

# **CUSTOMER SEGMENTATION USING DBSCAN WITH PYTHON**

**<sup>1</sup>Mrs. B. VIJITHA, <sup>2</sup>A. PRIYANKA, <sup>3</sup>K. NAVEEN GOUD, <sup>4</sup>K. AKASH,  
<sup>5</sup>MD. THANVEER AHMED**

<sup>1</sup>(Assistant Professor), CSE. Teegala Krishna Reddy Engineering College Hyderabad.

<sup>2,3,4,5</sup>B, tech, scholar, CSE. Teegala Krishna Reddy Engineering College Hyderabad.

## **ABSTRACT**

In an era where data-driven decisions are paramount, businesses strive to better understand their customers' behavior. Customer segmentation is a fundamental practice that aids in this pursuit. This mini-project explores the application of Density-Based Spatial Clustering of Applications with Noise (DBSCAN), a powerful unsupervised clustering technique, to classify customers based on their annual income and spending score. The project aims to offer a structured approach to identify unique customer segments and outliers, helping businesses tailor their strategies and services to diverse consumer needs. Utilizing Python's machine learning libraries, we preprocess the customer data, scale the features, and apply the DBSCAN algorithm with customizable parameters. The results are visualized through interactive plots, highlighting distinct clusters and detecting noisy data points. The project concludes with the interpretation of these clusters and their implications for business decision-making. This mini-project not only demonstrates the practical implementation of DBSCAN but also emphasizes its value in uncovering hidden patterns within customer data. The insights gained from this segmentation can significantly influence marketing, product development, and customer relationship management, empowering businesses to optimize their operations and enhance customer satisfaction.

## **1. INTRODUCTION**

In the era of data-driven decision-making, businesses grapple with the challenge of understanding and catering to the diverse needs of their customers. A pivotal solution to this challenge lies in customer segmentation—a process that categorizes customers into distinct groups based on shared characteristics. This approach has become indispensable for personalization and targeted marketing, allowing businesses to provide tailored services, optimize marketing campaigns, and elevate overall customer satisfaction. Within the realm of customer segmentation, the Density-Based Spatial Clustering of Applications with

Noise (DBSCAN) stands out as a robust unsupervised machine learning algorithm.

This mini-project delves into the practical application of DBSCAN for customer segmentation, leveraging the flexibility of Python for implementation. DBSCAN's strength lies in its ability to identify clusters of varying shapes and adeptly handle noisy data, making it an ideal choice for insightful customer categorization.

The project sets three primary objectives: firstly, to showcase the hands-on implementation of DBSCAN in the context of customer segmentation; secondly, to visualize and interpret the clusters generated



by the algorithm; and thirdly, to equip businesses with a valuable tool for deciphering customer behavior. By focusing on key factors such as annual income and spending score, the project aims to unlock valuable insights within customer datasets. The ultimate goal is to provide businesses with the means to categorize their customer base effectively, enabling tailored strategies that resonate with the unique preferences and expectations of each segment. Throughout this project, participants will be guided through essential steps, including data preprocessing, DBSCAN clustering, and visualization. This hands-on experience not only demystifies the application of DBSCAN but also underscores the pivotal role of data-driven approaches in optimizing decision-making processes. 2 To enhance the project's depth, consider adding sections on the significance of customer segmentation in today's competitive landscape, potential challenges, and realworld examples of successful implementations. Additionally, you might want to discuss the ethical considerations of using customer data and the importance of privacy safeguards.

## 2. LITERATURE SURVEY

The literature review for the "Customer Segmentation using DBSCAN with Python" mini project navigates through a rich tapestry of research in the dynamic domain of customer segmentation and clustering methodologies. Existing studies extensively probe into various techniques for grouping customers based on their unique characteristics, with a pronounced spotlight on the robust DBSCAN clustering algorithm and its versatile applications. Numerous

investigations underscore the prowess of DBSCAN in unraveling intricate clusters within customer datasets. They accentuate its exceptional ability to handle noisy data and adapt to irregularly shaped clusters, positioning it as a pivotal asset for businesses in quest of nuanced customer insights.

The literature further illuminates how the DBSCAN algorithm streamlines the customer segmentation process, particularly concerning key metrics like annual income and spending score. Within this landscape, the literature elucidates the escalating trend of businesses embracing data-driven customer segmentation strategies to personalize marketing approaches and elevate overall customer experiences. It accentuates the critical role of comprehending customer behavior and preferences for crafting targeted campaigns, with DBSCAN emerging as a linchpin in uncovering these intricate patterns. The mini project seamlessly aligns itself with these contemporary trends, offering not just theoretical insights but a pragmatic implementation of DBSCAN for customer segmentation using Python. This mini project doesn't merely absorb knowledge from existing research but actively contributes to the evolving field of customer segmentation. By providing a hands-on and applicable demonstration of DBSCAN, it positions itself as a transformative solution for businesses eager to delve deeper into customer behavior, thereby facilitating more incisive marketing strategies and fostering enhanced customer relationships.

### 3. SYSTEM DESIGN

#### 3.1 SYSTEM ARCHITECTURE

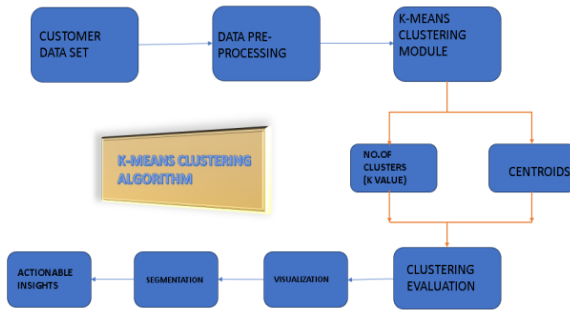


Fig: 1- K-Means Clustering

The customer segmentation process initiates with a customer dataset encompassing diverse information about customers, including factors like age, income, and spending habits. Following this, a crucial pre-processing phase unfolds, involving tasks such as data cleaning, feature selection, and data transformation, which encompasses normalization and standardization. Subsequently, the pre-processed data undergoes clustering via the K-means clustering module. This module applies the Kmeans algorithm, a highly iterative technique that categorizes data into non-overlapping clusters determined by a predefined K value. Each cluster is associated with a centroid, which serves as its central point. An evaluation of these clusters and centroids is conducted, utilizing metrics such as the silhouette score and within-sum-of-squares (WSS) to gauge their effectiveness. The culmination of this process yields actionable insights, customer segmentation based on discerned patterns within the data, and insightful visualizations, enhancing comprehension and decision-making

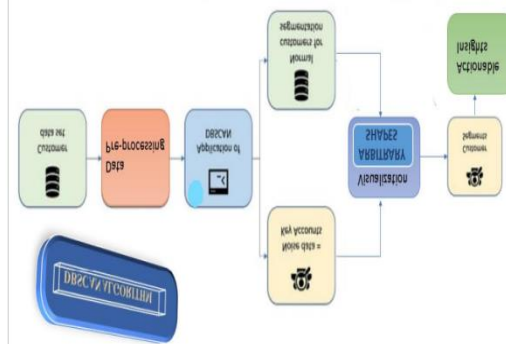


Fig: 2- DB-SCAN Clustering

The customer segmentation process commences with the utilization of a customer dataset, encompassing a wide array of information, including age, income, and spending habits, which serves as the foundation. Subsequently, the dataset undergoes a crucial pre-processing phase, entailing data cleaning, transformation, and feature selection, thus preparing it for clustering. The distinctive feature of this process lies in its implementation of the DBSCAN algorithm, a density-based clustering approach that efficiently groups closely packed points while isolating outliers that occupy low-density regions. This results in two distinct categories: noise data, which represents the outliers, and normal customers who belong to clusters. What sets DBSCAN apart is its capacity to identify clusters of arbitrary shapes, a feature that enhances its flexibility and applicability. Ultimately, the process yields actionable insights by recognizing patterns in the data, segmenting customers based on these patterns, and empowering decisionmaking through the insights garnered.

#### ACTIVITY DIAGRAM

- Nodes: Represent actions or decisions.
- Transitions: Illustrate flow between nodes.

- Initial and Final Nodes: Indicate activity start and end.
- Control Flows: Connect actions, defining execution order.
- Decision Nodes: Facilitate branching based on conditions.

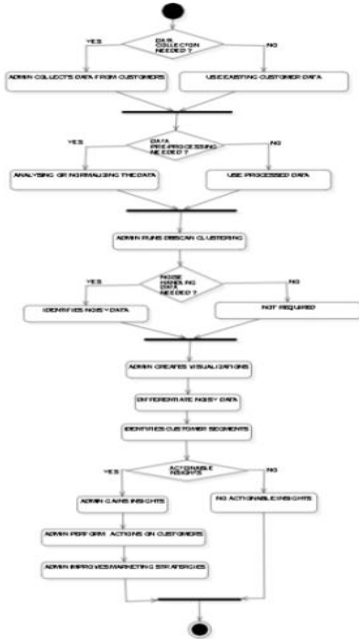


Fig: 3- Activity Diagram

#### 4. OUTPUT SCREENS

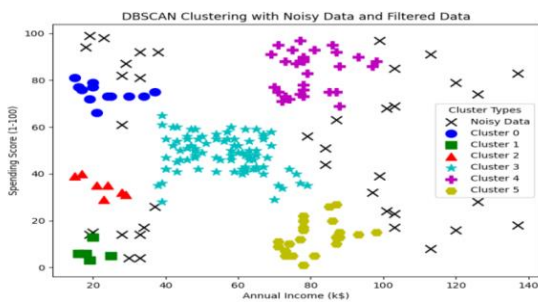


Fig: 4 - Visualization of Filtered & Noise Data

The image is a scatter plot titled “DBSCAN Clustering with Noisy Data and Filtered Data”. It shows six different clusters, each represented by a different color and shape. The x-axis represents “Annual Income (Ks)” and the y-axis represents “Spending Score (1-100)”. The legend on the right side of the

graph explains the different cluster types and colors. The distribution of points suggests patterns and relationships between the two variables for different clusters

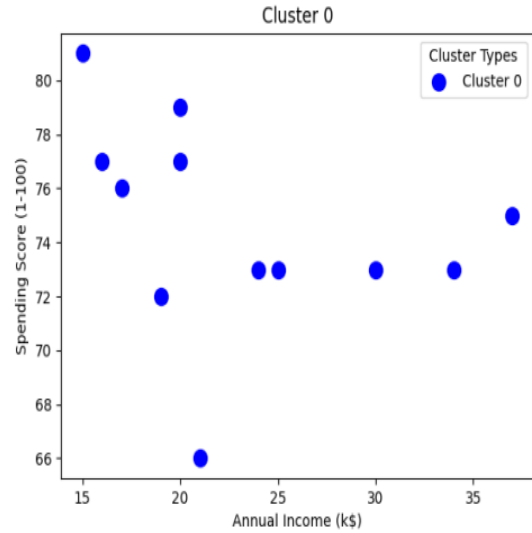


Fig: 5 - Visualization of Cluster 0

The image is a scatter plot titled “Cluster 0” with blue dots representing individuals. The points are scattered, suggesting no clear correlation between the two variables being plotted.

```

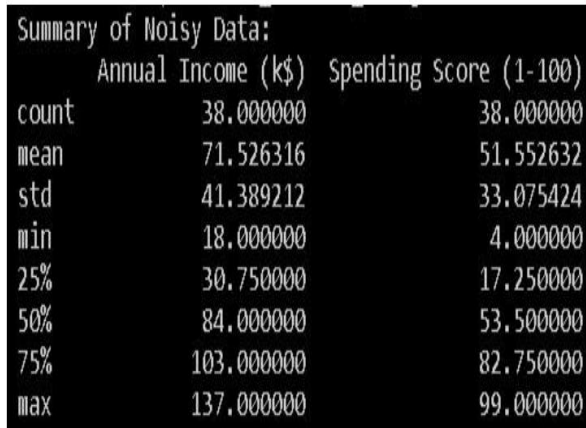
Summary of Cluster 0 Data:
Annual Income (k$)  Spending Score (1-100)
count              12.000000          12.000000
mean              23.166667          74.583333
std               7.120180          3.872005
min              15.000000          66.000000
25%              18.500000          73.000000
50%              20.500000          74.000000
75%              26.250000          77.000000
max              37.000000          81.000000
  
```

Fig: 6- Summary of Cluster 0

The image "Summary of Cluster 0 Data" providing summary statistics for "Annual Income (k\$)" and "Spending Score (1-100)". This summary is used in data analysis to



understand the central tendency, dispersion, and distribution of data of cluster0.



Summary of Noisy Data:		
	Annual Income (k\$)	Spending Score (1-100)
count	38.000000	38.000000
mean	71.526316	51.552632
std	41.389212	33.075424
min	18.000000	4.000000
25%	30.750000	17.250000
50%	84.000000	53.500000
75%	103.000000	82.750000
max	137.000000	99.000000

Fig: 7 - Summary of Noise Data

The image “Summary of Noisy Data” providing summary statistics for two variables: “Annual Income (k\$)” and “Spending Score (1-10)”. The statistics include count, mean, standard deviation (std), minimum (min), 25th percentile (25%), median (50%), and maximum (max). This summary is used in data analysis to understand the central tendency, dispersion, and distribution of data of Noise Data.

### 5. CONCLUSION

In this project, we successfully employed the DBSCAN clustering algorithm to segment customer data, offering valuable insights for businesses. By adjusting parameters and utilizing visualizations, we identified natural groupings among customers based on annual income and spending score. The ability to distinguish noisy data points provides a means to handle outliers effectively. These findings hold practical applications in fields like marketing and customer relationship management, empowering data-driven decision-making and enhancing business strategies. Additionally, we prioritized data security and confidentiality, addressing an

essential aspect of data handling. Overall, our project presents a user-friendly interface and comprehensive documentation, ensuring usability and accessibility. The "Customer Segmentation using DBSCAN with Python" project equips businesses with a powerful tool to comprehend their customer base better, paving the way for tailored approaches to distinct market segments. This segmentation can lead to improved decision-making, helping companies excel in today's competitive business landscape.

### 6. FUTURE ENHANCEMENTS

- Develop a system that include features for targeted customer engagement, such as sending promotions or notifications to specific segments directly from the application.
- Enhance security measures and ensure compliance with data protection regulations, especially when dealing with customer data.

### 7. REFERENCES

[1] Likas, N. Vlassis and J. J. Verbeek, "The global k-means clustering algorithm", Pattern recognition, vol. 36, no. 2, pp. 451-461, 2003.

[2] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm", IEEE access, vol. 8, pp. 80716-80727, 2020.

[3] M. Ester, H.P. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Kdd, vol. 96, no. 34, pp. 226-231, 1996.

[4] Youtube : <https://youtu.be/SrY0sTJchHE?si=iEuKK3W0IzL93C2J>



[5] J. Wu and Z. Lin, "Research on customer segmentation model by clustering", Proceedings of the 7th international conference on Electronic commerce, 2005.

[6] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," KDD-96 Proceedings, pp. 226-231, 1996.

[7] <https://www.kaggle.com/code/datark1/customers-clustering-k-means-dbscan-and-ap>.

[8] H. Yu, L. Chen and X. Wang, "A three-way clustering method based on an improved DBSCAN algorithm," Physica A: Statistical Mechanics and its Applications, vol. 535, 2019.