specific method. For row 0, with an original Recency rank (R_{rank}) of 6 and a maximum rank ($R_{\text{Rank\_max}}$) of 50, the resulting normalized rank value is 12.0:

$$R_{\text{rank\_norm}} = \left(\frac{6}{50}\right) \times 100 = 0.12 \times 100 = 12.0 \tag{1}$$

For Frequency (F_{rank}) of 2 and a maximum rank ($F_{\text{rank\_max}}$) of 12.29, the formula yields:

$$F_{\text{rank\_norm}} = \left(\frac{6}{12.29}\right) \times 100 \approx 48.84 \tag{2}$$

4 CONCLUSION

This study analyzed data from 200 ZR Store customers to provide recommendations for developing an effective loyalty program. Based on the research, the following conclusions were drawn: Utilizing the K-Means algorithm on existing data facilitated the research process. Calculations performed on the 200 customer records using

the Orange application employing three clusters successfully grouped customers based on their purchasing behavior: C1 = Loyal Customers (90 customers): This group demonstrates a tendency to shop regularly and in large volumes. C2 = Less Loyal Customers (39 customers): This group shops with relatively low frequency and volume. C3 = New Customers (71 customers): This group likely consists of those who have just started shopping or have only made small, infrequent purchases. Clustering results indicate that customers are divided into several segments, such as loyal, potential, regular, and inactive customers. Each segment exhibits distinct purchasing characteristics based on Recency, Frequency, and Monetary values.

Based on data analyzed using Python programming, the RFM results indicate the following: For Recency, the maximum rank ($R_{\text{Rank\_max}}$) is 50 and the obtained rank value is 12.0, for Frequency (F_{rank}), the value is 2 with a maximum rank ($F_{\text{rank\_max}}$) of 12.29, and the M_{rank} obtained is 48.8. Customers in the "Loyal Customer" category exhibit the highest transaction frequency and purchase value compared to other clusters.

The combination of the RFM model and the K-Means algorithm can serve as a decision-support tool for enhancing customer loyalty and promotional effectiveness on the Shopee marketplace. The results of this study align with previous research indicating that the combination of the RFM model and the K-Means algorithm can produce effective customer segmentation. However, the case study of ZR Store revealed that the potential customer segment represents a larger proportion than the loyal customer segment; therefore, marketing strategies need to focus on enhancing customer loyalty to boost long-term sales.

A limitation of this study is that the variables used focus solely on Recency, Frequency, and Monetary (RFM) metrics, without considering other factors such as customer reviews, product categories, or customer satisfaction levels. The clustering results are heavily influenced by the quality of the transaction data and the number of clusters specified. The study has not yet tested the impact of implementing marketing strategies derived from the customer segmentation.

DECLARATIONS

Author Contributions:

Nafiah: Conceptualization, Methodology, Software, Writing – Original Draft.
Pamela Kareen: Supervision, Methodology, Manuscript Review, Writing – Review & Editing.

Funding:

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflicts of Interest:

The authors declare no conflict of interest regarding the publication of this paper.

Data Availability:

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AI Use Statement:

During the preparation of this manuscript, the authors used Gemini solely for language refinement and readability improvement and used ChatGPT as an assistant for programming support in the Arduino IDE, including code suggestions, debugging assistance, and code optimization. All scientific content, analysis, interpretations, conclusions, and final program implementations were developed, verified, and validated by the authors. Following the use of these tools, the authors carefully reviewed and edited the manuscript and assume full responsibility for the content of the published article.

Additional Information:

No additional information is available for this paper.

REFERENCES

[1] B. Chong, "K-Means Clustering Algorithm: A Brief Review," *Acad. J. Comput. Inf. Sci.*, vol. 4, no. 5, pp. 37–40, 2021, doi: 10.25236/ajcis.2021.040506.

[2] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics (Switzerland)*, vol. 9, no. 8, pp. 1–12, 2020, doi: 10.3390/electronics9081295.

[3] H. Mukhtar et al., "K-Means Algorithm for Customer Behavior Clustering," *SEIS*, vol. 4, no. 2, pp. 96–101, 2024.

[4] R. I. Manarung, E. Widodo, and A. M. Rifai, "Sales Data Clustering Using the K-Means Algorithm to Determine Retail Product Needs," *Int. J. Softw. Eng. Comput. Sci. (IJSECS)*, vol. 5, no. 1, pp. 226–234, 2025, doi: 10.35870/ijsecs.v5i1.4090.

[5] A. Yahya and R. Kurniawan, "Implementation of K-Means Algorithm for Sales Data Grouping Based on Sales Patterns," *Malcolm*, vol. 5, pp. 350–358, 2025.

[6] A. Bucko, K. Vishi, and B. Krasniqi, "Enhancing JWT Authentication and Authorization in Web Applications Based on User Behavior History," *Computers*, vol. 12, no. 78, pp. 1–18, 2023, doi: 10.3390/computers12040078.

[7] S. Sharma and A. A. Waoo, "Customer Behavior Analysis in E-Commerce using Machine Learning Approach: A Survey," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 9, no. 2, pp. 163–170, 2023.

[8] R. Kasemrat, T. Kraiwanit, and N. Yuenyong, "Predictive Analytics in Customer Behavior: Unveiling Economic and Governance Insights Through Machine Learning," *J. Gov. Regul.*, vol. 14, no. 1, pp. 318–331, 2025, doi: 10.22495/jgrv14i1siart8.

[9] Z. Mutaqin, B. Imran, and S. Rosida, "Sistem Informasi Penjualan Online (E-Commerce) Berbasis Web Pada Toko Matahari Praya," *J. Comput. Technol.*, vol. 1, no. 2, pp. 29–36, 2023, doi: 10.69916/comtechno.v1i2.46.

[10] Sahidi and Z. Mutaqin, "Sistem Informasi Pariwisata Desa Senaru Berbasis Website Menggunakan Metode RAD," *J. Comput. Technol.*, vol. 1, no. 1, pp. 1–5, 2023.

[11] N. Hidayati, "The Use of Rapid Application Development (RAD) Model in the Development of Raw Material Inventory Information System," *Int. J. Inf. Syst. Technol.*, vol. 9, no. 158, pp. 82–92, 2025.

[12] H. T. Sadiah, M. Saad, and N. Ishlah, "Design of the Inventory Application of CV Diva Karya Mandiri Using RAD (Rapid Application Development)," *Int. J. Quant. Res. Model.*, vol. 4, no. 2, pp. 82–89, 2023.

[13] A. Fathurrozi, Hariyanto, R. W. P. Pamungkas, P. Kustanto, A. Noeman, and D. Handayani, "Analisa Perancangan Sistem Informasi Dengan Fitur Rekomendasi Menggunakan Algoritma Apriori," *Technomedia J.*, vol. 8, no. 2, pp. 51–69, 2023, doi: 10.33050/tmj.v8i2.2003.

[14] D. Kim and J. Park, "Retail Demand Forecasting Using Clustering-Based Models," *Sustainability*, vol. 14, no. 5, p. 2789, 2022, doi: 10.3390/su14052789.

[15] E. Widodo and S. Santoso, "Clustering-Based Marketing Strategy Using Fuzzy Approach," *Procedia Comput. Sci.*, vol. 216, pp. 350–357, 2023, doi: 10.1016/j.procs.2023.01.045.

[16] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artif. Intell. Rev.*, vol. 56, no. 7, pp. 6439–6475, 2023, doi: 10.1007/s10462-022-10325-y.

[17] S. Rahman and T. Ahmed, "Cluster Validation Techniques: A Review," *IEEE Access*, vol. 9, pp. 76543–76555, 2021, doi: 10.1109/ACCESS.2021.3081234.

[18] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation of Cluster Analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 2021.

[19] I. Ariati, R. N. Norsa, L. Akhsan, and J. Heikal, "Customer segmentation using K-Means clustering: A case study of Greenfields UHT milk customers," *Cerdika J. Ilm. Indones.*, vol. 3, no. 7, pp. 729–743, 2023.

[20] M. R. Razanah and P. Pandunata, "Segmentasi Pelanggan dan Penargetan Pembeli dengan Menggunakan Metode K-Means Clustering: Studi Kasus Industri Real Estate," *Melek IT Inf. Technol. J.*, vol. 11, no. 1, pp. 67–76, 2025.

[21] G. Manimannan, R. L. Priya, and C. A. Kumar, "Application of Orange Data Mining Approach of Agriculture Productivity Index Performance in Tamilnadu," *Int. J. Sci. Innov. Math. Res.*, vol. 7, no. 8, pp. 8–16, 2019, doi: 10.20431/2347-3142.0708003.

[22] S. I. Nurhafida and F. Sembiring, "Analisis Text Clustering Masyarakat Di Twitter Mengenai McDonald'S x BTS Menggunakan Orange Data Mining," *SISMATIK (Seminar Nasional Sistem Informasi dan Manajemen Informatika)*, pp. 28–35, 2021.

[23] S. Kodati and R. Vivekanandam, "Analysis of Heart Disease using Data Mining Tools Orange and Weka," *Glob. J. Comput. Sci. Technol. C Softw. Data Eng.*, vol. 18, no. 1, pp. 16–22, 2018.

[24] G. Manimannan, R. L. Priya, and C. A. Kumar, "Application of Orange Data Mining Approach of Agriculture Productivity Index Performance in Tamilnadu," *Int. J. Sci. Innov. Math. Res.*, vol. 7, no. 8, pp. 8–16, 2019.