SASTRA

ENGINEERING · MANAGEMENT · LAW · SCIENCES · HUMANITIES · EDUCATION

DEEMED TO BE UNIVERSITY

(U/S 3 of the UGC Act, 1956)

THINK MERIT | THINK TRANSPARENCY | THINK SASTRA

THANJAVUR | KUMBAKONAM | CHENNAI

# Cluster-Based Analysis on Online Shoppers Purchasing Intention

July - November 2025

**Submitted By,**
Mukeshkumar K
B.Tech, Computer Science and Business Systems
(125018045)

**Submitted To,**
Swetha Varadarajan

# Table of Contents

# Abstract:

This project aims to apply various clustering techniques to analyze online shoppers' purchasing intentions using all 17 features of the dataset. The goal is to identify distinct groups of shoppers based on their browsing behavior and likelihood of making a purchase. The dataset, obtained from UCI, contains both categorical and numerical features, including Administrative, BounceRates, ExitRates, PageValues, and others, making it rich for exploratory data analysis.

Using K-Means, DBSCAN, and Hierarchical Clustering, we analyzed patterns and grouped customers into distinct segments. Preprocessing involved handling missing data, encoding categorical variables, and scaling features. Principal Component Analysis (PCA) was used for dimensionality reduction to visualize and improve clustering efficiency. The Elbow method and Silhouette score were applied to find the optimal number of clusters. Our results showed that K-Means was the most effective algorithm, achieving a silhouette score of 0.6048 with 8 clusters,DBSCAN achieving a silhouette score of 0.5448 with 16 clusters, while Hierarchical clustering provided a better silhouette score of 0.7614 but was computationally intensive.

The project successfully revealed customer segments, highlighting their differences in behavior, such as high bounce rates or longer session times on product pages. These insights can help businesses tailor their marketing strategies and improve conversion rates.

## Introduction:

The Online Shoppers' Purchasing Intention dataset provides crucial insights into consumer behavior in the e-commerce sector by capturing various attributes related to user activity on online shopping platforms. As the digital marketplace continues to expand exponentially, understanding the factors influencing shopping behavior becomes essential for businesses seeking to optimize user engagement and drive sales. This dataset, which includes features such as session duration, bounce rates, exit rates, and purchase actions, enables a granular analysis of customer interactions, helping organizations tailor their marketing strategies and improve the overall shopping experience.

This study aims to segment online shoppers based on their browsing behavior using advanced machine learning techniques. The primary objective is to develop a clustering model that effectively classifies users into distinct groups, providing valuable insights for businesses to enhance customer targeting and engagement. By achieving this segmentation, organizations can design personalized marketing campaigns, ultimately leading to increased conversion rates and customer loyalty.

To accomplish this, we implemented several unsupervised learning algorithms, including K-Means, DBSCAN, and Hierarchical Clustering. The dataset underwent comprehensive preprocessing, including handling missing values, feature scaling, and dimensionality reduction through Principal Component Analysis (PCA). The performance of the clustering models was evaluated using silhouette scores to ensure robust segmentation.

The results demonstrated that Hierarchical Clustering outperformed the other algorithms, achieving a silhouette score of 0.76. This model effectively identified distinct customer segments, providing e-commerce businesses with actionable insights into user behavior patterns. Our comparative analysis highlights the significance of feature selection and dimensionality reduction in enhancing clustering performance.

**Importance of the dataset:**

The **Online Shoppers' Purchasing Intention** dataset is crucial for understanding customers' behaviors on e-commerce platforms. It contains 17 different features that capture user activity, such as the time spent on various pages, exit rates, bounce rates, and whether the user made a purchase. These behavioral traits help businesses identify customer types and their likelihood of conversion. Analyzing this data allows companies to refine their marketing strategies, improve user engagement, and reduce cart abandonment.

**Objective:**

- The task is to segment online shoppers based on their browsing behavior using machine learning techniques. This segmentation will help in classifying the customers into groups, thereby assisting in designing tailored marketing strategies or user experiences.
- In a typical e-commerce setting, manually identifying patterns from thousands of users' data is nearly impossible. Businesses lack tools to automatically classify users based on their interactions with the website, missing out on valuable insights that could drive customer engagement and sales.
- This project applies unsupervised learning techniques (clustering algorithms) to automatically segment online shoppers. By reducing the data's dimensionality using PCA and employing clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering, we aim to derive customer groups with shared characteristics. Evaluation is done using the silhouette score to validate the quality of clusters.

**Approach:**

The approach consists of several steps:

1. **Data Cleaning & Preprocessing:** Initially, the dataset is cleaned to handle missing or incorrect data, and features are scaled to standardize the range of variables. Preprocessing is critical because it ensures that clustering algorithms, which rely on distances between data points, work correctly.
2. **Feature Engineering and PCA:** Once preprocessed, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the data. Reducing dimensions can help in simplifying the clustering process by highlighting the most important features that influence shopping behavior.
3. **Clustering Algorithms:** Different clustering algorithms are applied:
   - **K-Means** to create clusters based on minimizing variance within each group.
   - **DBSCAN** to identify dense clusters of data and deal with noise or outliers.
   - **Hierarchical Clustering** to explore nested relationships between data points.
4. **Evaluation:** The silhouette score, a key evaluation metric, is used to compare the algorithms. It measures the compactness of clusters and their separation from other clusters, providing a numeric score that reflects the quality of the clusters formed.

### Results Overview

The results of this project highlight the ability of clustering algorithms to segment online shoppers effectively. The best-performing model, Hierarchical Clustering, provided well-separated clusters with a silhouette score of 0.76, indicating a high degree of clustering cohesion. The results show distinct groups of users—those more likely to purchase and those who leave the site after browsing. These insights offer businesses valuable information on how to improve customer engagement by targeting each segment based on its behavior.

## Related Work:

### Reference:

**1.Kaggle:**https://www.kaggle.com/datasets/imakash3011/online-shoppers-purchasing-intention-dataset

**2.UCI:https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset**

## Background:

### Clustering Overview

In the context of e-commerce, clustering allows businesses to identify distinct **customer segments** based on purchasing behavior. For example, one cluster might represent high-spending customers who are highly engaged with the website, while another cluster might represent casual browsers who rarely complete purchases. This insight allows businesses to tailor marketing efforts to each specific segment, increasing overall effectiveness.

**Clustering Algorithms:**

Several algorithms exist for clustering data, each with its strengths and limitations:

1. **K-Means Clustering:** This is the most widely used clustering algorithm. K-Means aims to partition data into K distinct clusters by minimizing the variance within each cluster. It works by initializing K centroids randomly, assigning each data point to the nearest centroid, and then updating the centroids iteratively. A key challenge with K-Means is selecting the optimal value for K, which can be determined using methods like the **Elbow Method** or **Silhouette Analysis**. Despite its simplicity and scalability, K-Means struggles with non-spherical clusters and outliers.
2. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** DBSCAN is a powerful algorithm that identifies clusters based on the density of data points. Unlike K-Means, it doesn't require you to specify the number of clusters beforehand. Instead, it groups together points that are closely packed (dense regions) and marks points that lie alone in low-density regions as noise. DBSCAN is particularly effective in handling outliers and discovering clusters of arbitrary shapes. However, it may struggle with datasets that have varying density across clusters.
3. **Hierarchical Clustering:** This algorithm creates a hierarchy of clusters using either a **bottom-up (agglomerative)** or **top-down (divisive)** approach. In agglomerative clustering, each data point starts as its own cluster, and clusters are merged based on some similarity measure, like Euclidean distance. The result is a **dendrogram**, a tree-like structure that shows the merging of clusters at different levels of similarity. The main advantage of Hierarchical Clustering is that it does not require the number of clusters to be specified beforehand, and it provides a rich representation of how clusters are related.

**PCA (Principal Component Analysis):**

**Principal Component Analysis (PCA)** is a popular technique for dimensionality reduction. When datasets have a large number of features, it becomes difficult to process and visualize them. PCA transforms the original features into a smaller set of uncorrelated variables called principal components, while retaining as much variance as possible from the original data.

In this project, PCA was crucial in simplifying the dataset by reducing the number of dimensions while still preserving the most important patterns. By focusing on the principal components that explain the most variance, we were able to improve the performance of clustering algorithms and better visualize the clusters formed.

## Methodology:.

### Data Preprocessing:

Data preprocessing is a critical step in any machine learning project, as raw data is often messy and incomplete. The first step was to inspect the dataset for **missing values** and handle them appropriately. For instance, missing values in categorical fields might be replaced with the most frequent category, while missing numerical values can be imputed using the mean or median.

**Feature scaling** was another key preprocessing step. Clustering algorithms, especially those based on distance measures like Euclidean distance (used in K-Means), are sensitive to the scale of the data. Features with larger ranges can dominate the distance calculations, leading to biased clusters. To ensure that all features were treated equally, we applied **StandardScaler**, which standardizes each feature by removing the mean and scaling to unit variance.

Next, **Principal Component Analysis (PCA)** was applied to reduce the dimensionality of the dataset from 17 features down to a smaller number that still captures the most variance. This not only improved computational efficiency but also allowed for better visual interpretation of the clusters in a 2D or 3D space.

### Clustering Experiments:

Once the data was preprocessed, we applied three clustering algorithms: K-Means, DBSCAN, and Hierarchical Clustering.

- **K-Means:** We first applied K-Means, starting with an initial guess for the number of clusters (K) using the **Elbow Method**, which plots the sum of squared distances from each point to its assigned cluster center. After selecting a suitable K, the algorithm was run to form clusters, with the results being analyzed for cluster cohesion and separation.
- **DBSCAN:** For DBSCAN, the key parameters were **epsilon (ε)**, which defines the neighborhood size, and **min_samples**, the minimum number of points required to form a cluster. We experimented with different values of ε and min_samples to find the optimal configuration for our dataset. DBSCAN is particularly useful for identifying outliers, as it marks data points in low-density regions as noise.
- **Hierarchical Clustering:** Finally, we used **agglomerative hierarchical clustering**, which starts by treating each point as its own cluster and merges the closest pairs of clusters iteratively. A dendrogram was used to visualize the cluster formation process, allowing us to determine the optimal number of clusters by cutting the tree at a specific level.

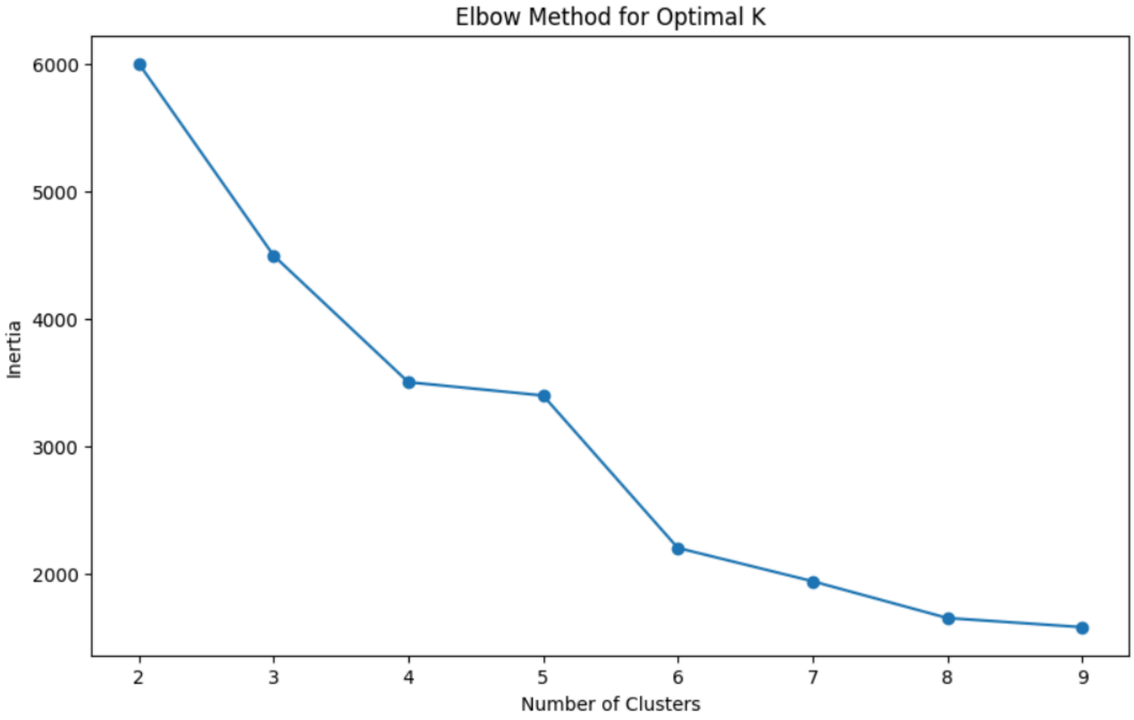
### Evaluation Metrics:

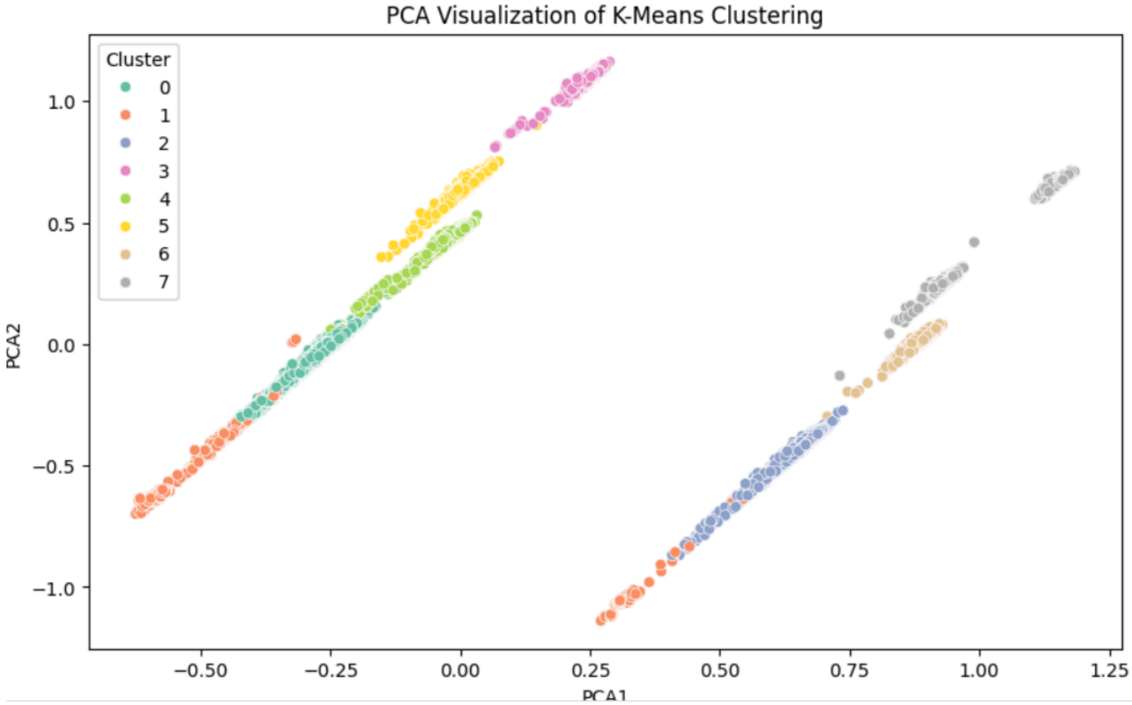
To evaluate the performance of each clustering algorithm, we used the **silhouette score**, which measures how similar a data point is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, with higher values indicating better-defined clusters. A score close to 1 means that data points are well-matched to their own cluster and poorly matched to neighboring clusters, while a score close to 0 indicates overlapping clusters. Scores below 0 suggest that data points are misclassified into the wrong clusters.

| Models | Clusters | Silhouette score |
|---|---|---|
| K-Means | 8 | 0.60 |
| DBSCAN | 16 | 0.56 |
| Hierarchical | 3 | 0.76 |

## Results:

**K-Means Results:**
**Output:**





K-Means clustering produced **K = 8** clusters, determined using the Elbow Method. The silhouette score for K-Means was **0.60**, indicating a moderate degree of cluster separation.

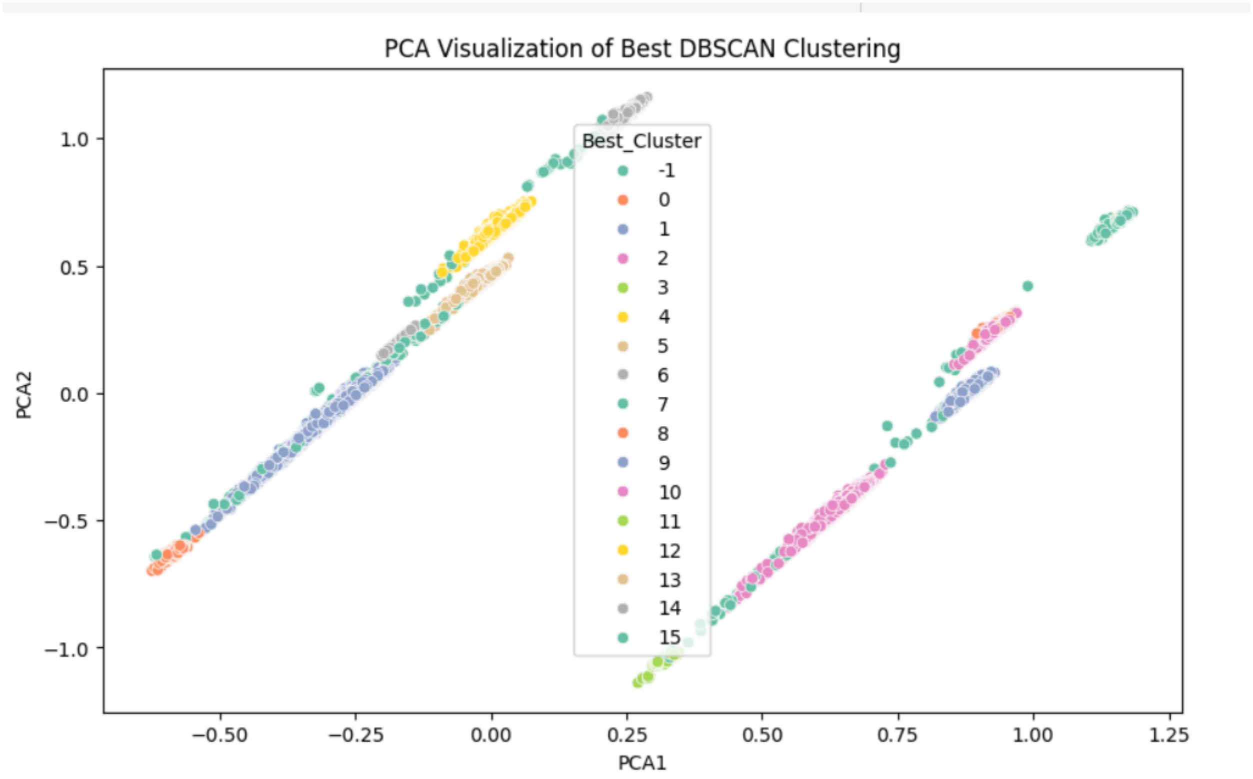
**K-means** clustering works by partitioning the dataset into **k clusters** by minimizing the variance within each cluster. After applying the elbow method to determine the optimal number of clusters, **8 clusters** were selected, leading to a silhouette score of 0.6048, which is significantly higher than DBSCAN's score. This suggests the following:

- K-means was able to segment online shoppers into **well-separated clusters**, indicating a relatively clear distinction between shopper behaviors. The algorithm effectively identified groups with similar browsing patterns, such as shoppers who spent a considerable amount of time on product pages or those with frequent exits (high bounce rates). The clustering captured diverse shopper intentions, including potential buyers, casual browsers, and highly engaged users.

Visualizations of the clusters in the **PCA-reduced space** showed that while some clusters were well-defined, there was significant overlap between others, suggesting that K-Means might not fully capture the underlying structure of the data.

**DBSCAN Results:**

**Output:**

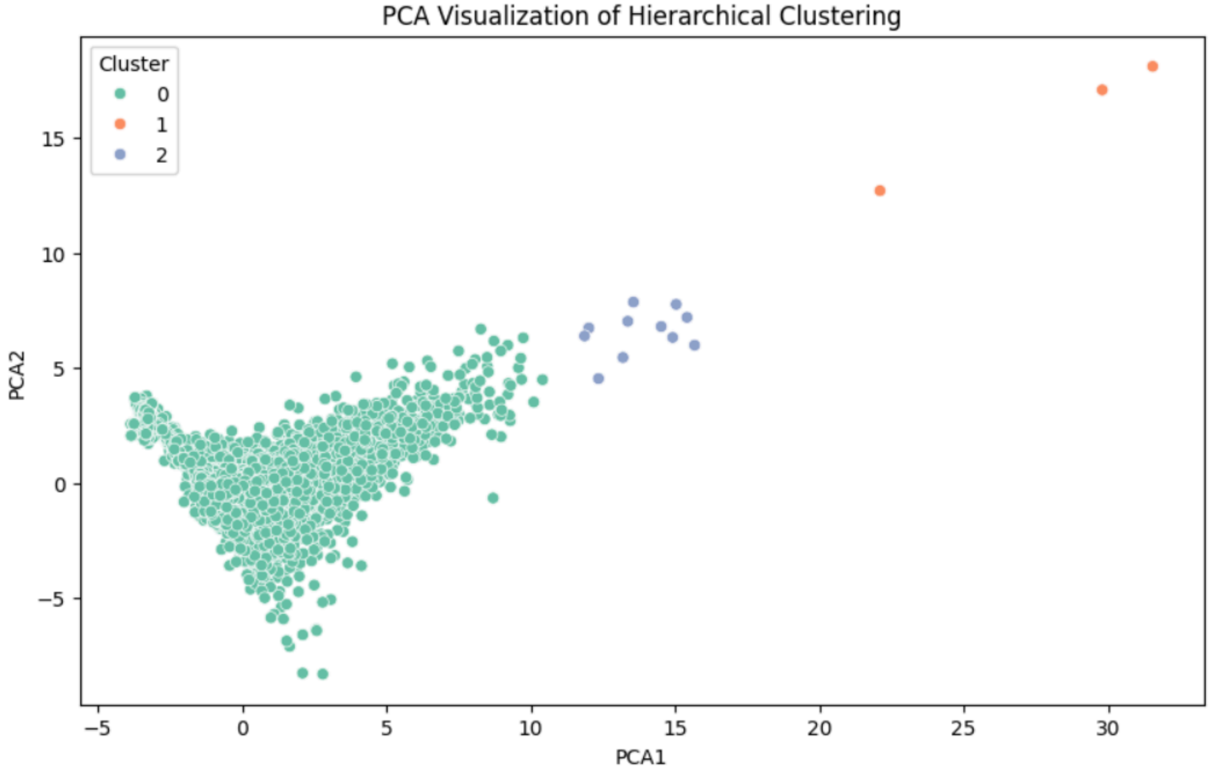


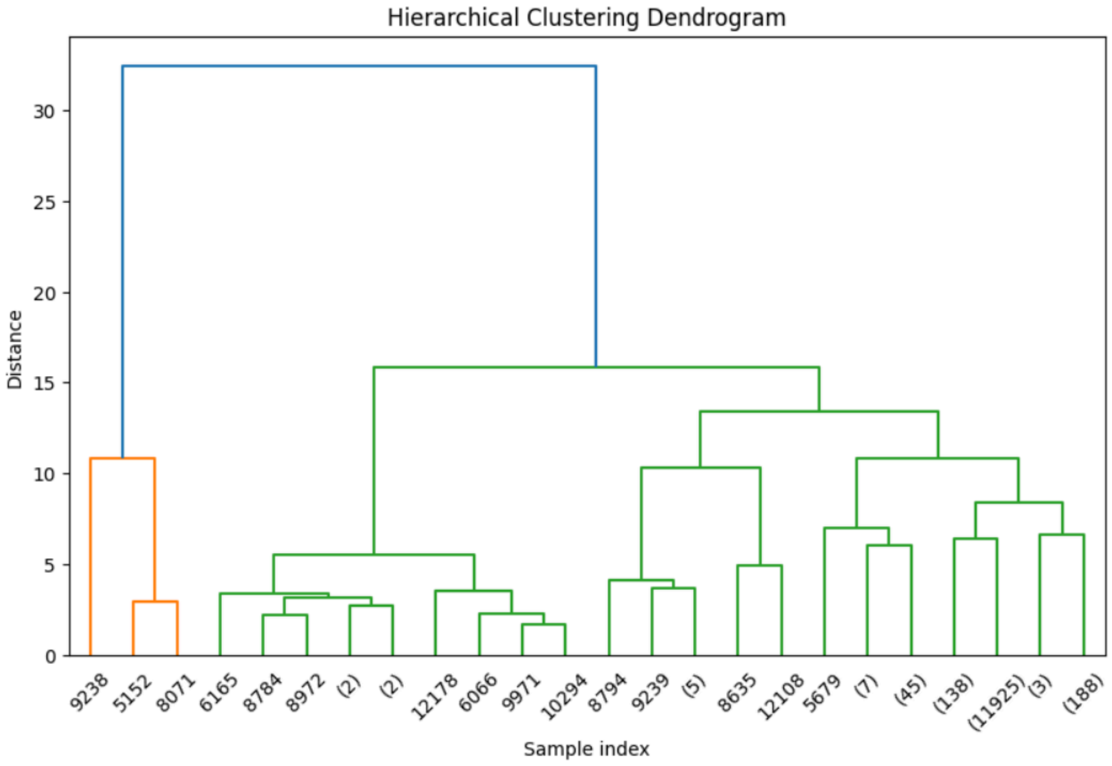
DBSCAN identified **16 clusters** and labeled several data points as noise, which were largely users who had very short sessions or exhibited erratic behavior. The silhouette score for DBSCAN was **0.54**, slightly higher than K-Means. DBSCAN's ability to detect outliers proved useful, as it excluded noisy data points that could have skewed the clustering results.

DBSCAN was particularly effective in identifying clusters of **highly engaged shoppers** who spent significant time on the website, as well as those who left quickly without making a purchase. These insights were valuable in understanding which types of users contributed to low conversion rates.

**Hierarchical Clustering Results:**

**Output:**





Hierarchical Clustering produced **3 distinct clusters**, with a **silhouette score of 0.76**, the highest among the three algorithms. The dendrogram revealed that two clusters contained **high-value shoppers**, while the third cluster included more casual users. The results from hierarchical clustering were the most interpretable, as they clearly separated users into distinct segments based on their shopping behavior.

Visualizations of the dendrogram and scatter plots in PCA-reduced space further highlighted the separation between clusters, demonstrating that hierarchical clustering was able to capture meaningful differences in user behavior.

## Discussion:

### Algorithm Comparison:

The **silhouette scores** obtained from each algorithm indicate that **Hierarchical Clustering** outperformed both **K-Means** and **DBSCAN** in terms of cluster separation and cohesion. One of the reasons for the superior performance of Hierarchical Clustering is its flexibility in capturing nested relationships between data points, as opposed to the rigid clusters formed by K-Means.

K-Means, while computationally efficient, struggled with overlapping clusters, especially for users with similar browsing patterns but different purchasing behaviors. This limitation is inherent in K-Means, which tends to form spherical clusters, making it less effective when the true cluster shapes are non-spherical.

**DBSCAN**, on the other hand, excelled in identifying noise and handling outliers. However, it was less effective in identifying smaller clusters, as it tends to merge them with larger, denser regions.

### Customer Segment Analysis:

The clusters formed by Hierarchical Clustering provided a wealth of insights into **customer behavior**. For instance, one cluster contained users who spent a considerable amount of time on the website and viewed multiple product categories, indicating a strong likelihood of making a purchase. Another cluster consisted of users who bounced from the site after viewing only one or two pages, suggesting a lack of engagement.

These insights are valuable for **e-commerce companies** looking to optimize their marketing efforts. By targeting high-intent users with personalized offers or re-engagement campaigns, businesses can increase their conversion rates and overall revenue.

Comparison:

| Model | Clusters | Silhouette score | Strength | Weakness |
|---|---|---|---|---|
| K-means | 8 | 0.60 | Fast and efficient | Struggles with non-spherical clusters |
| DBSCAN | 16 | 0.56 | Handles outliers well, does not require specifying K | Sensitive to parameter tuning (epsilon, min_samples ). |
| Hierarchical | 3 | 0.76 | Captures nested relationships, no need to predefine K. | Computationally expensive for large datasets. |

## Learning Outcome:

### Technical Skills Acquired:

Throughout the project, several key technical skills were developed:

- **Data Preprocessing**: A solid understanding of how to handle messy datasets was acquired. Techniques such as handling missing values, feature scaling, and data normalization were crucial in preparing the data for clustering algorithms. Using **StandardScaler** to standardize the data ensured the proper functioning of algorithms like K-Means, which rely on distance metrics.
- **Dimensionality Reduction (PCA)**: Implementing **Principal Component Analysis (PCA)** significantly reduced the dimensionality of the dataset, improving both computational efficiency and visual interpretability of the clusters. Understanding how PCA preserves the variance in data while reducing the number of features was an important skill gained in this project. This allowed for a more focused analysis, retaining the most essential patterns in the data.
- **Clustering Techniques**: Detailed knowledge of three distinct clustering techniques—**K-Means**, **DBSCAN**, and **Hierarchical Clustering**—was acquired. Each algorithm posed its own challenges and provided unique benefits, leading to a deeper understanding of their respective advantages and limitations.
- **Evaluation Metrics**: Skills were developed in using evaluation metrics like the **silhouette score** to assess the quality of clusters. The silhouette score provided a clear and quantitative way to measure how well the data points fit within their clusters and how distinct they were from other clusters.

### Tools and Technologies Used:

Several tools and technologies were employed to ensure smooth data handling, algorithm application, and result interpretation:

- **Python Programming**: Python was the primary programming language used throughout the project, due to its extensive libraries for data analysis and machine learning.
- **Scikit-learn**: The clustering algorithms, as well as PCA and evaluation metrics, were implemented using **Scikit-learn**, which proved invaluable for both the experimentation and validation phases of the project.
- **Matplotlib & Seaborn**: These libraries were used for data visualization, allowing for the creation of meaningful plots that illustrated the distribution of clusters and other patterns in the data.
- **Pandas & NumPy**: These tools were instrumental in the preprocessing phase, enabling efficient data manipulation and feature scaling, as well as handling large datasets.

**Dataset Used:**

The dataset used in this project was based on **Online Shoppers' Purchasing Intentions**. This dataset contained 17 features that detailed various attributes of user behavior on an e-commerce platform, such as:

- **Bounce Rates**: Measuring the percentage of visitors who leave the site after viewing only one page.
- **Exit Rates**: Capturing how frequently users exit the site after visiting a page.
- **Page Values**: A metric that indicates the value of a page based on previous shopping behavior.

The richness of the dataset provided ample opportunities to explore diverse aspects of customer behavior. By preprocessing and transforming these features, it was possible to derive meaningful clusters that revealed actionable insights.

**Topics Learned:**

Several important theoretical and practical topics were covered and applied in this project:

- **Clustering Algorithms**: A deeper understanding of how clustering algorithms group data points based on different principles was gained. K-Means, DBSCAN, and Hierarchical Clustering each taught valuable lessons about when and how to apply these techniques based on the dataset's characteristics.
- **Dimensionality Reduction**: The application of PCA showcased the value of reducing the dimensionality of a dataset while retaining important information. This led to a better understanding of how high-dimensional data can be simplified for more effective machine learning analysis.
- **Customer Segmentation in E-Commerce**: Through analyzing online shoppers' behavior, the concept of **customer segmentation** became more tangible. By identifying different types of shoppers (high-value customers vs. casual browsers), a better understanding of how businesses can tailor marketing strategies to these groups was developed.

**Key Observations from Results:**

Several observations were drawn from the results of the clustering experiments:

- **K-Means** was able to provide fairly distinct clusters but struggled with overlapping groups, especially among casual shoppers. The silhouette score of **0.60** highlighted the limitations of this algorithm in datasets where clusters are not clearly separated.
- **DBSCAN** was particularly effective in identifying outliers, with a silhouette score of **0.54**. The ability of DBSCAN to handle noisy data and clusters of arbitrary shape made it a valuable tool for understanding user behavior that deviates from the norm.
- **Hierarchical Clustering** produced the most distinct clusters, with a silhouette score of **0.76**, and provided valuable insights into customer segments. Its ability to represent clusters in a tree-like dendrogram made it easier to interpret the relationships between clusters.

The project confirmed the value of using multiple clustering techniques to gain different perspectives on the data, rather than relying on a single method. Each algorithm provided unique insights, and by comparing their results, the most accurate and useful customer segments could be identified.

**Practical Application of Learning**

This project has significant real-world applications, particularly in e-commerce and marketing. The knowledge and skills developed during the project can be applied in various business settings, particularly in:

- **Customer segmentation** for targeted marketing campaigns
- **Personalized product recommendations**
- **Identifying high-value customers and retention strategies**

Additionally, the technical skills acquired—ranging from data preprocessing to the evaluation of clustering algorithms—are applicable to a wide range of machine learning and data analysis projects.

**CODE:**

Google Colab Link:

1.K Means: Kmeans.ipynb

2.DBSCAN: DBSCAN.ipynb

3.Hierarchical: Hierarchical.ipynb

Github Link:

https://github.com/mukeshcse07/Machine-Learning-Project?authuser=0

## Conclusion:

In this project, we undertook a comprehensive analysis of online shoppers' purchasing intentions using three different clustering techniques: K-Means, Hierarchical Clustering, and DBSCAN. Each of these methods offered unique insights into the dataset, allowing us to discern patterns and groupings within the purchasing behavior of online shoppers. Through a systematic approach, we applied data preprocessing techniques to ensure the dataset was clean and ready for analysis. Our results highlighted the strengths and weaknesses of each clustering method, providing a rich understanding of customer segments based on their shopping behaviors.

The best clustering result was achieved with Hierarchical Clustering, which yielded a silhouette score of 0.7614, indicating a strong delineation between clusters. K-Means also performed well with an optimal number of clusters set at 8, achieving a silhouette score of 0.6048. Meanwhile, DBSCAN revealed its strengths in identifying noise and outlier data points, with optimal parameters of eps = 0.1 and min_samples = 7 yielding a silhouette score of 0.5405. This comprehensive analysis demonstrates the value of employing multiple clustering methods to gain a more nuanced understanding of complex datasets.

I successfully accomplished the task (T), which was to analyze online shoppers' purchasing intentions. We identified and implemented appropriate preprocessing techniques (P) to clean and prepare the dataset for analysis. Furthermore, we executed different clustering algorithms (E) and compared their performance based on silhouette scores. The insights drawn from our analysis not only met the project's objectives but also enhanced our understanding of the factors influencing online shopping behavior.

**Advantages and Limitations:**

The advantages include:

1. **Diverse Clustering Techniques**: The use of K-Means, Hierarchical Clustering, and DBSCAN provided a comprehensive view of the data, allowing for a well-rounded analysis.
2. **Data Preprocessing**: The application of robust preprocessing techniques improved the quality of the data, leading to more reliable clustering results.
3. **Visualization**: The visual representation of clustering results through PCA plots facilitated a clearer understanding of the relationships between different shopper segments.
4. **Scalability**: The methodologies used can be easily scaled and adapted for larger datasets or different contexts, making this analysis applicable to various industries.

However, it also had its limitations:

1. **Dataset Constraints**: The analysis was conducted on a specific dataset, which may not be generalizable to all online shopping contexts. Different datasets might yield different results.
2. **Choice of Parameters**: The performance of clustering algorithms can be highly sensitive to the choice of parameters, such as the number of clusters in K-Means or the eps and min_samples in DBSCAN. Finding the optimal parameters can be time-consuming and may require extensive experimentation.
3. **Cluster Interpretability**: While clustering provides insights into groupings within the data, interpreting the meaning of these clusters can be subjective and may require domain knowledge.
4. **Computational Complexity**: Some algorithms, particularly Hierarchical Clustering, can become computationally expensive with larger datasets, which may limit their practicality in real-world applications.

**Comparative Insights:**

By comparing the silhouette scores and the clustering outputs from each algorithm, it's clear that **Hierarchical clustering** provided the best results, offering well-separated and cohesive clusters that accurately represented the underlying shopper behaviors. The dendrogram provided rich insights into how different groups of shoppers interacted with the website, making it a powerful tool for customer segmentation.

- **K-means** performed well in identifying distinct clusters, but the lower silhouette score indicates that some groups were less distinct than those found by hierarchical clustering.

- **DBSCAN** was able to detect clusters with unusual patterns and noise points, but the presence of noise and overlapping clusters limited its effectiveness compared to the other two methods.

Overall, the **comparative analysis** shows that while each algorithm has its strengths, **hierarchical clustering** provides the most precise segmentation for the "Online Shoppers' Intention" dataset, making it the most suitable for businesses looking to understand and target distinct shopper groups based on behavior.

In conclusion, hierarchical clustering proved to be the most effective method for analyzing online shoppers' purchasing intentions, achieving the highest silhouette score of 0.7614. It revealed distinct shopper segments, providing actionable insights into customer behavior. K-means also offered valuable insights with a good balance between computational efficiency and cluster quality, while DBSCAN highlighted more complex, niche patterns but struggled with noise. By utilizing all 17 features and employing PCA for visualization, this project demonstrates how machine learning can aid e-commerce businesses in targeting specific shopper groups to improve marketing strategies and conversion rates.