iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** VI **Month of publication:** June 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.53911

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 11 Issue VI Jun 2023- Available at www.ijraset.com*

# An Unsupervised Learning Based System Employing K-Means Clustering to Perform Customer Segmentation

Atharva Gupta[1], Dr. Pooja Khanna[2]
*Department of Computer Science & Engineering, Amity School of Engineering and Technology, Amity University, Lucknow Campus*

***Abstract:*** ***The job of marketers is to get the right product in front of the right consumer at the moment that they're most likely to buy that product. The ability for marketers to do just that and to drill down to more and more specific niches of customers has grown exponentially over time.***
***AI technology is now being used to help marketers get even more specific with predictive targeting and personalization. Targeted advertising is a form of online advertising which micro-targets its customers. It is based on the traits and behavioral patterns of different people.***
***Nowadays, people, knowingly or unknowingly, are churning out personal data at an unprecedented scale because of the use of all electronic devices.***
***Targeted advertising has been gaining importance ever since the start of the century. This is because people are becoming more and more diversified and it is not feasible or even possible to fit them all in one campaign. Thus, the organization today are turning towards the concept of Targeted advertising as a way to boost their revenues.***
***Keywords: Machine Learning, Unsupervised Learning, Targeted Advertising, Feature Extraction,Customer Segmentation,K-means clustering.***

## I. INTRODUCTİON

The advertising industry has been going through a period of significant transition, with digital marketing strategies rapidly replacing older ways. One of the most important aspects of digital marketing is targeted advertising, which is sending personalized messages to certain audiences on the basis of their demographics, interests, behaviors, and other data elements. This constitutes one of the most important aspects of targeted advertising. It has been demonstrated that targeted advertising is more effective than traditional types of advertising. This is due to the fact that focused advertising enables more effective communication with the audiences who are being targeted.

During the course of the past few years, machine learning has emerged as an effective method that has the potential to be applied in the production of tailored advertising. Algorithms that are designed to learn through machine learning are capable of analyzing vast volumes of data, picking up on patterns and trends as they go, and then making predictions based on the information they have learned.

Because Python is one of the most widely used programming languages for machine learning, developers have access to a huge selection of tools and frameworks. This helps developers to construct advertising systems that are trustworthy and effective.

The purpose of this research is to investigate, from the point of view of machine learning, the ways in which Python is used to the practice of targeted advertising. Following introductory discussions of the terms "targeted advertising" and "machine learning," the remainder of this article will go into the fundamental ideas and approaches that are utilized in each of these respective fields. Following this, the article will investigate the ways in which these two technologies may be utilized to produce advertising campaigns that are more successful.

The paper is set up as follows. The Introduction is Section 1, and the Literature Review, which summarizes and provides the results of earlier work that served as the background research for this paper, is Section 2. The Methodology used for this project and a flowchart of the suggested procedure are both found in Section 3

The Experimental Analysis in Section 4 , the Result and Dicussion is highlighted in Section 5 and Section 6 is where this paper reaches its Conclusion.

## II. LİTERATURE REVİEW

The authors [1] describe a three-layer architecture for the targeting of advertisements, which includes the processes of extracting features from the data, preparing the data, and classifying the information using machine learning. Each layer is responsible for a distinct aspect of the overall process of targeting. The challenges that are linked with ad targeting, such as the vast amount of data, the wide range of user behaviors, and the requirement for real-time processing, have been addressed by the architecture that has been established.

In their work titled "A Deep Learning-Based Personalized Advertising Recommendation System" the authors [2] demonstrate a deep learning-based system that may be used to provide individualized advertising suggestions. This strategy involves making use of data relating to the activities of users in order to create predictions about the preferences of those users.

Implementing Machine Learning Algorithms for Targeted Advertising in Python is the title of a book written by [3] in which the authors [3] discuss machine learning algorithms for targeted advertising that are based on Python.The authors discuss a wide variety of machine learning strategies and then illustrate how to use Python to put those strategies into action. Logistic regression, decision trees, random forests, and gradient boosting are some of the methods that fall under this category.

An in-depth analysis of the effect that AI has had on advertising can be found in Yimam and Alqahtani's essay titled "The Impact of AI on Advertising" (2020), which was published in 2020. The authors [4] discuss the potential benefits of AI in advertising, some of which include greater targeting and personalization, higher efficiency and efficacy of advertising campaigns, enhanced ad creativity, and improved ad targeting. The article [5] presents a complete overview of the ethical difficulties that are linked with targeted advertising.The paper by [5] offers a comprehensive analysis of the moral concerns that are connected to tailored advertising. The authors conducted a search of the relevant published literature and analysed 41 papers that discussed ethical concerns about tailored advertising. Privacy, openness, fairness, autonomy, and informed consent were determined to be the five most important ethical concerns by the writers.The article contains a discussion of each of these ethical dilemmas, highlighting the difficulties that are involved with each of them.

The study by [6] provides a comprehensive review of sixty-five research articles that focus on the application of ML in the advertising industry. The authors highlighted three primary applications of machine learning in the advertising industry, which are as follows: ad targeting; ad optimisation; and ad creation. The paper by [6] draws attention to a number of difficulties that are connected to the application of ML in advertising. The requirement for enormous volumes of data, the possibility of algorithms including bias, and the requirement for transparency and accountability in the application of machine learning in advertising are some of these problems.

The writers [7] address the possible benefits of artificial intelligence in advertising, some of which include increased targeting and personalisation, higher efficiency and efficacy of advertising campaigns, enhanced ad inventiveness, and improved ad targeting. They also study the obstacles that are brought by the application of AI in advertising, such as the demand for qualified employees in order to manage and optimize AI-based advertising campaigns, concerns surrounding the protection of data, and ethical considerations that need to be taken into account.

In the article [8], we analyze the difficulties that are connected with using traditional techniques of advertising, as well as how machine learning could assist to overcome these difficulties. According to the authors, machine learning algorithms are able to do analysis on massive volumes of data in real time, which enables marketers to produce advertisements that are more personalized and pertinent to the individual.

The paper [9] provides an examination of the similarities and differences between various machine learning algorithms for targeted advertising. An overview of targeted advertising and the importance of machine learning in this particular market are given in the article's first section. The authors [9] then go on to give a succinct overview of the various machine learning algorithm categories and how these algorithms are applied to targeted advertising.

The authors [10] conduct a comprehensive analysis of a wide range of papers and research publications that investigate the application of machine learning algorithms to advertising. The analysis focuses on the benefits and drawbacks of various machine learning algorithms, as well as how these algorithms might be used to the field of advertising.

## III. METHODOLOGY

1) *Data Collection:* The first step of the advertising campaign is to collect data on the demographics of the target audience. This data may consist of demographic information, data pertaining to transactions, behaviour pertaining to web browsing, interactions with social media, etc. The information can be obtained from a wide variety of resources, including customer relationship management (CRM) systems, social media platforms, Google Analytics, surveys, and so on.

2) *Data Preparation:* In order to analyse the acquired data, they must first be cleansed, then processed, and finally prepared. This entails removing duplicates, missing numbers, and outliers from the data, in addition to transforming it into a format that is acceptable for analysis. The dataset was properly prepared and did not contain any NA values which can be seen in the below Figure [3.1]. Therefore, we may get started by shaping the characteristics. In the later steps, we will calculate three characteristics for each customer_id, and those features will assist us with the visualisation (by making use of the Plotly library), as well as the explainability of the process. Pandas and numpy will be utilised in order to complete the data preparation.

```
In [3]: # first rows of the dataset
        customers_orders.head()
```

Out[3]:

| | product_title | product_type | variant_title | variant_sku | variant_id | customer_id | order_id | day | net_quantity | gross_sales | discounts | returns |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | DPR | DPR | 100 | AD-982-708-895-F-6C894FB | 52039657 | 1312378 | 83290718932496 | 04/12/2018 | 2 | 200.0 | -200.00 | 0.00 |
| 1 | RJF | Product P | 28 / A / MTM | 83-490-E49-8C8-8-3B100BC | 56914686 | 3715657 | 36253792848113 | 01/04/2019 | 2 | 190.0 | -190.00 | 0.00 |
| 2 | CLH | Product B | 32 / B / FtO | 68-ECA-BC7-3B2-A-E73DE1B | 24064862 | 9533448 | 73094559597229 | 05/11/2018 | 0 | 164.8 | -156.56 | -8.24 |
| 3 | NMA | Product F | 40 / B / FtO | 6C-1F1-226-1B3-2-3542B41 | 43823868 | 4121004 | 53616575668264 | 19/02/2019 | 1 | 119.0 | -119.00 | 0.00 |
| 4 | NMA | Product F | 40 / B / FtO | 6C-1F1-226-1B3-2-3542B41 | 43823868 | 4121004 | 29263220319421 | 19/02/2019 | 1 | 119.0 | -119.00 | 0.00 |

Figure [3.1] : Dataset rows

3) *Client Segmentation:* To segment the client base into groups based on the qualities that are shared by those groups, machine learning algorithms such as clustering and decision trees should be utilised. This segmentation may be based on demographic information, transactional data, or behavioural data. All three types of data may also be used. It is important that the segmentation be carried out in a manner that makes it possible to effectively target advertising messages towards each individual category. The K-means technique included in scikit-learn is going to be utilised by us.
4) *Feature Selection:* Determine which features are most applicable to each consumer segment and choose them. Finding the characteristics that are the best predictors of client behaviour and preferences is a necessary step in this process. Techniques for feature selection could include things like principal component analysis (PCA), correlation analysis, or mutual information analysis.The K-means algorithm sees each row in the customer data frame as a point in a three-dimensional space and interprets them as such. When it comes to grouping them, it takes into account the Euclidean distance between each of the data points and the group's epicentre. If the ranges are quite different from one another, the algorithm might not work very well and might not be able to generate the groups as intended.In order to improve the efficiency of the K-means algorithm, we are going to logarithmically convert the data before scaling it. This is the kind of transformation that works well with skewed data. This will cause a proportional reduction in the amount of 3D space that our data is spread throughout while maintaining the same level of proximity between the spots.
5) *Predictive Modelling:* When developing a model that can anticipate the actions or preferences of customers, it is important to make use of predictive modelling strategies such as logistic regression, decision trees, and neural networks. Utilising this technique allows for the generation of personalised marketing messages that are tailored to each customer category.
We are going to use K-means algorithm from scikit-learn.
Let's go down how the algorithm will divide up the customers into their respective groups:

*a)* Determine the initial k = n centroids = number-of-clusters either randomly or intelligently.
*b)* Using the euclidian distance as a guide, assign each data point to the centroid that is geographically closest to it, therefore creating the groupings.
*c)* Adjust the centre points such that they are equal to the average of all the points in cluster.
*d)* It is necessary to repeat steps 2 and 3 until convergence is reached.

We are going to use the elbow approach to select k, and we are going to decide against the optimisation criteria of the K-means algorithm, which is inertia. We are going to construct various K-means models with k values ranging from 1 to 15, and we are going to save the inertia values that correspond to those models.

**Hyperparameter tuning: Find optimal number of clusters**

```
In [94]: def make_list_of_K(K, dataframe):
             '''inputs: K as integer and dataframe
             apply k-means clustering to dataframe
             and make a list of inertia values against 1 to K (inclusive)
             return the inertia values list
             '''

             cluster_values = list(range(1, K+1))
             inertia_values=[]

             for c in cluster_values:
                 model = KMeans(
                     n_clusters = c,
                     init='k-means++',
                     max_iter=500,
                     random_state=42)
                 model.fit(dataframe)
                 inertia_values.append(model.inertia_)

             return inertia_values
```

Figure [3.2] : Hyperparameter Tuning

With the elbow method, As shown in the above Figure [3.2], we are going to select the k value where the decrease in the inertia stabilizes.When k=1 inertia is at the highest, meaning data is not grouped yet. Inertia decreases steeply until k=2. Between k=2 and 4, the curve continues to decrease fast.
At k=4, the descent stabilizes and continues linearly afterwards, forming an elbow at k=4. This points out the optimal number of customer group is 4.

6) *Visualization and Interpretation of the Results:* We will plug in the k=4 to K-means and visualize how customer groups are created:

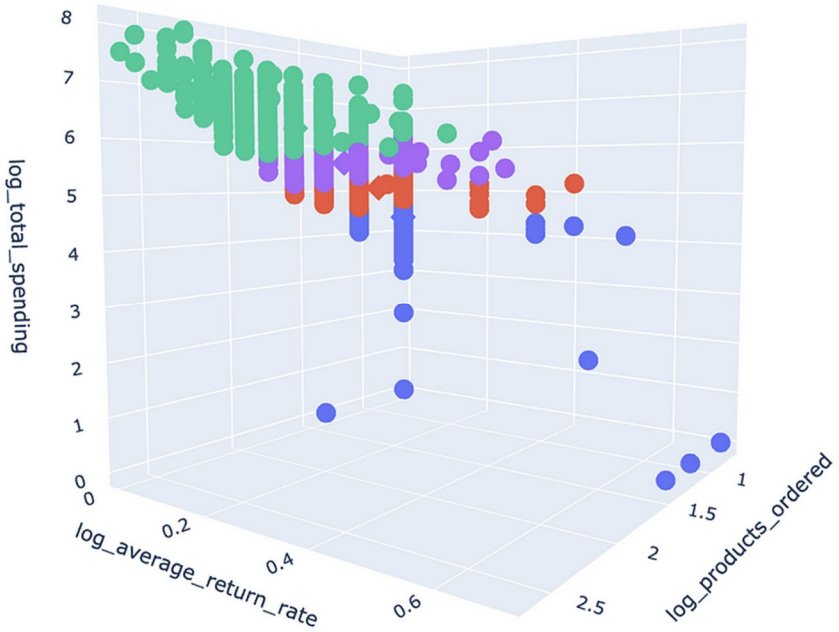



Figure [3.3] : 3D plot with centroids

As shown in the above Figure [3.3], The centroids of each group are displayed using cubes, while data points are displayed as spheres. The following four customer groups are:

*a)* *Blue:* Customers that made at least one purchase, spent up to a maximum of 100 dollars, and had the highest average return rate. They could be brand-new users of the online store.
*b)* *Purple:* Customers who ordered 1 to 4 things, with an average total expenditure of 300 and a maximum return rate of 0.5.
*c)* *Red:* Red represents customers who ordered 1 to 4 products and spent an average of 150 overall 0.5.
*d)* *Green:* Customers who purchased 1 to 13 items, on average spending 600 and returning 0 items on average. It creates the most advantageous consumer base for the business.

We will now look how many customers are there in each group — known as cluster magnitudes:

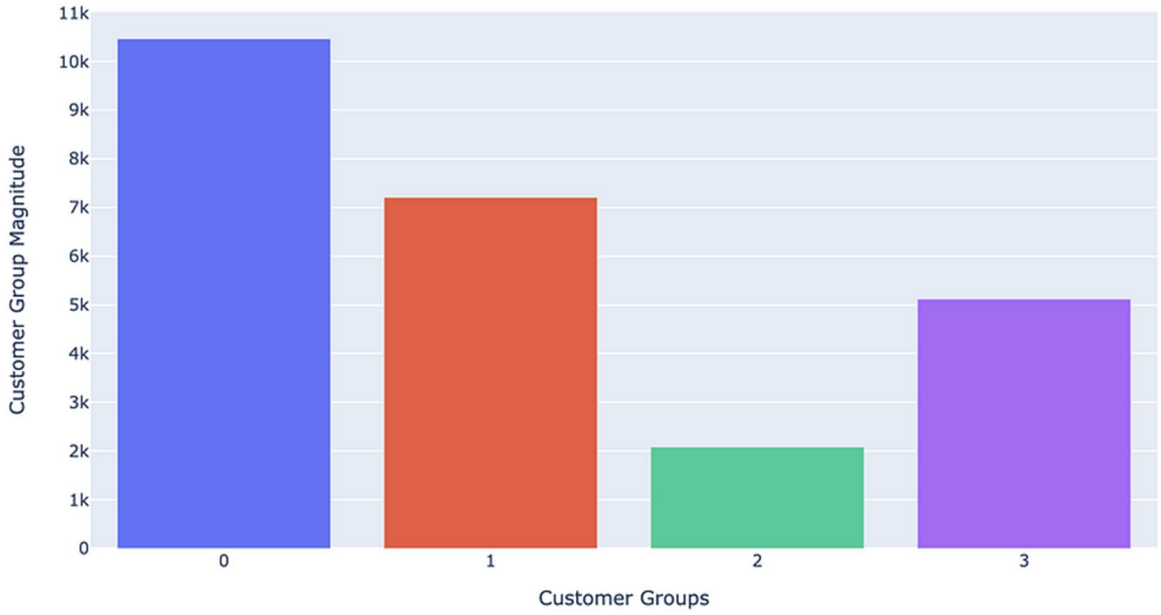



Figure [3.4] : Graph representing Customer Magnitudes

*7)* *Campaign Execution:* Execute the advertising campaign by creating personalized messages for each customer segment. The messages should be tailored to the preferences and behavior of each segment to maximize their effectiveness.
*8)* *Campaign Assessment:* Utilise data like as click-through rates, conversion rates, and return on investment (ROI) to assess the efficacy of the advertising campaign. Refine the segmentation and modelling strategies for upcoming campaigns after you've analysed the results to find areas that still need work.

## IV. RESULTS AND DİSCUSSİON

As can be seen from figure [3.4], The main plan would be to keep the green customer group, which is the most beneficial, while relocating the blue consumer group to the red and purple sectors.

42% of all customers are in the blue group; any enhancements made in this client group will significantly increase revenue. This consumer group can be moved to a low-average return-rate and high-total-spending area by eliminating high return rates and providing gift cards. Gift vouchers can hasten their return if we presume that they are newcomers.

The combined red and purple group includes 50% of all customers. From the views of the typical return rate and the products ordered, they exhibit the same traits, but their spending patterns are different. These people can be classified as those who have purchased several products and are familiar with the brand. With certain specialised messaging and promotions, the company may keep these clients informed about the product.

The most positive customer category for the company is made up of the 8% of customers who identify as green. They place many orders and are very likely to keep all of the items. Special offers and pre-product debuts may help to retain and conceivably grow this group. Additionally, they might attract new clients, which would increase the clientele.

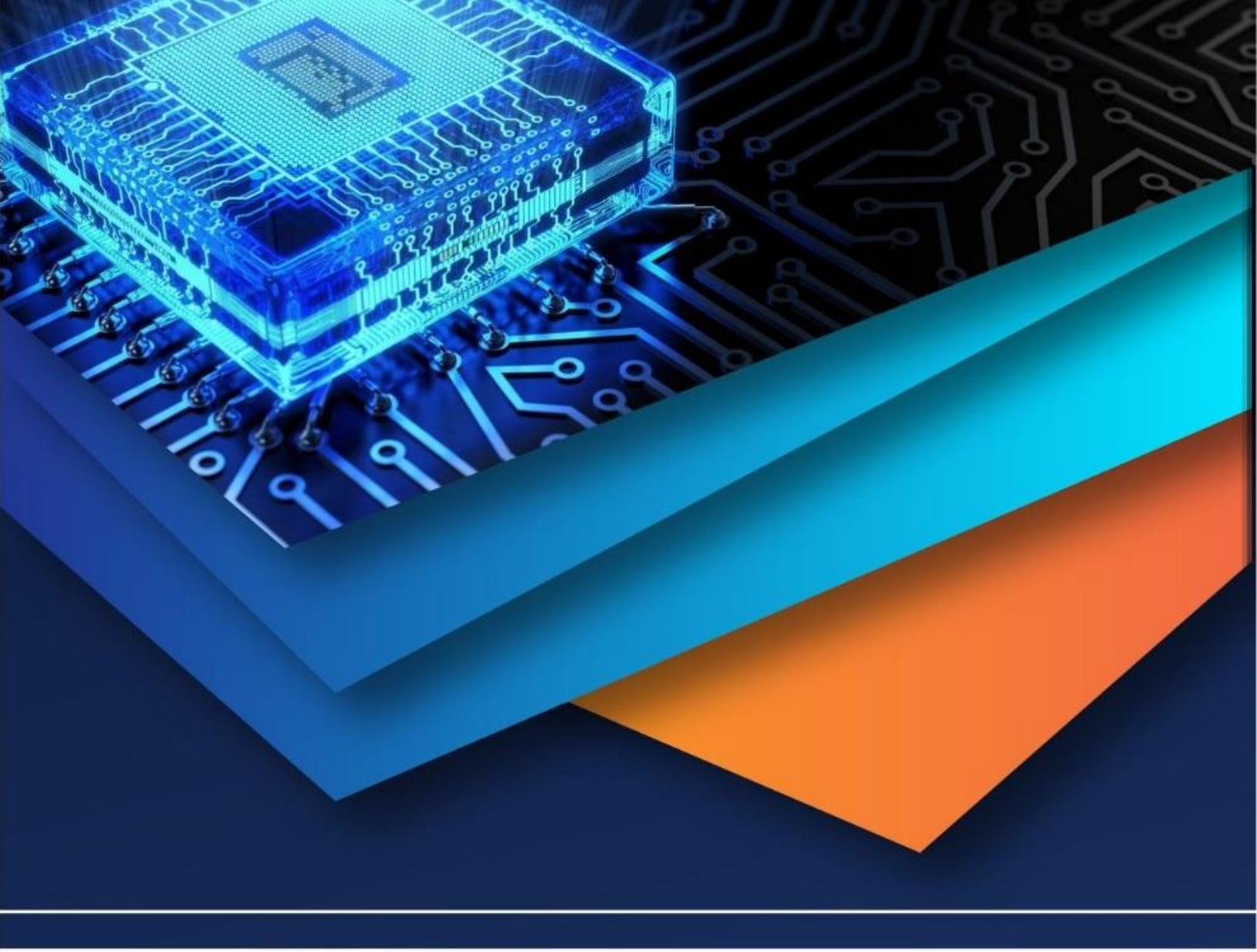
## V. CONCLUSİON AND FUTURE SCOPE

It is not a good idea to provide all of your clients with the same model of product, email campaign, text message campaign, or advertisement. Customers have a variety of requirements. A company strategy that takes the "one size fits all" approach will, in most cases, produce lower levels of engagement, lower click through rates, and eventually fewer sales. The solution to this issue is to divide customers into different groups.

Discovering the optimum quantity of distinct customer groups will assist you in comprehending the ways in which your clients differ and will enable you to supply them precisely what it is that they want. Customer segmentation results in an enhanced experience for the customer as well as more revenue for the organisation. Because of this, segmentation is an absolute necessity if you want to outperform your competitors and get a greater number of clients. Using machine learning to do this task is undeniably the best course of action to take.

## REFERENCES

[1] Liu, J., Jiang, S., Wei, Y., & He, X. (2020). An Intelligent Ad Targeting Framework with Machine Learning. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 2114-2121). IEEE.
[2] Zou, C., Zhang, X., Wang, Y., & Chen, L. (2021). A Deep Learning-Based Personalized Advertising Recommendation System. IEEE Access, 9, 57948-57958.
[3] Bera, S., & Varma, V. (2021). Implementing Machine Learning Algorithms for Targeted Advertising in Python. In Proceedings of the 5th International Conference on Intelligent Computing and Control Systems (pp. 797-801). Springer.
[4] Yimam, S. M., & Alqahtani, S. (2020). Impact of Artificial Intelligence on Advertising. Journal of Management and Business Administration, 28(1), 16-31.
[5] Esposito, M., Ficco, M., Palmieri, F., & Castiglione, A. (2020). Ethical Issues in Targeted Advertising: A Systematic Review. Journal of Business Research, 113, 278-288.
[6] Rasool, H., Alshabandar, R., & Hussain, A. (2019). Machine Learning in Advertising: A Systematic Literature Review. In 2019 8th International Conference on Industrial Technology and Management (ICITM) (pp. 36-41). IEEE.
[7] "Targeted Advertising Using Machine Learning" by A. L. A. M. Abood, A. A. Zaidan, B. B. Zaidan, and M. S. Saliem. In Proceedings of the 2020 2nd International Conference on Computer Applications & Information Security (ICCAIS), pp. 1-7, 2020.
[8] "Machine Learning for Targeted Advertising: A Review" by V. Bhatia, R. Arora, and R. Gupta. In Proceedings of the 2020 International Conference on Inventive Systems and Control (ICISC), pp. 2143-2148, 2020.
[9] "A Comparative Study of Machine Learning Algorithms for Targeted Advertising" by S. S. Mujawar and S. S. Kadam. In Proceedings of the 2018 3rd International Conference on Inventive Communication and Computational Technologies (ICICCT), pp. 1140-1145, 2018.
[10] "Python Machine Learning for Advertising: Predictive Modelling and Data Mining Techniques" by M. Vlahopoulou and V. Iosifidis. In Proceedings of the 2019 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 834-839, 2019.
[11] "Identifying machine learning techniques for classification of target advertising" by A. Sarkar, A. Ghosh, and S. Bandyopadhyay (2020)
[12] "The Impact of Customer Segmentation on Targeted Advertising Effectiveness" by David Bell, published in the journal "Journal of Marketing" in 2005.
[13] "The Use of Customer Data for Targeted Advertising: A Privacy Perspective" by David Hoffman, published in the journal "Journal of Consumer Research" in 2010.
[14] "The Future of Targeted Advertising" by Richard J. Lachmann, published in the journal "Journal of Advertising Research" in 2017.
[15] "Customer Segmentation and Targeted Advertising: A Review of the Literature" by Jie Zhang, published in the journal "Decision Support Systems" in 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY