Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 01, 2025, DOI 10.54209/infosains.v15i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

Multivariate Data Analysis for Customer Segmentation Using Principal Component Analysis and K-Means Clustering

Bosker Sinaga
Information Technology, Mahkota Tricom Unggul University, Jl. Perintis Kemerdekaan No. 3A, Medan, Indonesia

Article Info
Keywords:
Multivariate Data Analysis,
Customer Segmentation,
Principal Component Analysis,
K-Means Clustering,
Dimension Reduction.

ABSTRACT
This study discusses multivariate data analysis for customer segmentation using Principal Component Analysis (PCA) combined with the K-Means clustering method. The problem faced is the high dimension of customer data which makes it difficult to segment and make targeted marketing decisions. The solution offered is the implementation of PCA to reduce the data dimension without losing important information, then followed by K-Means to segment customers based on demographic attributes and shopping behavior. Using a dataset of 200 customers, three customer clusters with different characteristics in terms of age, annual revenue, and shopping score were found. The results of the PCA show that the first two main components are able to explain more than 78% of the data variation, making it easier to visualize and interpret the cluster. These findings provide the basis for a more targeted marketing strategy according to customer segments. In conclusion, the combination of PCA and K-Means is effective in simplifying complex data and resulting in meaningful customer segmentation.

Corresponding Author:
Bosker Sinaga
Mahkota Tricom Unggul University
Jl. Perintis Kemerdekaan No. 3A
boskersinaga@gmail.com

INTRODUCTION

The development of information technology and advances in data processing have opened up new opportunities for companies to understand customer behavior more deeply (Listy & Ilham, 2025). Data generated from customer transactions, interactions, and activities can now be collected and analyzed to generate strategic insights. One form of analysis that is widely used is customer segmentation, which is the process of grouping customers into groups that have similar characteristics (Awalina & Rahayu, 2023). This segmentation is an important basis for companies to design marketing strategies that are more targeted, increase customer satisfaction, and ultimately maximize profits.

Customer segmentation plays a strategic role in the modern business world because it allows companies to tailor products, services, and marketing strategies based on the specific needs of each customer group (Santoso et al., 2024). Companies that are able to identify high-value customer groups can allocate resources more efficiently and minimize marketing budget waste. In addition, proper segmentation also helps in building long-term relationships

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 01, 2025, DOI 10.54209/infosains.v15i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

with customers through personalization of relevant services and communication (Ramadian et al., 2025).

However, customer segmentation is not simple, especially if the data is multivariate, consisting of many variables that interact with each other. Multivariate data often has a high dimension, making it difficult to analyze and visualize directly (Sihombing, 2022). The existence of intercorrelated features can cause information redundancy, slow down the analysis process, and decrease the accuracy of segmentation results. Therefore, an effective dimension reduction method is needed to simplify the data without eliminating the important information in it.

Principal Component Analysis (PCA) is one of the widely used statistical methods for dimension reduction in multivariate data (Shalih et al., 2025). PCA works by converting the original variable into a new set of independent variables (principal components), sorted by the contribution of variance to the overall data (Bharadiya, 2023). By using PCA, data complexity can be reduced making it easier to analyze and visualize. In addition, PCA helps reduce noise and address multicollinearity issues between variables, which often appear in customer data (Santosa, 2023).

After the data dimension is reduced using PCA, the segmentation process can be carried out using the K-Means Clustering method. K-Means is an effective unsupervised learning algorithm to group data into k groups based on the closest distance to the center of the cluster (Adawiyah & Defit, 2024). This method is known for its simplicity, computational efficiency, and ability to generate clear segmentation (Dubey & Rajavat, 2023). The combination of PCA and K-Means has been proven to improve the quality of segmentation results by minimizing interference from irrelevant variables, while making it easier to visualize results.

Although the PCA and K-Means methods have been widely used separately, the application of both in an integrated manner in customer segmentation still requires in-depth study, especially on customer data with varying demographic and behavioral variables. Many previous studies have only focused on the use of K-Means without first optimizing the data structure through dimension reduction. This has the potential to result in segmentation that is less accurate or difficult to interpret. Therefore, this study seeks to apply a combined approach of PCA and K-Means to multivariate data analysis to obtain more optimal and informative customer segmentation.

This study aims to develop a customer segmentation model based on multivariate data analysis by utilizing the Principal Component Analysis (PCA) method for dimension reduction and K-Means Clustering. Specifically, this study aims to identify the key components that represent the greatest variability in customer data, segment customers based on similarity characteristics in the PCA space, and produce visualizations that facilitate the interpretation of each segment's profile. The results of this research are expected to contribute to the development of more effective and data-driven marketing strategies.

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 01, 2025, DOI 10.54209/infosains.v15i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

METHODS

This study uses a quantitative approach based on data mining by combining the Principal Component Analysis (PCA) method for dimension reduction and K-Means Clustering for customer data grouping. The research process is carried out systematically through several main stages, ranging from data collection, data preprocessing, algorithm application, to visualization and model storage.

The research was designed with the aim of segmenting customers based on three main variables, namely Age, Annual Income (k$), and Spending Score (1-100). The PCA method is used to reduce the dimensions of the data so that variables that have a high correlation can be compacted into key components without losing important information. Furthermore, the K-Means Clustering method is used to form customer groups based on data projections in the PCA space.

The data used comes from Mall_Customers.csv public datasets downloaded from the Kaggle platform. This dataset contains demographic and behavioral information on customers as many as 200 entries. The selection of this dataset is based on the completeness of the relevant attributes for customer segmentation analysis. The pre-processing stage begins with the selection of three relevant features from the dataset, namely Age, Annual Income (k$), and Spending Score (1-100). Next, correlation analysis between features was carried out using heatmaps to see the relationships between variables. After that, feature scaling is carried out using the StandardScaler from scikit-learn so that each variable has a distribution with an average of 0 and a standard deviation of 1. This process is important to prevent bias due to scale differences between variables in PCA and K-Means.

The PCA method is applied to convert the original variable into an uncorrelated main component. Initially, a fit transform was carried out on all features to calculate the explained variance ratio of each component, which was visualized in the form of a scree plot. Based on the results of the analysis, two main components (n_components=2) were determined to be used for data representation. The total variance that can be explained by these two components is calculated to ensure that the missing information is minimal.

The general equation of PCA can be written as follows (Badri & Sari, 2021):

$$Z = XW \quad (1)$$

Once the data is reduced to two dimensions, the K-Means Clustering method is applied with the optimal number of clusters k = 3. The K-Means algorithm works by minimizing objective functions (Borlea et al., 2022):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n_j} |x_i^{(j)} - c_j|^2 \quad (2)$$

The visualization of the grouping results is carried out in three forms, namely scatter plots in PCA spaces with coloring based on clusters and centroid point marking to see the group separation clearly, pairplots on the original features used to observe the distribution of variables in each cluster, and plot bars that display the average of each feature in each cluster as a customer segmentation profile.

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 01, 2025, DOI 10.54209/infosains.v15i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

RESULTS AND DISCUSSION

In this section, the results of customer segmentation research based on Principal Component Analysis (PCA) and K-Means Clustering using customer datasets from Kaggle are described in detail. This study uses 200 customer data with five main attributes, namely CustomerID, Gender, Age, Annual Income (k$), and Spending Score (1–100). The entire analysis process is carried out in stages: data exploration, standardization, PCA dimension reduction, determination of the number of clusters, grouping using K-Means, and cluster profile analysis.

Description of Initial Data

The dataset used is the Mall Customers Dataset from Kaggle, containing 200 shopping center customer data. The dataset includes five variables: CustomerID (unique customer code), Gender (gender), Age (age in years), Annual Income (annual revenue in thousands of US dollars), and Spending Score (shopping behavior score from 1–100).

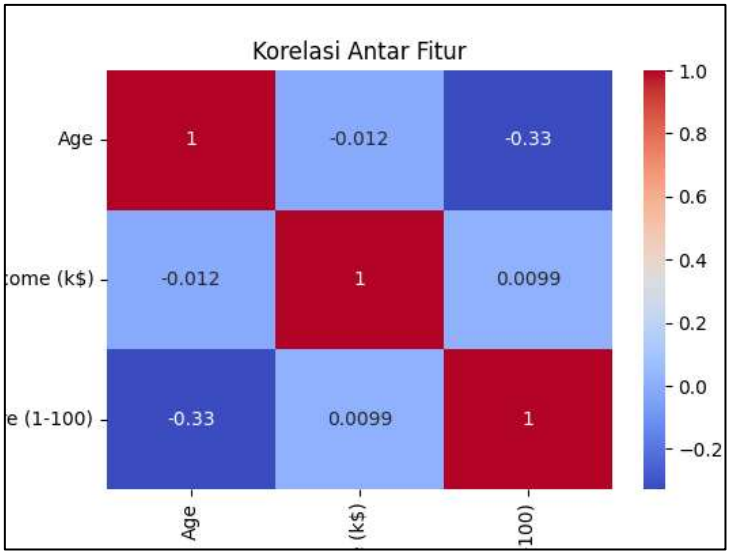

Figure 1. Intervariable Correlation Heatmap

Based on Figure 1, it can be seen that the correlation relationship between numerical variables is relatively low. The Age variable has a moderate negative correlation with a Spending Score of -0.33, which indicates that the older the customer, the lower the spending score. Meanwhile, Annual Income has a very low correlation with both Age (-0.012) and Spending Score (0.0099). This suggests that each variable contributes unique information, making it feasible to use together at the next stage of analysis without significant risk of multicollinearity.

Data Standardization and PCA Analysis

Before conducting further analysis, numerical data on the variables Age, Annual Income, and Spending Score were first standardized using the Z-score normalization method. The purpose of this standardization is to ensure that each variable has a mean of zero and a standard deviation of one, so that the difference in scale between variables does not affect the results of the analysis. This is important because methods such as PCA are very sensitive to differences in data scale. Once the data is standardized, Principal Component Analysis

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 01, 2025, DOI 10.54209/infosains.v15i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

(PCA) is performed to reduce the dimensions of the data while identifying linear combinations of variables that can explain the greatest variations in the dataset. PCA helps simplify data complexity, reduce noise, and make visualization easier, without losing too much important information.

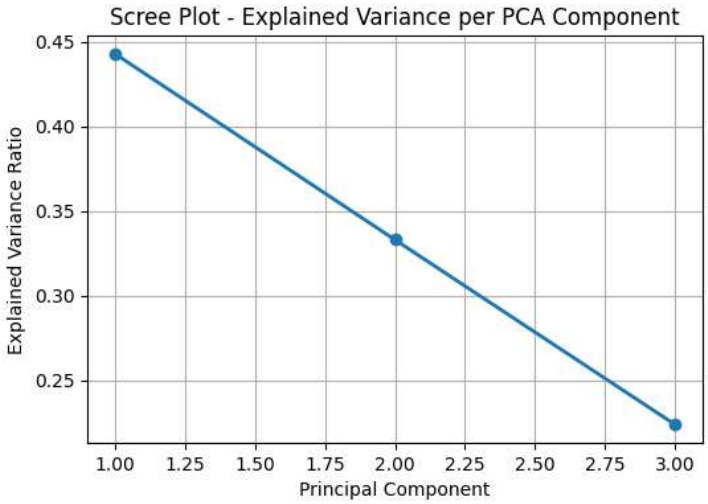

Figure 2. Scree Plot – Explained Variance per PCA Component

The results of the PCA are shown in Figure 2, in the form of a Scree Plot, which describes the explained variance ratio for each principal component. Based on the graph, Principal Component 1 (PC1) explains about 44.7% of the data variation, Principal Component 2 (PC2) explains 33.4%, and Principal Component 3 (PC3) explains 21.9%. Thus, the first two main components are able to explain about 78.1% of the total data variation, which is already representative enough to be used at the visualization and clustering analysis stages.

Visualization of Clustering Results with PCA

After determining the optimal number of clusters using the Elbow method, a clustering process is carried out with the K-Means algorithm on data that has been reduced in dimensions using PCA. This dimension reduction aims to simplify visualization and interpretation, while ensuring that the patterns that emerge are based on key variations in the data.

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 01, 2025, DOI 10.54209/infosains.v15i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

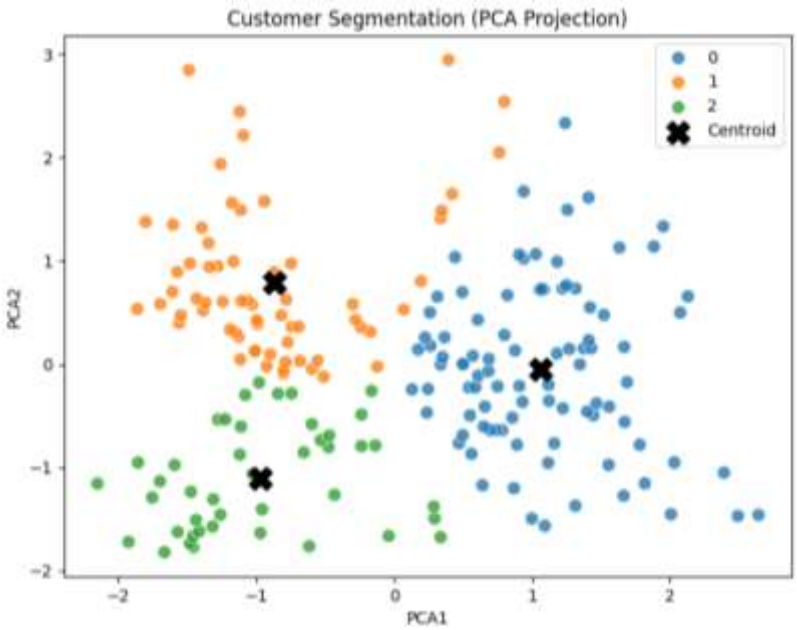

Figure 3. Visualization of K-Means Clustering in PCA Space

Based on Figure 3, it shows the results of customer segmentation in two-dimensional space based on Principal Component 1 (PCA1) and Principal Component 2 (PCA2). Each dot on the graph represents one customer, with different colors indicating the membership of each cluster, namely Cluster 0 (blue), Cluster 1 (orange), and Cluster 2 (green). The large black cross (X) on each cluster indicates the position of the centroid, which is the central point of each cluster calculated based on the average position of all cluster members in the PCA space.

This centroid position is used by the K-Means algorithm to determine the boundaries between clusters and minimize the average distance between points in the cluster. Based on the visualization, it can be seen that the distribution of each cluster is relatively separate, even though there are several points that are in the boundary area and adjacent to other clusters. This indicates that the resulting segmentation is quite effective in differentiating customer patterns based on the variables analyzed.

Visualization of Grouping Results

After the number of clusters is determined, grouping is carried out using the K-Means algorithm in the PCA space (PC1 and PC2). The results are visualized in the form of scatter plots with different coloring for each cluster, and special markings for the cluster centroids.

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 01, 2025, DOI 10.54209/infosains.v15i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

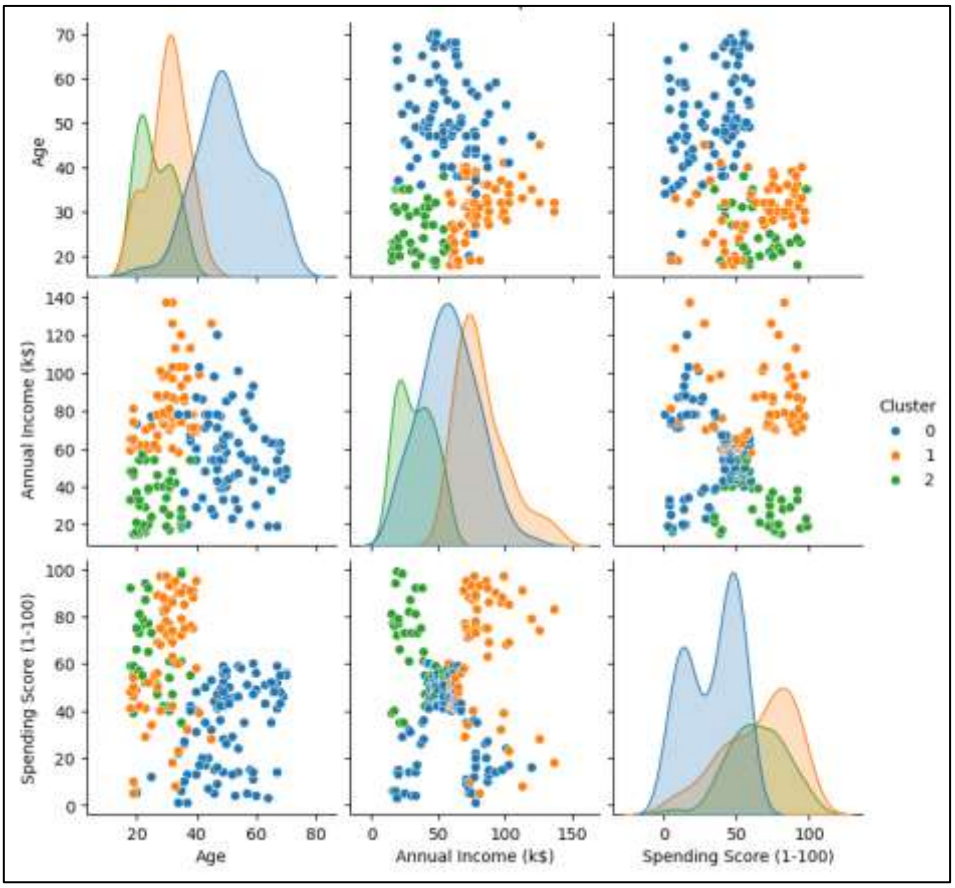

Figure 4. Visualization of Grouping Results

Based on Figure 4, it shows the relationship between Age, Annual Income (k$), and Spending Score (1–100) which has been grouped by the K-Means method into three clusters: Cluster 0 (blue), Cluster 1 (orange), and Cluster 2 (green). The diagonal graph shows the distribution of each variable, while the other graph shows the relationships between the variables. In general, Cluster 0 contains older customers with variations in income and spending scores, Cluster 1 is dominated by middle-aged with high income and moderate spending scores, while Cluster 2 contains young-middle-aged people with low income and varied spending scores.

Customer Cluster Profile Analysis

The average profile of each customer feature formed from the K-Means clustering process provides an overview of the unique characteristics of each group. The variables analyzed included Age, Annual Income (k$), and Spending Score (1–100). The average difference in each variable shows the difference in behavior and potential customer value that can be used in marketing strategy planning.

Jurnal Info Sains : Informatika dan Sains
Volume 15, Number 01, 2025, DOI 10.54209/infosains.v15i01
ESSN 2797-7889 (Online)
https://ejournal.seaninstitute.or.id/index.php/InfoSains

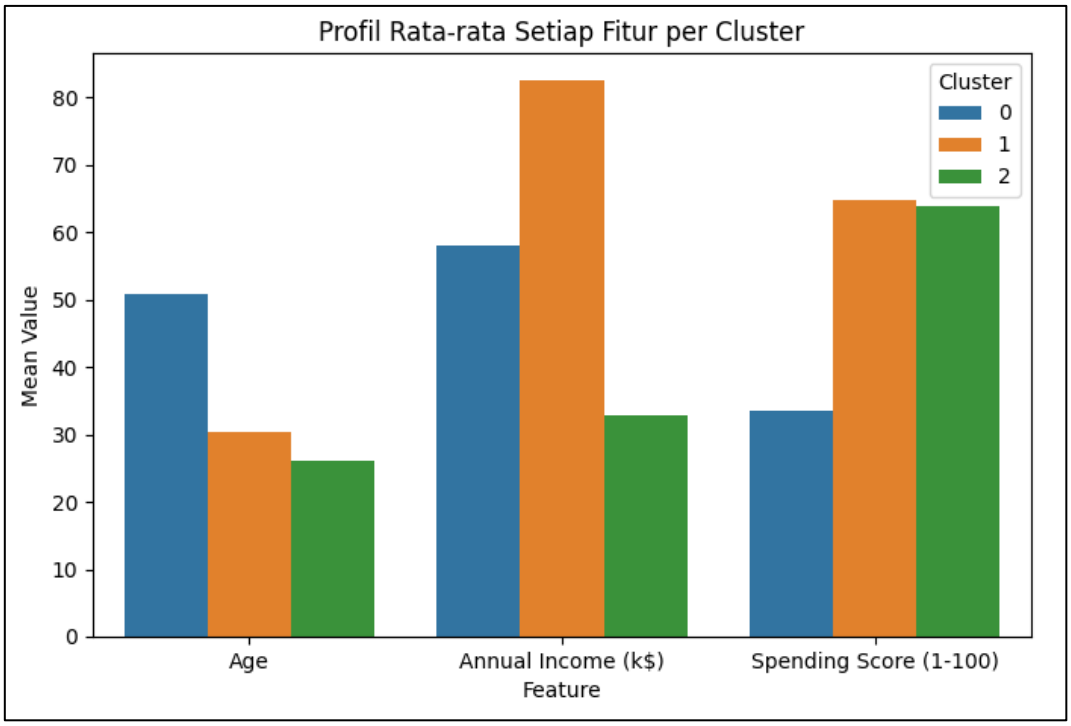

Figure 5. Average Profile of Each Feature per Cluster

From the visualization in Figure 5, it can be seen that Cluster 0 (blue) has the highest average lifespan, medium annual income, and low spending score. Cluster 1 (orange) stands out with the highest annual revenue, middle age, and fairly high spending score, making it a segment of high value customers. Meanwhile, Cluster 2 (green) is made up of customers with the youngest, lowest income, but high shopping scores. This difference shows the need for a different marketing approach for each segment, so that the strategy applied is more effective and on target.

CONCLUSION

This study successfully segmented customers using the K-Means clustering method combined with Principal Component Analysis (PCA) dimension reduction. The three clusters formed had different characteristics based on age, annual revenue, and customer shopping behavior scores. Cluster 0 consists of customers with the oldest age and low spending scores, Cluster 1 is the highest income customer but moderate shopping score, while Cluster 2 contains the youngest customer with the lowest income but high shopping score. This segmentation shows that there are differences in customer behavior patterns that can be used as the basis for a more targeted marketing strategy. Visualization of the PCA results and scatter plots confirms a fairly clear division of clusters despite the slight overlap in the boundary areas between groups.

REFERENCE

Adawiyah, Q., & Defit, S. (2024). Penerapan Algoritma K-Means Clustering untuk

Mengelompokkan Rekomendasi Metode Kontrasepsi Berbasis Machine Learning di Puskesmas. *Jurnal KomtekInfo*, 300–305.

Awalina, E. F. L., & Rahayu, W. I. (2023). Optimalisasi Strategi Pemasaran dengan Segmentasi Pelanggan Menggunakan Penerapan K-Means Clustering pada Transaksi Online Retail. *Jurnal Teknologi Dan Informasi*, *13*(2), 122–137.

Badri, F., & Sari, S. U. R. (2021). Penerapan metode Principal Component Analysis (PCA) untuk identifikasi faktor-faktor yang mempengaruhi sikap mahasiswa memilih melanjutkan studi ke Kota Malang. *Build. Informatics, Technol. Sci*, *3*(3), 426–431.

Bharadiya, J. P. (2023). A tutorial on principal component analysis for dimensionality reduction in machine learning. *International Journal of Innovative Science and Research Technology*, *8*(5), 2028–2032.

Borlea, I.-D., Precup, R.-E., & Borlea, A.-B. (2022). Improvement of K-means cluster quality by post processing resulted clusters. *Procedia Computer Science*, *199*, 63–70.

Dubey, P., & Rajavat, A. (2023). Effective K-means clustering algorithm for efficient data mining. *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 1–6.

Listy, V., & Ilham, I. (2025). Revolusi Sistem Informasi Manajemen di Era AI dan Big Data Mengubah Cara Bisnis Bekerja. *Simpatik: Jurnal Sistem Informasi Dan Informatika*, *5*(1), 27–36.

Ramadian, A., Judijanto, L., & Erwin, E. (2025). *Customer Relationship Management (CRM): Strategi Membangun Hubungan Pelanggan yang Kuat*. PT. Green Pustaka Indonesia.

Santosa, Y. P. (2023). Kombinasi Linier Target Data Untuk Regresi Multitarget Menggunakan Principal Component Analysis. *Jurnal Teknologi Terpadu*, *9*(1), 1–9.

Santoso, R. P., Ningsih, L. S. R., & Irawati, W. (2024). Implementation Of Segmenting Targeting And Positioning Strategies In Improving Marketing Performance. *BIMA: Journal of Business and Innovation Management*, *6*(2), 280–292.

Shalih, F. A., Ramadhan, R. A., & Syalaisa, N. (2025). Comprehensive Overview of Principal Component Analysis Applications and Developments. *Jurnal EurekaMatika*, *13*(1), 25–34.

Sihombing, S. O. (2022). *Pengantar metode analisis multivariat*. Penerbit NEM.