iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XI **Month of publication:** November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75316

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XI Nov 2025- Available at www.ijraset.com

Machine Learning Based Customer Personality Analysis

Smit Pendor[1], Shital Pawar[2]
Dept of E&T Engineering, Vishwakarma Institute of Technology, Pune, India

Abstract: ***This project provides a solid customer segmentation framework using machine learning techniques to improve targeted marketing strategies. By preprocessing and analyzing the marketing campaign dataset, we extracted and transformed key features like income, age, spending habits, and campaign response. We applied dimensionality reduction with PCA and used clustering techniques such as K-Means and Agglomerative Clustering to group customers based on similar behaviors. The segmented data allows businesses to tailor their marketing efforts more effectively, boosting engagement and conversion rates. Visualization and statistical analysis further confirm the clusters, giving valuable insights into consumer behavior and purchasing patterns.***

Keywords: Customer Segmentation, Machine Learning, Targeted Marketing, Data Preprocessing, Feature Engineering, Dimensionality Reduction, Principal Component Analysis (PCA), K-Means Clustering, Agglomerative Clustering, Consumer Behavior, Marketing Campaign, Data Visualization, Purchase Patterns, Business Intelligence, Behavioral Analysis

I. INTRODUCTION

Customer segmentation is a crucial process in modern marketing analytics, enabling organizations to tailor their strategies based on customer behavior and demographics. This study presents a data-driven approach to customer segmentation using machine learning and unsupervised clustering techniques on a real-world marketing dataset. The dataset contains diverse attributes, including demographic information, product spending patterns, and customer interaction history. To ensure data quality, preprocessing steps such as handling missing values, eliminating duplicates, encoding categorical features, and feature scaling were performed.

Feature engineering played a vital role in enhancing insights, with new features like total spending (Expenses), customer tenure (Customer_For), and aggregated campaign responses being derived. Outlier detection and removal were applied to prevent distortion in clustering outcomes. K-Means clustering and Agglomerative Hierarchical Clustering were employed to identify natural groupings in the customer base. Additionally, Principal Component Analysis (PCA) was used to reduce dimensionality and facilitate visualization.

Through visual analysis and cluster interpretation, meaningful patterns were uncovered, such as the correlation between spending behavior and age, income, education level, and purchase channels. This approach demonstrates the power of unsupervised learning in uncovering customer profiles, aiding targeted marketing strategies, and enhancing business decision-making. The methodology offers a scalable framework adaptable to various industries.

II. LITERATURE REVIEW

In A Novel Approach for Customer Segmentation and Product Recommendation to Boost Sales using Machine Learning, a hybrid model that combines collaborative filtering and K-Means clustering is proposed for customer segmentation and product recommendation. Using past sales data, it improves targeting accuracy and shows higher conversion rates.[1] A Supervised Clustering Method for Commercial Bank Customer Segmentation according to Customer Value integrates customer value dimensions like profitability and loyalty to present a supervised clustering technique designed for banking customers. When compared to conventional unsupervised techniques, the study shows improved segmentation accuracy.[2] K-Means Clustering and Seasonal ARIMA for Behavioral Segmentation with Product Estimation combines time series prediction (ARIMA) and unsupervised segmentation using K-Means to estimate product demand based on consumer behavior. The approach demonstrates accurate forecasting that is in line with seasonal patterns. [3] A Comparison of Supervised and Unsupervised Learning Technologies for Marketing Customer Segmentation compares unsupervised clustering (K-Means, DBSCAN) and supervised learning (decision trees, etc.). Findings indicate that integrating both improves personalization and segmentation quality.[4] Segmenting Customers Making use of K-means Grouping divides clients according to spending and demographics using standard K-Means; interpretation is done through visual analysis.

Acts as a starting point or standard for customer segmentation using straightforward but efficient methods.[5] Study on Clustering-Based Customer Segmentation Model explains the significance of feature selection and normalization while introducing a conventional clustering-based segmentation model. confirms that feature engineering and preprocessing are crucial before using clustering. .[6] Effective Digital Marketing Customer Segmentation Combining Swarm Intelligence and Deep Learning employs a hybrid model that combines swarm optimization for clustering and deep autoencoders for feature extraction. [7] Using Bibliometrics and Machine Learning to Segment Customers This meta-study uses bibliometric techniques to analyze trends in machine learning-based customer segmentation research. gives a more comprehensive view of the development and areas of emphasis in customer segmentation research.[8] Comparative Evaluation of Customer Segmentation Machine Learning Models uses the same dataset to test various machine learning models (K-Means, Hierarchical Clustering, DBSCAN, etc.) for segmentation. provides advice on how to choose the best clustering algorithm for various segmentation situations. [9] K-Means Clustering for Customer Segmentation in Banking for Targeted Marketing focuses on segmenting bank customers using K-Means and enhancing marketing campaigns according to their financial behavior. A domain-specific application that demonstrates how well clustering works in the banking and financial services industry.

III. METHODOLOGY

This section outlines the process used to apply unsupervised machine learning techniques to customer segmentation on a marketing campaign dataset. Finding unique consumer groups to support focused marketing initiatives is the aim. Data collection, preprocessing, feature engineering, exploratory data analysis (EDA), normalization, clustering, dimensionality reduction, and interpretation are some of the crucial stages in the process.

A. Gathering and Preparing Data

Customer demographics and behavioral characteristics, including age, income, education, marital status, product purchases, campaign responses, and enrollment dates, are included in the dataset. We start by loading the dataset and using pandas functions to analyze its structure. Duplicate entries are eliminated, and the median is used to impute null values in the Income column. To cut down on noise and dimensionality, uninformative columns like Z_CostContact and Z_Revenue are removed.

B. Engineering Features

To record consumer behavior, a number of new variables are developed:

- Customer Age: Determined by the customer's birth year.
- Costs: Totaled from purchases of meats, fruits, wines, and other goods.
- Children: a combination of Teenhome and Kidhome.
- Customer_For: Based on the date of enrollment.
- TotalAcceptedCmp: The sum of all campaigns that have been approved.
- NumTotalPurchases: The total of all kinds of purchases.

Richer information is provided by these engineered features, which improve model performance.

C. Encoding and Categorical Simplification

Education levels and marital statuses are combined into more general categories like "Post Graduate" and "Relationship" in order to streamline sparse categories. These are transformed into a numerical format appropriate for clustering algorithms using label encoding.

D. Analysis of Exploratory Data (EDA)

Matplotlib and Seaborn are used for EDA. We look at feature relationships, analyze distributions, and find outliers. To learn more about variables like Income, Expenses, and Customer Age across different categories, boxplots, bar plots, and scatter plots are used.

E. Standardization

StandardScaler is used to normalize numerical features because clustering algorithms are sensitive to scale. In order to guarantee consistent contributions to distance computations, scaled features include income, customer age, expenses, and other engineered variables.

F. Algorithms for Clustering

K-Means Grouping

The Elbow Method is used to determine the ideal number of clusters. Based on the WCSS plot's inflection point, K is set to 2. Each customer is given a cluster label by the K-Means algorithm according to how close they are to centroids. Clustering Agglomeratively

Alternatively, a bottom-up hierarchical method is used to carry out Agglomerative Clustering. It is used on data that has been reduced using Principal Component Analysis (PCA) and two clusters.

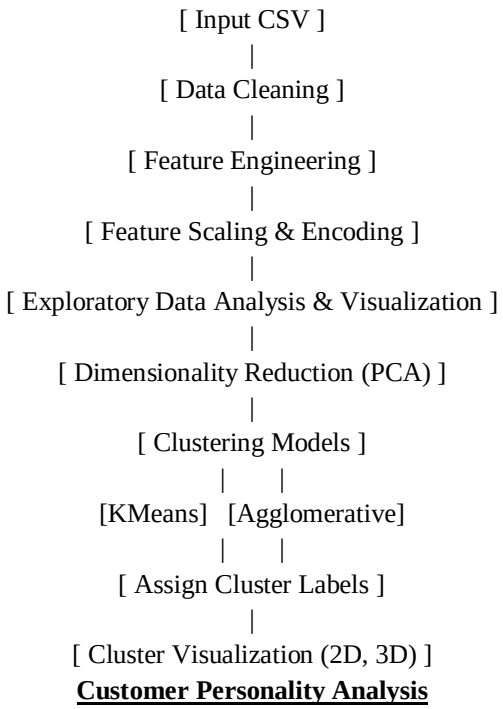
G. Visualization and Dimensionality Reduction

In order to visualize high-dimensional data, PCA breaks it down into three principal components. 3D scatter plots are useful for validating the effectiveness of clustering and evaluating cluster separation. Additional plots, such as Customer Age vs. Income and Income vs. Expenses, aid in the interpretation of segment characteristics.

H. Cluster Interpretation

The final step involves analyzing and visualizing the cluster profiles. Count plots and scatter plots reveal insights, such as identifying high-income, high-spending customers or younger, less engaged segments.

System architecture



Customer Personality Analysis

IV. RESULTS AND DISCUSSION

The goal of the customer segmentation project was to use clustering techniques to classify customers according to their behavioral and demographic characteristics. We processed the dataset, created useful features, and used clustering algorithms to find distinct customer groups using a thorough data analysis and machine learning pipeline. The main conclusions, revelations from the exploratory data analysis, clustering model outcomes, and their ramifications are covered in this section.

A. Data Cleaning and Exploration

2,240 records and 29 variables pertaining to consumer demographics, product purchases, marketing campaign reactions, and behavioral indicators made up the original dataset. Missing values in the Income field, which made up a tiny percentage of the data, were discovered during an initial examination. To reduce skewness and maintain the distribution, these missing values were imputed using the median income.

The integrity of the dataset was maintained by checking for duplicates and finding none.
In order to streamline the dataset, we also eliminated low-information features and redundant columns (Z_CostContact, Z_Revenue). Additionally, disparate categories in Education and Marital Status were combined: education levels were categorized as "Post Graduate" and "Under Graduate," and marital statuses were combined into "Single" and "Relationship." By ensuring categorical consistency and reducing dimensionality, these procedures improved the analysis's interpretability.

B. Engineering Features

To improve predictive ability and derive useful insights, new features were developed:
• Customer_Age: Determined by deducting the current year from the customer's birth year, this number provides an age distribution that primarily falls between 30 and 80 years old.
• Costs: The total amount spent on all product categories (wine, fruit, meat, fish, sweets, and gold), capturing the total amount spent
• Kids: The total number of kids and teens living at home.
• TotalAcceptedCmp: A summary of customer responsiveness based on responses to five distinct marketing campaigns.
• Customer_For: Indicates the number of days since enrollment for the customer.
In order to guarantee that the dataset represented accurate customer profiles, outlier treatment eliminated implausible records (such as clients who were older than 90 or had incomes greater than $300,000.

C. Analysis of Exploratory Data (EDA)

The cleaned data's visualization yielded insightful business information:
• The distribution of income was right-skewed, with the majority of clients making between $20,000 and $80,000.
• With a few high-spending anomalies, expenses were mostly under $2,000.
• Customers who were in a relationship and had more education tended to spend more.
• Costs rose slightly as the size of the household (number of children) increased.
• TotalAcceptedCmp and expenses showed a positive correlation, suggesting that customers who accepted more campaigns also tended to spend more money overall.
• It's interesting to note that there was no significant correlation between expenses and customer tenure (Customer_For), indicating that spending levels and loyalty duration were not directly related.
The idea that income and marketing responsiveness are important factors influencing spending was supported by a correlation heatmap, which showed weak to moderate correlations between income, expenses, and total accepted cents.

D. Results of Clustering

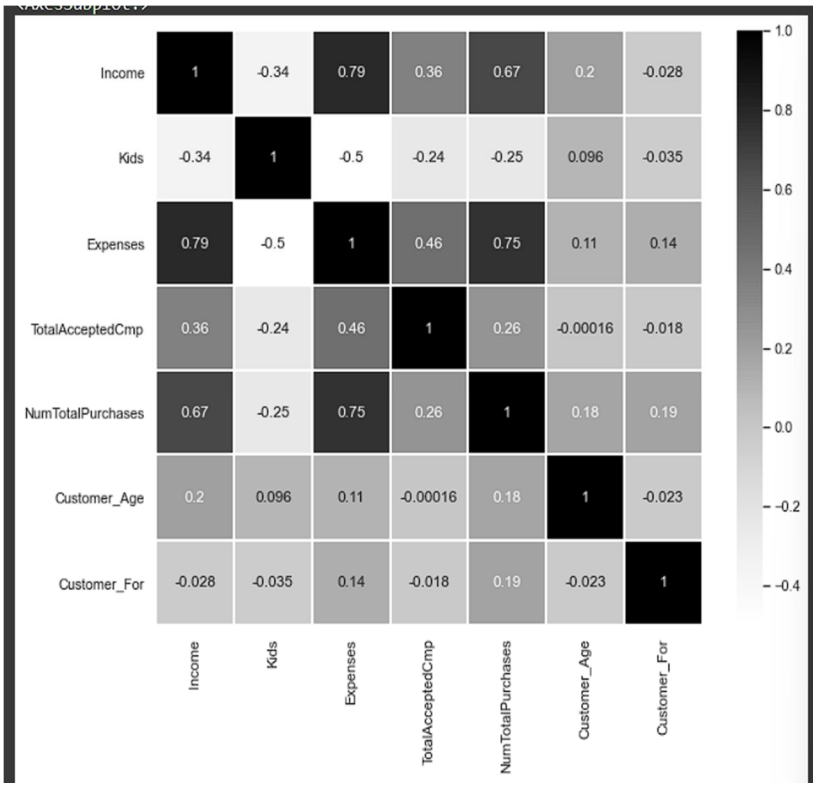
K-Means Grouping
We assessed cluster inertia between k=1 and k=10 using the Elbow Method. Two clusters offered the best balance between granularity and interpretability, according to the inflection point at k=2.
The clients were divided into two separate clusters using K-Means:
Above-average income, higher expenses, more campaign acceptances, and more cross-channel purchases are characteristics of Cluster 1 (Higher Spend, Higher Income).
• Cluster 2 (Lower Spend, Lower Income): Served clients who were less affluent, made fewer purchases, and responded less well to campaigns.
Customers in Cluster 1 are not only wealthier but also more active buyers, according to a 2D scatter plot of expenses versus income color-coded by cluster.
Clustering Agglomeratively
Agglomerative Clustering was used on a PCA-reduced dataset with three components in order to validate K-Means results. Hierarchical clustering revealed two main clusters, just like K-Means. A 3D PCA plot of these clusters revealed clearly defined groups that closely matched the K-Means findings.
Additional 2D scatterplots stratified by cluster (such as Income vs. Customer Age and Marital Status vs. Expenses) showed a consistent separation: middle-aged, high-income customers with a relationship status spent more than their lower-income, single-status counterparts.
5. Cluster Interpretation
Two major customer segments were consistently identified by the clustering algorithms:

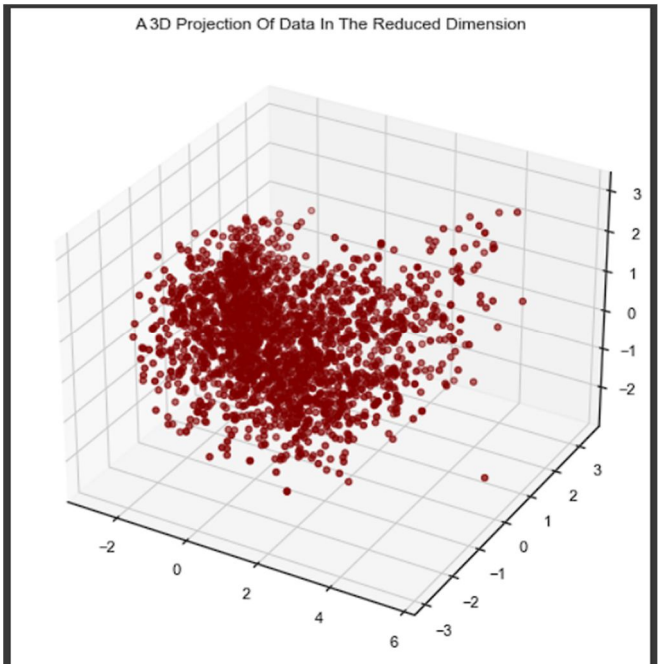
High Value Segment A:
• More likely to accept marketing campaigns;
• Slightly older (mid-40s to 50s);
• More likely to be in a relationship;
• Higher income;
• Higher expenses

Low Value Segment B:
Reduced income, low spending, poor marketing responsiveness, extremes in age, and a higher chance of being unmarried
Actionable insights for focused marketing are revealed by these segments:
• Customers in Segment A may be given preference for loyalty plans, upselling campaigns, or promotions of premium products.
• Rather than costly campaigns, Segment B might react better to entry-level products, discount offers, or retention tactics.

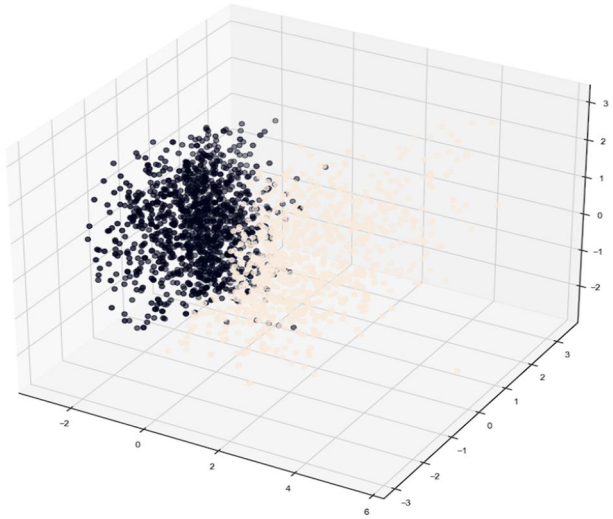
E. Analysis and Consequences

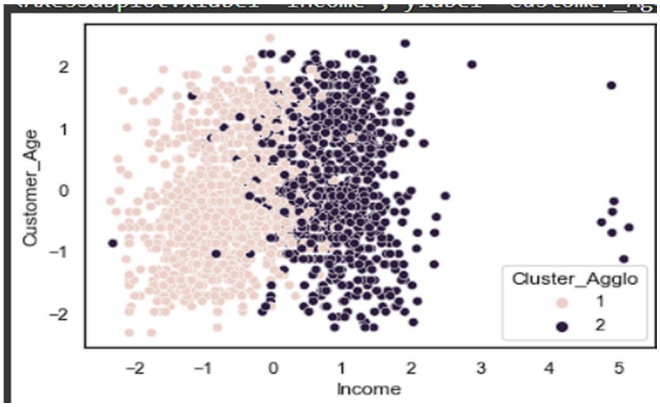
The customer base was effectively divided into significant groups in line with corporate goals thanks to the clustering analysis. Consistent patterns were produced by both clustering techniques (K-Means and Agglomerative), confirming the findings' resilience.
One noteworthy finding was that income and marketing responsiveness had a greater impact on spending than age or tenure. Tenure did not significantly correlate with expenses, indicating that behavioral segmentation may perform better than traditional demographic segmentation, despite presumptions that older, devoted customers spend more.
Additionally, the analysis was made simpler without sacrificing discriminatory power by combining the categories in Education and Marital_Status, demonstrating that simplified categorical encoding can preserve insights while lowering complexity.
Last but not least, dimensionality reduction through PCA made it possible to visualize complicated data in three dimensions without suffering a substantial loss of variance, which made it easier to understand the results of clustering.




A 3D Projection Of Data In The Reduced Dimension


Distribution Of The Clusters
count
1200
1000
800
600
400
200
0
1
2
Cluster_Agglo





V. CONCLUSION & FUTURE SCOPE

Analysis of Customer Personality Using Machine Learning: Conclusion and Prospects

In summary: Converting Information into Useful Knowledge (approximately 250 words)

Customer personality analysis based on machine learning (ML) is a paradigm shift away from conventional market segmentation. ML techniques use complex, multi-dimensional customer data to reveal latent behavioral patterns and unique psychological profiles, as opposed to depending on static, broad demographic categories.

Unsupervised learning algorithms, particularly clustering (e.g., K-Means, Hierarchical Clustering, DBSCAN), usually form the basis of this analysis. To create customer segments that are both internally homogeneous and externally heterogeneous, these algorithms analyze features obtained from purchase history (frequency, recency, monetary value), engagement metrics (clicks, time spent), and channel preference.

Several crucial business outcomes are produced by the effective use of ML in this field:

Hyper-Personalization: It enables companies to customize user experiences, product recommendations, and marketing messaging for each market segment, going beyond mass communication to provide pertinent, customized interaction.

Better Resource Allocation: Businesses can more efficiently allocate their marketing and retention budgets by precisely identifying high-value and at-risk segments, which significantly raises Return on Investment (ROI).

Strategic Product Development: Product teams use personality insights to guide the development of features and products that are sure to be in demand by learning about the unmet needs and preferences of important customer groups.

Enhanced CLV: By anticipating and reducing churn, an understanding of customer personality makes it possible to implement proactive retention tactics that target particular behavioral triggers.

Business strategy becomes incredibly customer-centric as a result of machine learning (ML), which essentially offers the quantitative tools required to convert vast amounts of customer data into a clear, data-driven understanding of the "who" and "why" behind purchase decisions.

Future Scope: Personality Analysis's Development (approximately 250 words)

There will likely be a lot of innovation in the future of ML-based customer personality analysis, leading to more dynamic, predictive, and morally sound systems:

Sequential Modeling and Time Series Data In order to analyze the context and sequence of customer actions over time, future models will make extensive use of Deep Learning architectures, including Transformer networks and Recurrent Neural Networks (RNNs). This will make it possible to anticipate changes in personality, anticipate client needs before they materialize, and transition from static segmentation to dynamic personality tracking.

Integration of Unstructured Data through NLP: The next generation of models will significantly integrate unstructured sources, whereas the current models primarily rely on structured transaction data. Customer reviews, social media posts, and support transcripts will all have their sentiment, tone, and semantic content examined by natural language processing (NLP). By connecting transactional behavior with customer attitudes and emotional states, this adds a deeper, psychological layer to the analysis.

Explainable AI (XAI) and Ethical Segmentation: Explainable AI (XAI) will become required as these models become more important in business decisions. Future systems must explain in detail how the model arrived at a particular recommendation and why a customer was assigned to a particular segment. In order to comply with regulations, reduce algorithmic bias against protected groups, and increase consumer confidence in personalized

REFERENCES

[1] E. Saraf, S. Pradhan, S. Joshi, and S. Sountharrajan, "Behavioral Segmentation with Product Estimation using K-Means Clustering and Seasonal ARIMA," in Proc. 6th Int. Conf. Trends Electron. Informatics (ICOEI), 2022, pp. 1641–1648. doi: 10.1109/ICOEI53556.2022.9776834

[2] Z. Li, J. Wang, and X. Yang, "A Supervised Clustering Approach for Customer Segmentation in Commercial Banks Based on Customer Value," in Proc. 2024 Int. Conf. Comput. Sci. Eng. Technol., Apr. 2024. [Online]. Available: https://www.researchgate.net/publication/382663007_A_Supervised_Clustering_Approach_for_Customer_Segmentation_in_Commercial_Banks_based_on_Customer_Value

[3] E. Saraf, S. Pradhan, S. Joshi, and S. Sountharrajan, "Behavioral Segmentation with Product Estimation using K-Means Clustering and Seasonal ARIMA," in Proc. 6th Int. Conf. Trends Electron. Informatics (ICOEI), 2022, pp. 1641–1648. doi: 10.1109/ICOEI53556.2022.9776834.

[4] S. Manimozhi, D. Ruby, and K. Biruntha, "Comparing Supervised and Unsupervised Learning Technologies for Customer Segmentation in Marketing," in Proc. 2024 Int. Conf. Comput. Sci. Eng. Technol., Apr. 2024. doi: 10.1109/ICONSTEM60960.2024.10568702.

[5] K. Gopalakrishnan, "Customer Segmentation Using K-Means Clustering for Targeted Marketing in Banking," Int. J. Artif. Intell. Mach. Learn., vol. 3, no. 2, pp. 89–94, 2024. [Online]. Available: https://iaeme.com/Home/article_id/IJAIML_03_02_006

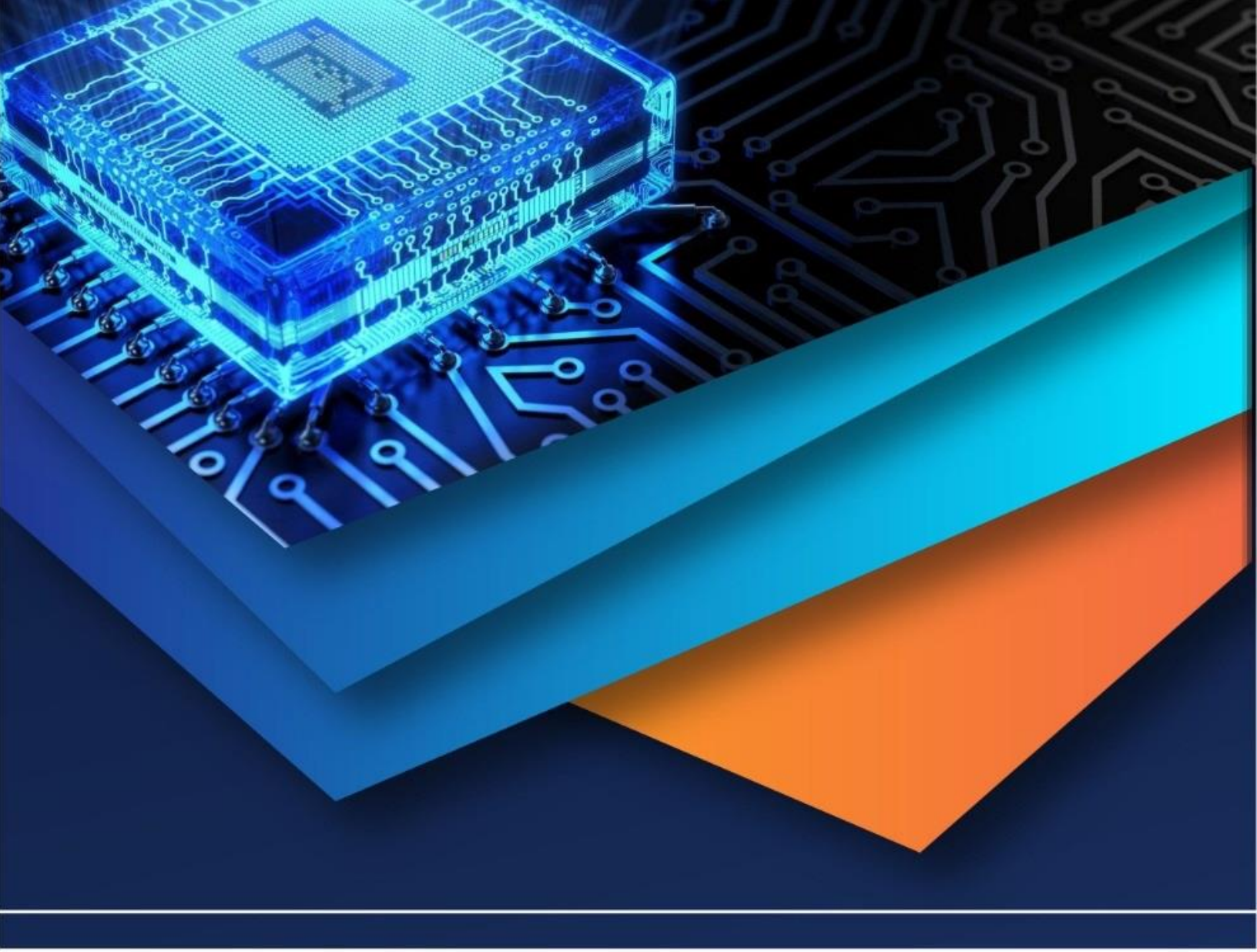
[6] Y. Chang and C. Tsai, "Customer Segmentation and Clustering Techniques," J. Bus. Res., vol. 58, no. 6, pp. 735–742, 2005. [Online]. Available: https://www.researchgate.net/publication/221550401_Research_on_customer_segmentation_model_by_clustering

[7] C. Wang, "Efficient Customer Segmentation in Digital Marketing Using Deep Learning with Swarm Intelligence Approach," Inf. Process. Manag., vol. 59, no. 6, p. 103085, 2022. doi: 10.1016/j.ipm.2022.103085.

[8] L. Behera, P. Nanda, S. Patnaik, and B. Panda, "Machine Learning for Customer Segmentation Through Bibliometric Approach," in Proc. 2020 Int. Conf. Mach. Learn. Cybern., 2020. [Online]. Available: https://www.semanticscholar.org/paper/Machine-Learning-for-Customer-Segmentation-Through-Behera-Nanda/cd8f2d03e5abd972656ba022f22dce354415918aHoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2015). Botnet in DDoS attacks: trends and challenges. IEEE Communications Surveys & Tutorials, 17(4), 2242-2270.

[9] Kumar and R. Singh, "Comparative Analysis of Machine Learning Models for Customer Segmentation," Int. J. Comput. Appl., vol. 182, no. 9, pp. 1–5, 2024. [Online]. Available: https://www.researchgate.net/publication/371186145_Comparative_Analysis_of_Machine_Learning_Models_for_Customer_Segmentation

[10] K. Gopalakrishnan, "Customer Segmentation Using K-Means Clustering for Targeted Marketing in Banking," Int. J. Artif. Intell. Mach. Learn., vol. 3, no. 2, pp. 89–94, 2024. [Online]. Available: https://iaeme.com/Home/article_id/IJAIML_03_02_006



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY