International Journal of Engineering, Science and InformationTechnology
Volume 4, No. 4 (2024) pp. 1-12
ISSN 2775-2674 (online)
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v4i4.558
Research Paper, Short Communication, Review, Technical Paper

Grouping Sales Levels Smartphone of Offline Store Using BIRCH Clustering Algorithm

Putri Dwi Rahmadani Sari*, Mukti Qamal, Lidya Rosnita

Department of Informatics, Faculty of Engineering, Universitas Malikussaleh, Aceh, Indonesia

**Corresponding author Email: putri.200170095@mhs.unimal.ac.id*

The manuscript was received on 1 March 2024, revised on 15 April 2024, and accepted on 10 September 2024, date of publication 28 September 2024

Abstract

From 2020 to 2024, TM_Store and Jaya Com exhibited different sales patterns based on cluster analysis using the BIRCH algorithm. The background of this research is to provide strategic insights to both stores to improve their sales performance through data analysis. The sales data includes brand, type, month, year, stock quantity, quantity sold, unit price, and total sales. The BIRCH method was chosen for its effectiveness in handling large datasets and providing accurate clustering results. The clustering results indicate a significant increase in the "Moderate" category, from 12 sales in 2020 to 354 in 2023. Meanwhile, the "Very High" category also increased from 5 sales in 2020 to 97 in 2023, with sales in the "Very Low" category remaining high at 70 in 2023. On the other hand, Jaya Com was dominated by the "Very High" category, sharply increasing from 25 sales in 2020 to 597 in 2023. The "High" category also showed significant growth, from 6 sales in 2020 to 98 sales in 2023. This data indicates that Jaya Com focuses on high-performance products, while TM_Store shows a more balanced distribution across various sales categories. Based on the analysis, Jaya Com had 1988 data points with 1984 cluster points, whereas TM_Store had 2012 data points with 1811 cluster points. Overall, the study concludes that the BIRCH algorithm can identify significant sales patterns in both stores, aiding in developing more effective and efficient promotional strategies tailored to each sales category's performance.

Keywords: *BIRCH, Smartphone, Sales, Shop, Clustering.*

1. Introduction

Technology and communication currently play a crucial role in daily life. One of the most evident impacts is the increased use of smartphones, which has spurred competition among various brands. Numerous brands and variants of smartphones are sold in different stores with varying sales levels, making the smartphone industry a significant and highly profitable business sector. As a result, competition in the smartphone market has intensified over time [1]. According to an analysis conducted by Counterpoint Research, global smartphone sales experienced an 8% decline in the third quarter of 2023 compared to the same period in the previous year. This decline is widespread across all smartphone manufacturers. The drop in sales volume occurred due to a slower-than-expected recovery in consumer demand [2]. Offline stores, as one of the smartphone providers, have become a primary purchasing alternative for consumers.

Jaya Com and TM_Store are businesses engaged in smartphone sales. Jaya Com has been operating since 2010 in Stabat City, Langkat Regency. In 2017, Jaya Com received the "The Best Royal Dealer" award for achieving the highest sales volume over three months, totaling approximately 4,000 smartphone sales. Meanwhile, TM_Store has operated since 2020 and has two branches in Stabat City, Langkat Regency. Both Jaya Com and TM_Store are among the most sought-after stores by consumers. In addition to good service, their prices are also affordable. However, they often face a common problem: excess stock for some models while experiencing shortages for others, indicating an imbalance in demand forecasting. This results in overstocking and accumulating unsold items in the warehouse, leading to losses.

By applying the BIRCH clustering algorithm to sales data, offline stores can categorize smartphones based on their sales levels. This allows for the determination of appropriate marketing strategies for each product cluster. Consequently, smartphone sales at Jaya Com and TM_Store are hoped to be maximized through strategies tailored to product sales patterns. The BIRCH algorithm (Balanced Iterative Reducing and Clustering using Hierarchies) is a clustering method introduced by Zhang, Ramakrishnan, and Livny in 1996. This algorithm is designed to handle large datasets that cannot fit into memory, making it suitable for big data processing applications. BIRCH employs a

hierarchical approach to clustering by building a tree structure called the Clustering Feature Tree (CF Tree). The CF Tree stores concise information about clusters, such as cluster centroids, the Number of objects, and sizes. This way, BIRCH can significantly reduce the amount of data that needs to be processed, enhancing memory efficiency and processing speed [3].
In previous research conducted by Wynecia Liwindi in 2022, titled "Implementation of DBSCAN and BIRCH Methods for Classifying COVID-19 Spread Zones in North Sumatra," the classification of spread zones was necessary to enhance preventive measures against COVID-19. The DBSCAN method eliminated noise data, while BIRCH categorized data into four zones: green, yellow, orange, and red. Testing results showed that the model achieved a precision of 100%, with recall and accuracy at 51.52% and an error rate of 48.48% [4].
Another study by Fanny Ramadhani, Muhammad Zarlis, and Saib Suwilo in 2020, titled "Enhancing the BIRCH Algorithm for Big Data Clustering," introduced improvements to the BIRCH algorithm for Big Data clustering by replacing static thresholds with dynamic values. This modification enhanced cluster quality, measured using the silhouette coefficient (SC). A comparison between the standard BIRCH algorithm and the modified version showed that the modified algorithm resulted in a 60% reduction in CF Nodes, total CF Entries, and CF Leaf Entries, indicating improved efficiency and clustering quality in Big Data management [5].
Based on the above background, the author intends to cluster smartphone sales levels in offline stores using the BIRCH clustering algorithm to determine categories for high, high, moderate, low, and shallow sales. Through this clustering, Jaya Com and TM_Store can identify which smartphone products are the best sellers and which are not. This will enable Jaya Com and TM_Store to implement more focused promotional strategies to maximize sales revenue in the future.

2. Literature Review

2.1. Previous Research

In a study conducted by Iftah Nur Fadlilah in 2022, titled "Data Clustering of Accident Tweets Using Text Mining Approach and BIRCH Algorithm," the research aimed to cluster tweets about accidents using the BIRCH algorithm after processing through text mining. Utilizing data from Kaggle and results from Twitter crawling, the BIRCH algorithm produced efficient clustering in a single data scan. The analysis showed that data from Kaggle resulted in 1,545 clusters with a silhouette coefficient of 0.116, while Twitter data yielded 487 clusters with a higher silhouette coefficient of 0.726. These findings indicate that the BIRCH algorithm effectively clusters accident data, with better cluster quality observed in the Twitter crawling data than Kaggle data [6].
In research by Zhaojin Yan, Guanghao Yang, Rong He, Hui Yang, Hui Ci, and Ran Wang in 2023, titled "Clustering Ship Trajectories Based on Trajectory Sampling and Enhanced BIRCH Algorithm," the study proposed a method for clustering ship trajectories using an enhanced BIRCH algorithm to analyze large and complex AIS data. Using 764,393 trajectory points from 13,845 ships in the Taiwan Strait, this method generated 832 ship trajectories clustered into 172 classes. Experimental results indicated that this method efficiently detected differences in trajectories with similar spatial characteristics and produced more accurate clustering than existing methods. Furthermore, the research established more directed and content-rich main navigation routes between ports, providing crucial support for ship route planning and maritime traffic management [7].
In a study by M. Aldo Shauma, Yudha Purwanto, and Astri Niovianty in 2019, titled "Traffic Anomaly Detection Using BIRCH and DBSCAN Algorithms on Streaming Traffic," the rapid growth of the internet has posed risks of attacks on networks, such as traffic anomalies marked by flashcrowds. This research developed a traffic anomaly detection system using BIRCH and DBSCAN algorithms, effectively classifying normal and abnormal traffic in streaming data. The results showed that the combination of these two algorithms was effective, achieving an average accuracy of 98.45% with a processing time of approximately 600 seconds for 30,000 data points. These findings highlight the importance of real-time detection for enhancing network security [8].
In a study conducted by Ahmad Alzu'bi and Maysarah Barham in 2022, titled "Automatic BIRCH Threshold with Feature Transformation for Breast Cancer Clustering," the research proposed an updated BIRCH algorithm for breast cancer clustering, addressing challenges in clustering patient records and selecting optimal thresholds. This algorithm enhanced the tree-based sub-clustering procedures by utilizing breast cancer screening features and automating the threshold initialization. Evaluation using the Wisconsin breast cancer benchmark dataset demonstrated a clustering accuracy of 97.7% in 0.0004 seconds, affirming the efficiency of this method in patient record clustering and rapid decision-making [9].
In a study by Martin C Nwadiugwu in 2020, titled "Gene-Based Clustering Algorithms: A Comparison Between Denclue, Fuzzy-C, and BIRCH," the research aimed to compare three clustering algorithms applicable in gene-based bioinformatics research to understand disease networks, protein-protein interaction networks, and gene expression data. The three selected gene-based clustering algorithms were Denclue, Fuzzy-C, and the BIRCH Hierarchical Balanced Clustering. These algorithms were explored concerning subfields of bioinformatics that analyze omics data, including but not limited to genomics, proteomics, metagenomics, transcriptomics, and metabolomics data. The results from the review showed that, unlike Denclue and Fuzzy-C, which were more efficient in handling noisy data, BIRCH could effectively manage datasets with outliers and exhibited better time complexity [10].

2.2. Sales

According to Isnandi and Wardati (2014), "sales" refers to the operational process to achieve organizational goals. It involves efforts to predict customer needs and manage the flow of goods and services to meet market demand. In this way, sales aim to connect producers with consumers efficiently and effectively while aligning supply with demand to achieve the expected objectives of society [11].
Sales are a crucial component of a company's business activities, serving to meet the demand for specific products. The goal is to attract customers to generate mutually beneficial transactions. The sales transaction process involves agreements on products, prices, and services. It is influenced by the sales location and the methods and techniques used, which intertwine between sellers and buyers in the market [12].

2.3. Data Mining

Data mining is the process of exploring significant information from data sets. It also involves examining patterns within the data. These models are obtained from various databases, such as relational databases, data warehouses, transaction data, and object data. Data mining can help businesses make decisions quickly and accurately [13]. Data mining can also be used to explore databases to identify patterns and rules and find helpful information for classifying new customers and determining electricity usage categories.

Data mining, often referred to as knowledge discovery in databases (KDD), is a process that involves collecting and analyzing historical data to identify patterns, regularities, or relationships within large datasets. Findings from this process can enhance the quality of decision-making in the future [14].

2.4 Clustering

Clustering is a technique in data mining that serves to group data into specific categories. Clustering algorithms are responsible for grouping data into clusters that share similarities. In the clustering process, determining or defining the quantitative Value of the degree of similarity or difference between data (proximity measure) plays a crucial role. Therefore, it is essential to compare various commonly used methods, such as Euclidean distance, Manhattan distance, and Minkowski distance [15].

Cluster analysis is the process of dividing data in a set into several groups, where the data within each group have a higher degree of similarity compared to data from other groups. In cluster analysis, the methods divide data into subsets based on predefined similarities or likenesses [16].

2.5. Algorithm BIRCH

The BIRCH algorithm (Balanced Iterative Reducing and Clustering using Hierarchies) is a clustering method designed to manage efficiently and cluster large-scale datasets, particularly in the context of big data. It was introduced in 1996 by Tian Zhang, Raghu Ramakrishnan, and Miron Livny. BIRCH is categorized as a hierarchical clustering algorithm. It utilizes a CF Tree (Clustering Feature Tree) structure to store summarized information about clusters, including the Number of data points and central statistics. One of the advantages of the BIRCH algorithm is its ability to find good clusters using just a single scan of the data. Additionally, BIRCH can improve cluster quality with additional scans [17].

The BIRCH algorithm is designed to cluster large amounts of data in a scalable, efficient, and memory-efficient manner. It can handle large datasets by leveraging hierarchical data structures. This integrated hierarchical clustering method is specifically tailored for managing big data. Within the BIRCH algorithm, two main concepts exist: Clustering Feature and Clustering Feature Tree (CF-Tree), which function to describe the summary of clusters [18].

BIRCH Algorithm Process

The BIRCH algorithm processes data through four main stages. The first stage involves storing data in memory by building a CF tree (Clustering Feature Tree). In this phase, the initial step is to construct the CF tree from the clustered data, where the resulting tree structure has a uniform height. During this phase, the BIRCH algorithm introduces two key concepts: Clustering Feature (CF) and Clustering Feature Tree (CF-Tree), which present summaries of the formed clusters.

Clustering Feature

The Clustering Feature represents information that encompasses subclusters of data objects. It is a set of summary statistics that define a set of data points within a cluster. Using Clustering Feature to summarize clusters, there is no need to store detailed information about every data object. This algorithm requires constant storage space for each built Clustering Feature, which is why BIRCH is considered highly effective and efficient in space utilization. The formula for the Clustering Feature is as follows [19]:

$$CF = (N, LS, SS) \quad (1)$$

$$LS = \sum_{i=1}^{n} xi \quad (2)$$

$$SS = \sum_{i=1}^{n} xi^2 \quad (3)$$

Description:

N = Number of data points in the cluster
LS = Linear Sum of all data points in the cluster
SS = Squared Sum of all data points in the cluster
n = Number of data points in the cluster (equal to N)
i = Index of each data point, ranging from 1 to n
xi = Value of the i-th data point in the cluster.

Calculating Euclidean Distance

The following is an explanation of the process for calculating Euclidean Distance. The Euclidean Distance between two data points, A and B, in an n-dimensional space is calculated using the following formula:

$$d(A,B) = \sqrt{\sum_{i=1}^{n} (B_i - A_i)^2} \quad (4)$$

Description:

d(A, B) = Euclidean Distance between points A and B.
A_i = The i-th coordinate of point A.
B_i = The i-th coordinate of point B.
n = Number of dimensions (features) of the data.

With its Radius:

$$R = \sqrt{\frac{SS - LS^2}{n}} \quad (5)$$

Description:
SS = Squared Sum of all data points in the cluster.
LS = Linear Sum of all data points in the cluster.

3. Research Methods

3. 1. Place and Time

This study was conducted at offline stores, namely Jaya Com and TM_Store, located in Stabat City, Langkat Regency. The data variables used include Brand, Type, Month, Year, Stock Quantity, Units Sold, Unit Price, and Total Sales. The Number of clusters to be analyzed in this research is five, categorized as very high, high, moderate, low, and very low. The study will commence in December 2023 and continue until completion.

3. 2. Research Steps

The steps of the research are as follows:

1. Observation
   The observation technique involves the researcher visiting the research sites to directly observe and collect sales data for smartphones obtained from Jaya Com and TM_Store.
2. Interviews
   This stage includes conducting direct question-and-answer sessions with critical informants, precisely Jaya Com and TM_Store owners, to gather the necessary data.
3. Literature Study
   This involves collecting information from various sources, including books, journals, articles, websites, and other supporting documents related to the research. This literature will serve as a theoretical reference to aid the research.

3.3. System Schematic

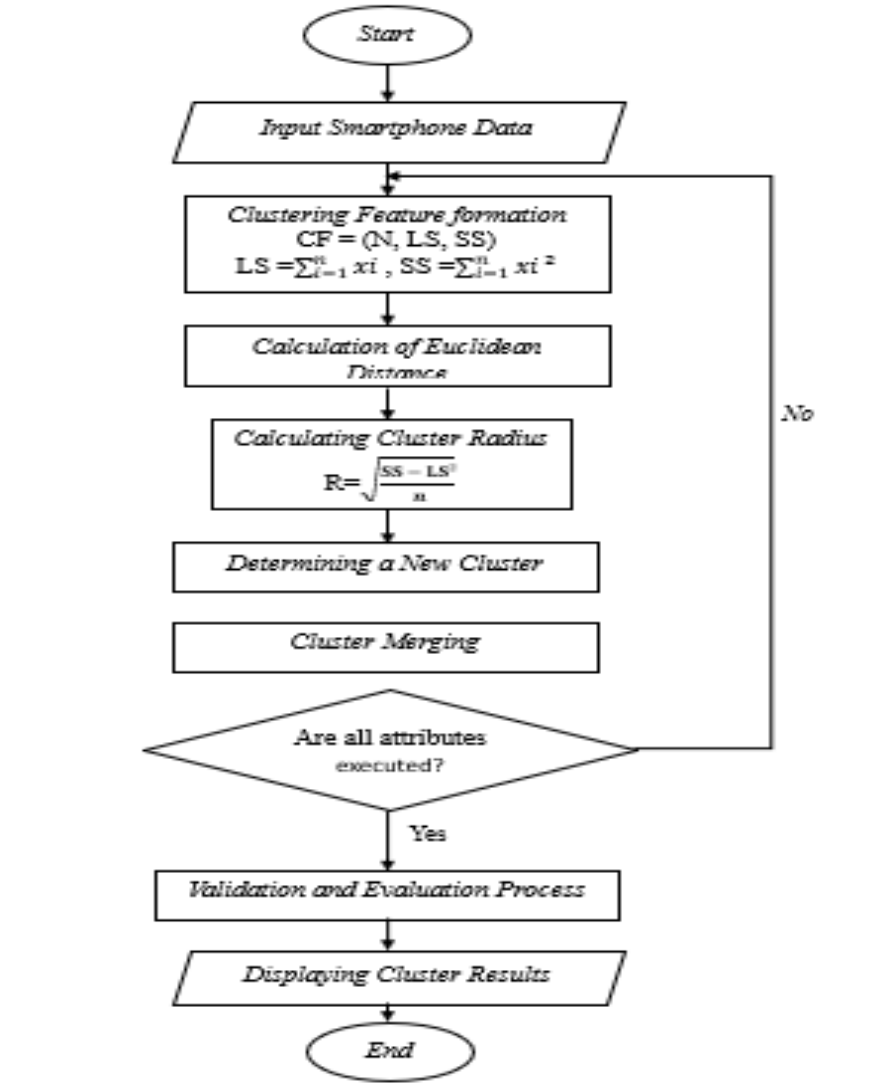

Fig 1. System Schematic

4. Results and Discussion

4.1. Research Results

Below are the results of the design model analysis, database analysis, and the implementation process of BIRCH. This analysis encompasses various critical aspects for developing a smartphone data management system, including database structure, relationships between entities, and the clustering methods used to group data based on specific characteristics.

4.2. Results of System Requirements Identification

Functional Requirements

The system to be developed must meet several functional requirements to manage sales data and smartphone specifications effectively. Functionally, the system should be able to connect to a MySQL database to access the necessary data. Users will enter connection details, and the system will process requests to ensure a successful connection. Once connected, the system must retrieve sales data and specifications from the database, execute SQL queries, and display the results in a tabular format for users. Additionally, the system should be capable of creating and displaying data visualizations in the form of scatter plots based on user requests and performing data clustering using the BIRCH algorithm to group data based on specific characteristics. The clustering results should then be saved and presented to the user.

Non-Functional Requirements

For non-functional requirements, the system must ensure the security of data stored and transferred between system components through user authentication and data encryption. The system should also be responsive, with minimal response times, achieved through optimized SQL queries and clustering algorithms.

4.3. Analysis and Discussion of Research Results

1. Dataset

   The dataset used in this research comprises a collection of data points tested with the BIRCH clustering algorithm. Four 4,000 entries were utilized, with 1,988 coming from Jaya Com and 2,012 from TM_Store. The following table summarizes the dataset for this study:

Table 1. Jaya Com Store Dataset

No	Brand	Type	Month	Year	Stock Quantity	Units Sold	Unit Price	Total Sales
1	Realme	8 Pro	1	2021	310	234	4,999,000	1,169,766,000
2	Realme	8 Pro	2	2021	481	444	4,999,000	2,219,556,000
3	Realme	8 Pro	3	2021	25	23	4,999,000	114,977,000
4	Realme	8 Pro	4	2021	255	243	4,999,000	1,214,757,000
5	Realme	8 Pro	5	2021	562	486	4,999,000	2,429,514,000
...	...	...	...	...	...	...	...	...
1984	Vivo	Vivo Y36	8	2024	345	340	3,299,000	1,121,660,000
1985	Vivo	Vivo Y36	9	2024	168	129	3,299,000	425,571,000
1986	Vivo	Vivo Y36	10	2024	437	391	3,299,000	1,289,909,000
1987	Vivo	Vivo Y36	11	2024	144	56	3,299,000	184,744,000
1988	Vivo	Vivo Y36	12	2024	517	438	3,299,000	1,444,962,000

Table 2. TM_Store Dataset

No	Brand	Type	Month	Year	Stock Quantity	Units Sold	Unit Price	Total Sales
1	Apple	iPhone 13	1	2021	18	3	14,999,000	44,997,000
2	Apple	iPhone 13	2	2021	19	10	14,999,000	149,990,000
3	Apple	iPhone 13	3	2021	21	4	14,999,000	59,996,000
4	Apple	iPhone 13	4	2021	23	6	14,999,000	89,994,000
5	Apple	iPhone 13	5	2021	14	2	14,999,000	29,998,000
...	...	...	...	...	...	...	...	...
2008	Samsung	Galaxy S21	8	2022	17	7	12,999,000	90,993,000
2009	Samsung	Redmi Note 10	9	2022	11	3	12,999,000	38,997,000
2010	Samsung	Redmi Note 10	10	2022	15	11	12,999,000	142,989,000
2011	Samsung	Redmi Note 10	11	2022	18	4	12,999,000	51,996,000
2012	Samsung	Redmi Note 10	12	2022	13	10	12,999,000	129,990,000

2. Test Data

   Determining the initial cluster centers is defined as test data taken randomly. The data used for testing BIRCH clustering can be seen in Table 3.

Table 3. Test Data

No	Brand	Type	Month	Year	Stock Amount	Units Sold	Unit Price	Total Sales
1	Realme	8 Pro	2	2021	481	444	4,999,000	2,219,556,000
2	Samsung	Galaxy S21	11	2021	194	121	12,999,000	1,572,879,000
3	Infinix	12 Play NFC	5	2023	143	70	1,899,000	132,930,000
4	Huawei	P40 Pro	10	2021	21	3	1,399,000	41,997,000
5	Infinix	12 Play NFC	11	2022	20	1	1,899,000	1,899,000

4.4. Implementation of BIRCH for Smartphone Sales Clustering

At this stage, calculations will be performed using the BIRCH clustering algorithm to group smartphone sales. The data used is from datasets in Tables 1 and 2, which include information on Brand, Type, Month, Year, Stock Amount, Units Sold, Unit Price, and Total Sales.

1. Calculation of Sales Ratio Percentage and Price Ratio with Total Sales Amount

   Below, we will calculate the sales and revenue ratios based on smartphone prices. Each sales percentage will be categorized into several categories based on specific standards. This categorization aims to provide a clearer picture of the sales performance of each smartphone type. The sales percentage is calculated based on the Number of smartphones sold compared to the available stock. Once the sales percentage is calculated, it will be categorized using the following standards:

   Very Low: If the sales percentage is less than 10%.
   Low: If the sales percentage is between 10% and less than 20%.
   Moderate: If the sales percentage is between 20% and less than 50%.
   High: If the sales percentage is between 50% and less than 70%.
   Very High: If the sales percentage is 70% or more.

   First Data (Realme, Feb 2021)
   Sales Percentage:
   "Percentage Sold" = "Units Sold" / "Stock Amount" = 444 / 481 = 0.92
   Price Ratio to Total Sales:
   "Price Ratio" = ("Total Sales" (Rp)) / ("Unit Price" (Rp)) = 2,219,556,000 / 4,999,000 = 444

   Second Data (Samsung Galaxy S21, Nov 2021)
   Sales Percentage:
   "Percentage Sold" = "Units Sold" / "Stock Amount" = 121 / 194 = 0.62
   Price Ratio to Total Sales:
   "Price Ratio" = ("Total Sales" (Rp)) / ("Unit Price" (Rp)) = 1,572,879,000 / 12,999,000 = 121

   Third Data (Infinix 12 Play NFC, May 2023)
   Sales Percentage:
   "Percentage Sold" = "Units Sold" / "Stock Amount" = 70 / 143 = 0.48
   Price Ratio to Total Sales:
   "Price Ratio" = ("Total Sales" (Rp)) / ("Unit Price" (Rp)) = 132,930,000 / 1,899,000 = 70

   Fourth Data (Huawei, Oct 2021)
   Sales Percentage:
   "Percentage Sold" = "Units Sold" / "Stock Amount" = 3 / 21 = 0.14
   Price Ratio to Total Sales:
   "Price Ratio" = ("Total Sales" (Rp)) / ("Unit Price" (Rp)) = 41,997,000 / 13,999,000 = 3

   Fifth Data (Infinix 12 Play NFC, Nov 2022)
   Sales Percentage:
   "Percentage Sold" = "Units Sold" / "Stock Amount" = 1 / 20 = 0.05
   Price Ratio to Total Sales:
   "Price Ratio" = ("Total Sales" (Rp)) / ("Unit Price" (Rp)) = 1,899,000 / 1,899,000 = 0

   Below is a summary table of the results for the sales percentage and price ratio calculations based on the provided data:

Table 4. Calculation Results Summary

No	Brand	Type	Month	Year	Percentage Sold	Price Ratio
1	Realme	8 Pro	Feb	2021	0.92	444
2	Samsung	Galaxy S21	Nov	2021	0.62	121
3	Infinix	12 Play NFC	May	2023	0.48	70
4	Huawei	P40 Pro	Oct	2021	0.14	3
5	Infinix	12 Play NFC	Nov	2022	0.05	0

2. Clustering Feature (CF) Formation
   For each data point, we need to calculate the Number of points (N), linear sum (LS), and square sum (SS). Below are the detailed calculations for each data point.
   Data Point 1 (Realme, Feb 2021)
   N = 1
   LS = [0.92, 444]
   SS = [0.92^2, 444^2] = [0.8464, 197136]
   Data Point 2 (Samsung Galaxy S21, Nov 2021)
   N = 1
   LS = [0.62, 121]
   SS = [0.62^2, 121^2] = [0.3844, 14641]
   Data Point 3 (Infinix 12 Play NFC, May 2023)
   N = 1
   LS = [0.48, 70]
   SS = [0.48^2, 70^2] = [0.2304, 4900]
   Data Point 4 (Huawei P40 Pro, Oct 2021)
   N = 1
   LS = [0.14, 3]
   SS = [0.14^2, 3^2] = [0.0196, 9]
   Data Point 5 (Infinix 12 Play NFC, Nov 2022)
   N = 1
   LS = [0.05, 0]
   SS = [0.05^2, 0^2] = [0.0025, 0]

Table 5. Clustering Feature Results

No	Brand	Type	Month	Year	N	LS	SS
1	Realme	8 Pro	Feb	2021	1	[0.92, 444]	[0.8464, 197136]
2	Samsung	Galaxy S21	Nov	2021	1	[0.62, 121]	[0.3844, 14641]
3	Infinix	12 Play NFC	May	2023	1	[0.48, 70]	[0.2304, 4900]
4	Huawei	P40 Pro	Oct	2021	1	[0.14, 3]	[0.0196, 9]
5	Infinix	12 Play NFC	Nov	2022	1	[0.05, 0]	[0.0025, 0]

3. Calculating Euclidean Distance
   Below are the Euclidean distance calculations between data points used in the clustering process with the BIRCH algorithm. This distance calculation is essential for determining the proximity between data points, which is then used to group sales data and smartphone specifications into appropriate clusters. The results will provide the foundation for the BIRCH algorithm to identify patterns and trends in the analyzed data.
   Distance Between Data 1 and Data 2:
   $$\sqrt{(0.92-0.62)^2+(444-121)^2+(0.8464-0.3844)^2+(197136-14641)^2}=182.495$$
   Distance Between Data 1 and Data 3:
   $$\sqrt{(0.92-0.48)^2+(444-70)^2+(0.8464-0.2304)^2+(197136-4900)^2=192.236}$$
   Distance Between Data 1 and Data 4:
   $$\sqrt{(0.92-0.14+(444-3)^2+(0.8464-0.0196)^2+(197136-9)^2=197.127}$$
   Distance Between Data 1 and Data 5:
   $$\sqrt{(0.92-0.05)^2+(444-0)^2+(0.8464-0.025)^2+(197136-0)^2=197.136}$$
   Where the results are as follows:

Table 6. Euclidean Calculation Results

Data Point	Distance
Data 1 – Data 2	182.495
Data 1 – Data 3	192.236
Data 1 – Data 4	197.127
Data 1 – Data 5	197.136

4. Calculating the Radius of a Cluster
   The cluster radius determines whether a new data point can be added to an existing cluster. Let's take an example of a cluster consisting of three data points:

Data in the Cluster

Table 7. Radius of the Cluster

Data	LS1	LS2	SS1	SS2
Data1	0.92	444	0.8464	197136
Data2	0.62	121	0.3844	14641
Data3	0.48	70	0.2304	4900

Calculation Steps
Linear Sum (LS)
$LS = [0.92 + 0.62 + 0.48, 444 + 121 + 70] = [2.02, 635]$
Squared Sum (SS)
$SS = [0.8464 + 0.3844 + 0.2304, 197136 + 14641 + 4900] = [1.4612, 216677]$
Number of Points (N)
N = 3
Radius (R)

$$R = \sqrt{\frac{1.4612 - \frac{(2.02)^2}{3}}{3}}$$

$$R = \sqrt{\frac{1.4612 - \frac{4.0804}{3}}{3}}$$

$$R = \sqrt{\frac{1.4612 - 1.3601}{3}} = \sqrt{\frac{0.1011}{3}} = \sqrt{0.0337} \approx 0.1835$$

5. Determining Cluster for New Data Type
After calculating the radius of the cluster, the next step is to determine whether the new data type should be added to the existing cluster or if a new cluster should be created. This is done by comparing the new data's Euclidean distance to the cluster's centroid with a specific threshold.
Manual Calculation Steps:
   a. Calculate the Centroid of the Cluster
      The centroid is calculated as the average of the values within the cluster.
   b. Calculate the Euclidean Distance from New Data to the Cluster Centroid
      Use the Euclidean distance formula discussed earlier.
   c. Compare with Threshold
      Add the data to the existing cluster if the distance is less than the threshold.

If not, create a new cluster.
Let's consider an example with the existing cluster data and a new data point we want to add.
Data in the Cluster :

Table 8. Data in the Cluster

Data	LS1	LS2	SS1	SS2
Data1	0.92	444	0.8464	197136
Data2	0.62	121	0.3844	14641
Data3	0.48	70	0.2304	4900

New Data

Table 9. New Data

Data	LS1	LS2	SS1	SS2
Data4	0.14	3	0.0196	9

First, we calculate the cluster's centroid, which is determined as the average of the values within the cluster.
$Centroid\ LS1 = (0.92 + 0.62 + 0.48)/3 = 2.02/3 = 0.6733$
$Centroid\ LS2 = (444 + 121 + 70)/3 = 635/3 = 211.66$

Next, we calculate the Euclidean distance from the new data to the cluster centroid using the Euclidean distance formula:

$$D = \sqrt{(0.14 - 0.6733)^2 + (3 - 211.66)^2 + (0.0196 - 0.6733)^2 + (9 - 14641)^2}$$

$$D = \sqrt{(0.14 - 0.6733)^2 + (3 - 211.66)^2 + (0.0196 - 0.3844)^2 + (9 - 4900)^2}$$

$$D = \sqrt{(-0.5333)^2 + (-208.66)^2 + (-0.3648)^2 + (-14632)^2}$$

$$D = \sqrt{0.284 + 43538 + 0.133 + 214095}$$

$$D = \sqrt{257.633.655} \approx 16.05$$

Compare this with a threshold; let's assume our threshold is 50. Since 16.05<50, it indicates that the distance is less than the threshold, allowing the new data (Data4) to be added to the existing cluster. Since the distance of 16.05 is less than the threshold, the new data (Data4) can be added to the existing cluster.
Final Results in Tabular Form

Table 10. Final Results for New Cluster Determination

Cluster	Data	LS1	LS2	SS1	SS2	Centroid LS1	Centroid LS2	Radius	Distance to Centroid
1	Data1	0.92	444	0.8464	197136	0.6733	211.66	0.1835	-
1	Data2	0.62	121	0.3844	14641	0.6733	211.66	0.1835	-
1	Data3	0.48	70	0.2304	4900	0.6733	211.66	0.1835	-
1	Data4	0.14	3	0.0196	9	0.6733	211.66	0.1835	16.05

Summary of Steps:
1. Calculating the Centroid of the Cluster:
   Centroid LS1: 0.6733
   Centroid LS2: 211.66
2. Calculating the Euclidean Distance from New Data to the Cluster Centroid:
   Distance: 16.05
3. Comparing with the Threshold:

The distance of 16.05 is less than the threshold of 50, allowing the new data to be added to the existing cluster.
With these steps, we can determine whether a new data point should be added to the existing cluster or if a new cluster should be created based on the radius and distance to the centroid.

6. Updating Clusters and Merging Clusters

After determining whether the new data can be added to the existing cluster, the next step is to update the cluster and, if necessary, merge clusters too close. Here are the detailed steps for this process:
Add New Data to the Existing Cluster:
If the new data is added to the existing cluster, we need to update the Linear Sum (LS), Square Sum (SS), and Centroid of that cluster.
Example of Updated Data in the Cluster:

Table 11. Updated Data in the Cluster

Data	LS1	LS2	SS1	SS2
Data1	0.92	444	0.8464	197136
Data2	0.62	121	0.3844	14641
Data3	0.48	70	0.2304	4900
Data4	0.14	3	0.0196	9

Updating Linear Sum (LS) and Square Sum (SS) of the Cluster

$$LS = [0.92 + 0.62 + 0.48 + 0.14, 444 + 121 + 70 + 3] = [2.16, 638]$$

$$SS = [0.8464 + 0.3844 + 0.2304 + 0.0126, 197136 + 14641 + 4900 + 9 = [3.845, 216686]$$

Updating the Centroid :

$$Centroid \text{ LS1} = \frac{2.16}{4} = 0.54$$

$$Centroid \text{ LS2} = \frac{638}{4} = 159.5$$

Next, we combine the clusters by checking the distance between them, following these steps:
a. Calculate the Euclidean distance between the centroids of different clusters.
b. If the distance between two clusters is less than the merging threshold, combine the two clusters.

Let's consider the following two clusters:

Cluster A :

Table 12. Cluster A

Data	LS1	LS2	SS1	SS2
Data1	0.92	444	0.8464	197136
Data2	0.62	121	0.3844	14641

Cluster B:

Table 13. Cluster B

Data	LS1	LS2	SS1	SS2
Data3	0.48	70	0.2304	4900
Data4	0.14	3	0.0196	9

Centroids of Cluster A and B:

$$\text{Centroid A} = \left[\frac{0.92 + 0.62}{2}, \frac{444 + 121}{2}\right] = [0.982, 282.5]$$

$$\text{Centroid B} = \left[\frac{0.48 + 0.14}{2}, \frac{70 + 3}{2}\right] = [0.62, 36.5]$$

Euclidean Distance Between Centroids A and B:

$$D = \sqrt{(0.982 - 0.62)^2 + (282.5 - 36.5)^2}$$

$$D = \sqrt{(0.362)^2 + (246)^2}$$

$$D = \sqrt{0.1310 + 60516} = \sqrt{60.516.131} \approx 1.7046$$

If the distance DDD is less than the merging threshold, for example, 3, the two clusters will be combined.

Table 14. Updated Cluster Results

Data	**LS1**	**LS2**	**SS1**	**SS2**
Data1	0.92	444	0.8464	197136
Data2	0.62	121	0.3844	14641
Data3	0.48	70	0.2304	4900
Data4	0.14	3	0.0196	9

Update of Linear Sum (LS) and Square Sum (SS) for the Combined Cluster:

$$LS = [0.92 + 0.62 + 0.48 + 0.14, 444 + 121 + 70 + 3] = [2.16, 638]$$

$$SS = [0.8464 + 0.3844 + 0.2304 + 0.0126, 197136 + 14641 + 4900 + 9 = [3.845, 216686]$$

Updated Centroid Calculation:

$$\textit{Centroid}\ \text{LS1} = \frac{2.16}{4} = 0.54$$

$$\textit{Centroid}\ \text{LS2} = \frac{638}{4} = 159.5$$

Updated Radius Calculation:

$$R = \sqrt{\frac{3.845 - \frac{(2.16)^2}{4}}{4}}$$

$$R = \sqrt{\frac{3.845 - \frac{4.6656}{4}}{4}} = \sqrt{\frac{3.845 - 1.1664}{4}} = \sqrt{0.66965} \approx 0.818$$

Here's how to structure the final results summary in table format:

Table 15. Final Results

Cluster	**Data**	**Centroid LS1**	**Centroid LS2**	**Radius**	**Distance to Centroid**
1	Data1	0.7605	6.5	0.032	-
1	Data2	0.7605	6.5	0.032	-
1	Data3	0.7605	6.5	0.032	-
1	Data4	0.7605	6.5	0.032	16.05

5. Conclusion

Based on the clustering results of smartphone sales at TM_Store, there is a clear difference in monthly and annual sales categories distribution. TM_Store experienced a significant increase in the "Moderate" category, with sales skyrocketing from 12 units in 2020 to 354 in 2023. Meanwhile, sales in the "Very Low" and "Low" categories remained substantial, with 70 and 86 units sold in 2023, respectively. TM_Store has achieved a more balanced distribution across various categories year by year. In contrast, the clustering results for smartphone sales at Jaya Com are dominated by the "Very High" category, which saw an extraordinary rise from only 25 units sold in 2020 to 597 units in 2023.

Additionally, the "High" category at Jaya Com has consistently grown from 6 units in 2020 to 98 in 2023. This trend indicates that Jaya Com consistently focuses on selling products with powerful performance. The testing results found that the BIRCH algorithm identified 1,984 cluster points for the Jaya Com dataset, while TM_Store identified 1,811 cluster points. This indicates that the BIRCH algorithm can effectively detect significant sales patterns in both stores, aiding in the development of more effective and efficient promotional strategies tailored to the performance of each sales category.

Acknowledgment

To deepen the market segmentation, future research should focus on a detailed analysis of the characteristics of customers who purchase products in the high-sales category at TM_Store and Jaya Com. Understanding the factors influencing these customer preferences will aid in designing more targeted marketing strategies, enabling each store to retain and increase its market share more effectively.

References

[1] Aditya, A., Jovian, I., & Sari, B. N. (2020). Implementasi K-Means *Clustering* Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019. *Jurnal Media Informatika Budidarma*, *4*(1), 51-58. https://doi.org/10.30865/mib.v4i1.1784

[2] Afiasari, N., Suarna, N., & Rahaningsi, N. (2023). Implementasi Data Mining Transaksi Penjualan Menggunakan Algoritma Clustering dengan Metode K-Means. *Jurnal SAINTEKOM*, *13*(1), 100–110. https://doi.org/10.33020/saintekom.v13i1.402

[3] Alzu'Bi, A., & Barham, M. (2022). Automatic *BIRCH Threshold*ing With Features Transformation For Hierarchical Breast Cancer *Clustering*. *International Journal of Electrical and Computer Engineering*, *12*(2), 1498–1507. https://doi.org/10.11591/ijece.v12i2.pp1498-1507

[4] Anam, C., & Iswari, R. (2021). Kemampuan Pengusaha Dari Perspektif Orientasi Kewirausahaan dan Konsep Penjualan Di Masa New Normal. *Jurnal Ilmu Manajemen*, *10*(2), 93-100. https://doi.org/10.32502/jimn

[5] Dinata, R. K., Safwandi, S., Hasdyna, N., & Azizah, N. (2020). Analisis k-means clustering pada data sepeda motor. *INFORMAL: Informatics Journal*, *5*(1), 10-17. https://doi.org/10.19184/isj.v5i1.17071

[6] Harahap, L. M., Fuadi, W., Rosnita, L., Darnila, E., & Meiyanti, R. (2022). Klastering Sayuran Unggulan Menggunakan Algoritma K-Means. *Jurnal Teknik Informatika Dan Sistem Informasi*, *8*(3). https://doi.org/10.28932/jutisi.v8i3.5277

[7] Hasibuan, F. P. A., Sumarno, S., & Parlina, I. (2021). Penerapan K-Means pada Pengelompokan Penjualan Produk Smartphone. *SATESI: Jurnal Sains Teknologi Dan Sistem Informasi*, *1*(1), 15–20. https://doi.org/10.54259/satesi.v1i1.3

[8] Hidayat, I., Darnila, E., & Afrillia, Y. (2023). Clustering Zonasi Daerah Rawan Bencana Alam di Kabupaten Mandailing Natal menggunakan Algoritma K-Means. *G-Tech: Jurnal Teknologi Terapan*, *7*(3), 1218–1226. https://doi.org/10.33379/gtech.v7i3.2880

[9] Karyono, K., Violin, V., Osman, I., Rao, D. G., & Apramilda, R. (2024). Analysis of The Interrelationship of Human Resource Performance, Digital Service Quality, Perceived of Service Value and Customer Loyalty. International Journal of Engineering, Science and Information Technology, 4(3), 66-71. .https://doi.org/10.52088/ijesty.v4i3.527

[10] Kushariyadi, K., Yani, I., Silamat, E., Sari, T. N., & Aulia, M. R. (2024). Analysis of The Influence of Market Consumption Behavior and Economic Conditions on SME Business Performance. International Journal of Engineering, Science and Information Technology, 4(3), 35-40. https://doi.org/10.52088/ijesty.v4i3.521

[11] Maulida, L. (2020). Penerapan *Data Mining* dalam Mengelompokkan Kunjungan Wisatawan ke Objek Wisata Unggulan di Prov. DKI Jakarta dengan K-Means. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *2*(3), 167-174. https://doi.org/10.14421/jiska.2018.23-06

[12] Mina, M., & Kartika, K. (2023). Monitoring System for Levels of Voltage, Current, Temperature, Methane, and Hydrogen in IoT-Based Distribution Transformers. International Journal of Engineering, Science and Information Technology, 3(1), 22-27. https://doi.org/10.52088/ijesty.v3i1.414

[13] Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT*, *4*(1), 20–24. https://doi.org/10.30591/jpit.v4i1.1253

[14] Nwadiugwu, M. C. (2020). Gene-Based *Clustering* Algorithms: Comparison Between Denclue, Fuzzy-C, and *BIRCH*. *In Bioinformatics and Biology Insights* (Vol. 14), 1-6. https://doi.org/10.1177/1177932220909851

[15] Oktory, H. D., & Hadiwandra, T. Y. (2024). Penerapan Algoritma Apriori untuk Penentuan Pola Pembelian Kacamata pada Optik Indah Optikal: Application of an Apriori Algorithm to Determine Eyeglass Purchasing Patterns at Optik Indah Optik. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(4), 1275-1281. https://doi.org/10.57152/malcom.v4i4.1353

[16] Qamal, M., Syah, F., & Parapat, A. Z. I. (2023). Implementasi *Data Mining* Untuk Rekomendasi Paket Menu Makanan Menggunakan Algoritma Apriori Implentasi *Data Mining* Untuk Rekomendasi Paket Menu Makanan Menggunakan Algoritma Apriori. *TECHSI-Jurnal Teknik Informatika*, *14*(1), 42-53. https://doi.org/10.29103/techsi.v14i1.6747

[17] Ramadhani, F., Zarlis, M., & Suwilo, S. (2020). Improve *BIRCH* Algorithm For Big Data *Clustering*. *In IOP Conference Series: Materials Science and Engineering*, *725*(1), 1-10. https://doi.org/10.1088/1757-899X/725/1/012090

[18] Suciana, D. A., & Syahputra, E. (2023). Analisis Strategi Promosi Dalam Meningkatkan Penjualan Produk Pada Resto Dan Pusat Oleh-Oleh Putra Nirwana Magetan Di Era Pandemi Covid. *Digital Bisnis: Jurnal Publikasi Ilmu Manajemen dan E-Commerce*, *2*(2), 95–115. https://doi.org/10.30640/digital.v2i2.1058

[19] Venkateswarlu, B., & Raju, P. G. S. V. P. (2022). *Mine Blood Donors Information through Improved K-Means Clustering*. *International Journal of Computational Science and Information Technology*, *1*(3), 9-15. https://doi.org/10.48550/arXiv.1309.2597
[20] Yan, Z., Yang, G., He, R., Yang, H., Ci, H., & Wang, R. (2023). Ship Trajectory Clustering Based on Trajectory Resampling and Enhanced BIRCH Algorithm. *Journal of Marine Science and Engineering*, *11*(2). https://doi.org/10.3390/jmse11020407