CUSTOMER SEGMENTATION

University of Missouri – Kansas City

Lecturer: Syed Jawad Hussain Shah

3rd May 2024

Group Members
Praveen Kumar Reddy Kadapala
Manoj Kumar Suggala
Venkata Sai Veeramalla
Sreevardhan Reddy Soma

INTRODUCTION:

In today's dynamic business environment, gaining profound insights into customer behavior is crucial for customizing products and services effectively. This project aims to utilize diverse clustering algorithms to categorize customers according to their characteristics, including age, gender, annual income, and spending score. By segmenting customers, businesses can refine their marketing strategies to target specific customer groups, thereby enhancing greater customer satisfaction and driving revenue growth.

RELATED WORK:

Numerous clustering techniques have been applied in the past to address particular challenges and offer unique advantages in the field of customer segmentation research. Because of their ease of use and scalability, traditional methods like K-Means clustering have long been preferred. They efficiently divide data into discrete clusters according to similarity metrics. Hierarchical clustering, on the other hand, provides a hierarchical view of the data, enabling a more thorough comprehension of cluster relationships and streamlining exploratory data analysis. Density-based clustering algorithms, like DBSCAN, are excellent at locating clusters of any size and shape while withstanding data noise and anomalies with resilience. With its probabilistic framework, Gaussian Mixture Model (GMM) clustering provides flexibility in modeling complex data distributions, allowing for complex cluster assignments.

METHODOLOGY:

Our methodology consists of several steps. To ensure data consistency and integrity, the dataset containing customer attributes was first preprocessed to fill in missing values and encode categorical variables. After preprocessing, features were analyzed and their distribution and characteristics were visualized, through visualization. which gave important insights into the structure of the dataset. The relative importance of each feature was then determined, and possible relationships between features were revealed, by feature importance analysis.

After that, the preprocessed data was subjected to four different clustering algorithms, K-Means, Agglomerative Hierarchical, DBSCAN, and Gaussian Mixture Model (GMM). The goal of these algorithms was to divide the customer base into meaningful segments according to their attributes. Finally, metrics like the silhouette score were used to evaluate the quality of the resulting clusters.

Scaling

A common preprocessing step in clustering algorithms is scaling, which makes sure that each feature contributes equally to the clustering process. It involves converting the feature values to a comparable scale, which is usually 0 to 1 or has a mean of 0 and a standard deviation of 1. Larger magnitude features are kept from influencing the distance calculations during clustering by this normalization, or standardization, of feature scales.

Gaussian Mixture Model Clustering:

Gaussian Mixture Model (GMM) clustering is a probabilistic model used for clustering data by assuming that they are composed of a combination of multiple Gaussian distributions. Each Gaussian distribution within the model corresponds to a distinct cluster within the dataset, allowing for the identification of underlying patterns and structures by which the different clusters of customers are formed based on their attributes.

The Bayesian Information Criterion (BIC) is commonly used in GMM clustering to identify the ideal number of components, or gaussian distributions or clusters. The final GMM model was trained with this number of components after the ideal number of clusters was established.

The Expectation-Maximization (EM) algorithm is used in GMM clustering to estimate the Gaussian distributions' parameters. The Expectation step (E-step) and the Maximization step (M-step) are the two steps that the EM algorithm iteratively switches between.

In E-step (Expectation Step) using the current estimates of the parameters, the algorithm calculates the probabilities (responsibilities) of each data point that belongs to each cluster in the E-step. In M-step (Maximization Step) considering the current responsibilities, the algorithm modifies the parameters of the Gaussian distributions to maximize the likelihood of the data.

Up until a convergence criterion is satisfied, the EM algorithm alternates between the E-step and the M-step. Following convergence of the parameters, each data point is allocated to the cluster whose probability (responsibility) for belonging to that cluster is highest.

K-Means Clustering:

Initialization of the K-Means Clustering is done with a predetermined number of clusters (K). The elbow method was used to calculate the ideal number of clusters (K). Using this method, the within-cluster sum of squares (WCSS) is plotted against the number of clusters (K) to determine the "elbow" point, or the point at which the WCSS decreases at a significantly slower rate. This figure represents the ideal number of clusters. The final K-Means model was trained with this value of K after the ideal number of clusters was determined.

Then using the Euclidean distance as a guide, K-Means iteratively assigns each data point to the closest cluster centroid and updates the centroids based on the average of the data points within each cluster. Convergence was reached, which was demonstrated by little centroid movement or completing the maximum number of iterations. The resulting clusters formed represent the distinct groups of customers based on their attributes.

Hierarchical clustering:

Hierarchical clustering is a clustering technique for organizing data points into clusters according to their proximity or similarity. It constructs a hierarchy of clusters through a process of iteratively merging or dividing data points until certain criteria are satisfied. This method groups together the most similar data points first and progressively combines them into larger clusters, creating a hierarchical structure.

Hierarchical clustering yields a dendrogram, a graphical representation illustrating the cluster hierarchy. This dendrogram showcases the series of mergers or divisions, enabling users to comprehend the interconnections among clusters at varying levels of detail.

Agglomerative clustering:

Agglomerative clustering is a bottom-up hierarchical clustering technique that starts from the bottom, aiming to group similar data points into clusters. It initiates with each data point forming its own cluster and then progressively merges the nearest clusters based on a specified distance measure.

Agglomerative clustering offers flexibility in selecting the distance metric and linkage criterion, which govern the process of merging clusters. Popular linkage criteria such as single linkage, complete linkage, and average linkage exert distinct influences on both the clustering outcomes and the structure of the dendrogram. Agglomerative clustering is advantageous for exploratory data analysis as it eliminates the need to predetermine the number of clusters.

Density Based Spatial Clustering of Applications with Noise (DBSCAN):

DBSCAN stands out for its ability to automatically determine the number of clusters and handle datasets with clusters of varying shapes and sizes. Its approach revolves around grouping densely populated data points into clusters, without requiring a predefined number of clusters. This method allows for the identification of clusters that may have irregular shapes and sizes, enhancing its applicability to diverse datasets.

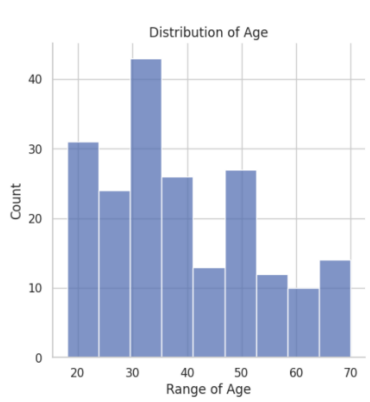
To apply the DBSCAN clustering we need to find the suitable ‘eps’ value (i.e., the maximum distance between two points for them to be considered as in the same neighborhood) and “min_samples” (i.e., The minimum number of points required in a neighborhood for a point to be considered a core point). To get “eps” value we used the K-distance plot to get idea about how data points are distributed and calculated distance matrix if there is need to perform any further analysis.

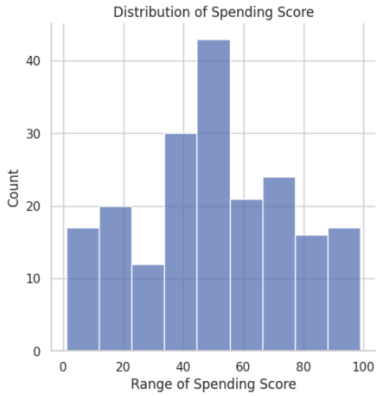
Silhouette score:

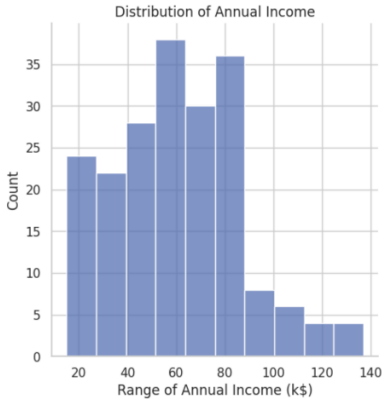
The silhouette score is a metric used to evaluate the quality of clusters produced by clustering algorithms, including K-Means, Hierarchical Clustering, DBSCAN, and others. It measures how well-separated clusters are and provides insights into the compactness and separation of clusters within the dataset.

RESULTS AND DISCUSSION:

Data Visualization:







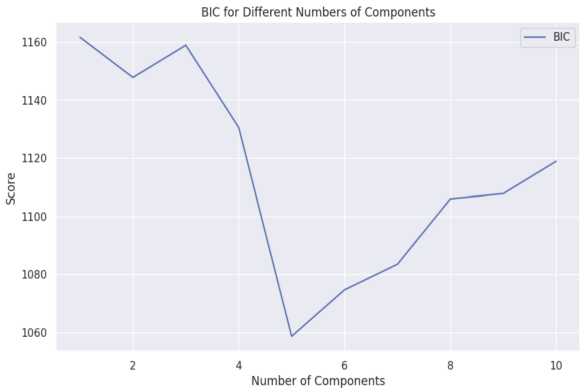
These visualizations show the frequency of each attribute (age, spending score and Annual Income) in a specific range. From the plots we can see that there are more people in the age between 30 to 35 and spending score of most customers is 40 to 50 and Annual Income of most of the customers is between 50 to 80 k$.

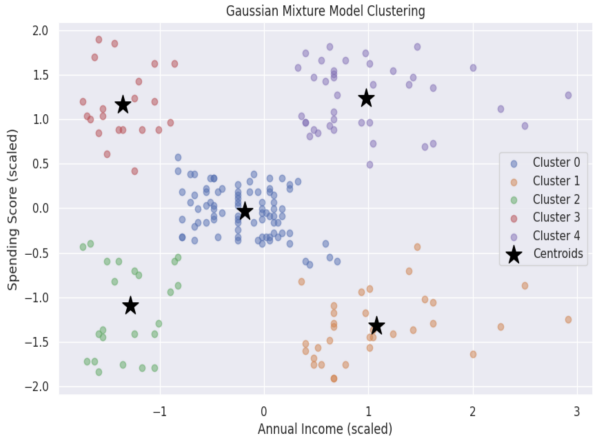
These visualizations show the Age vs Spending Score distribution based on Gender. From these plots we can infer that customers between the age group 20 to 40 has higher spending score than other age customers.

This visualization shows the Annual Income vs Spending Score distribution based on Gender. From this plot we can infer that almost all customers with annual income between 40 to 60 have spending score between 40 to 60.

Gaussian Mixture Model Clustering:

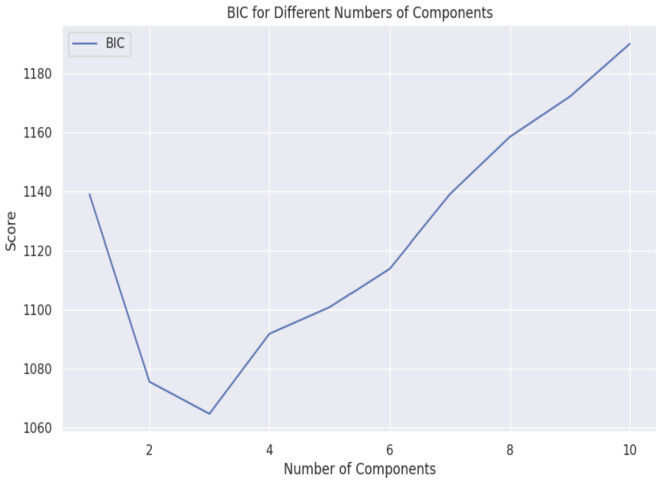
BIC curve plotting for Annual Income and Spending score with Components ranging from 1 to 10. As in the curve we can clearly see that BIC score is low with components so fitting GMM model with 5 components gives the best clusters formation for Age and Spending Score attributes.

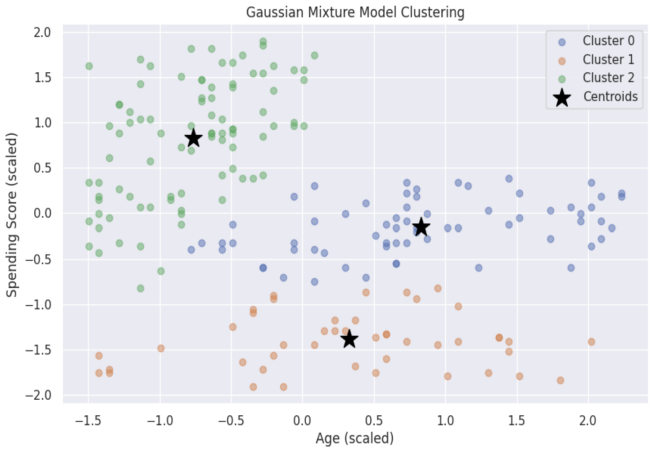




Above plotting shows the clusters formation for Annual Income and Spending Score attributes. Since we are fitting the model with 5 components the model has resulted with the 5 different components or clusters. Stars in the plotting represent the centroids of each cluster.

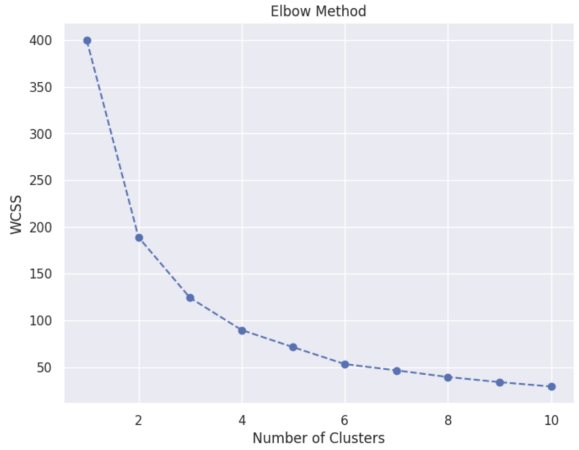
BIC curve and visualizing the clusters formed using Gaussian Mixture Model Clustering for Age and Spending Score. Since BIC score is lowest for 3 components the GMM model is fitted with 3 components and resulted with 3 cluster formation.

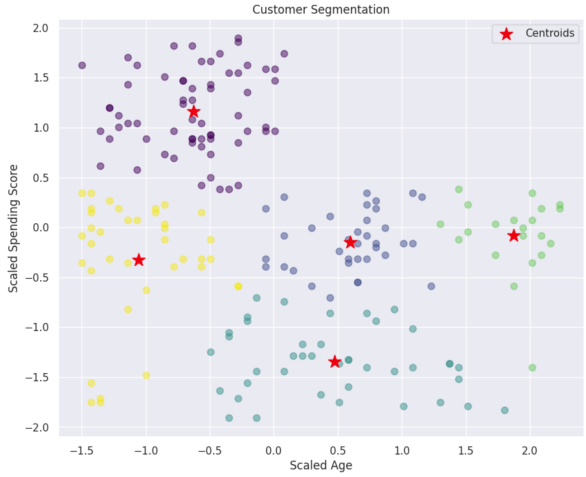




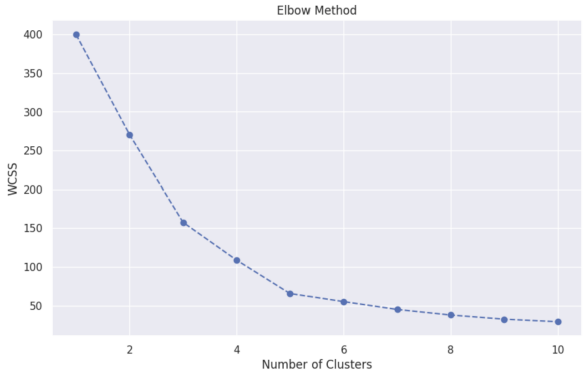
K-Means Clustering:

Elbow curve and visualizing the clusters formed using K-Means Clustering for Age and Spending Score. The "elbow" point, or the point at which the WCSS decreasing significantly slower rate is at 5 so we consider K=5 (No of clusters) and fitted the K-means model with 5 clusters and obtained the following clusters for age and spending score.



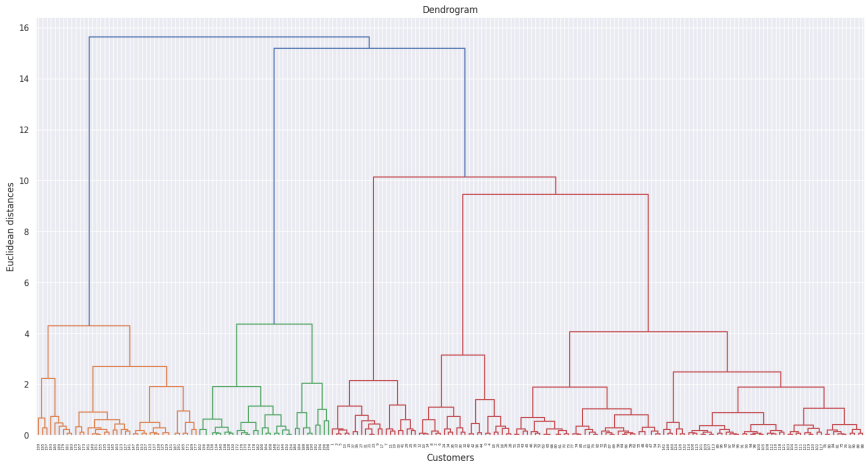


Elbow curve and visualizing the clusters formed using K-Means Clustering for Annual income and Spending Score. The "elbow" point, or the point at which the WCSS decreasing significantly slower rate is at 5 so we consider K=5 (No of clusters) and fitted the K-means model with 5 clusters and obtained the following clusters for Annual Income and spending score.



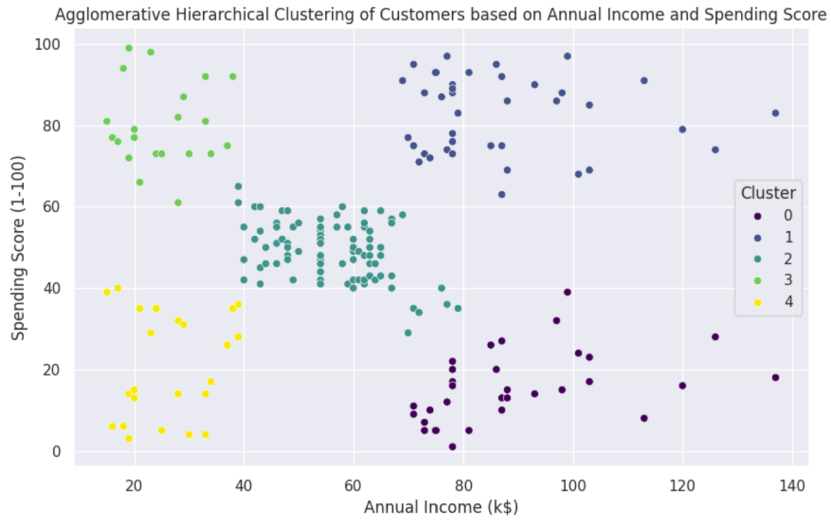
Hierarchical Clustering:

Organized the data into a dendrogram-based tree structure from whichwe can understand that and interprets the clustering patterns without the need for predefining the number of clusters.

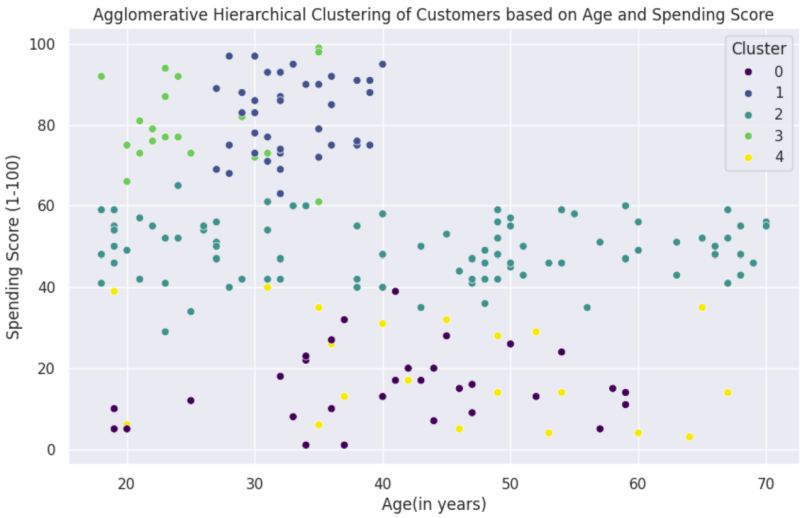


Agglomerative Hierarchical Clustering:

Assigns customers to clusters based on their annual income and spending score, revealing distinct patterns in their behavior across different demographic dimensions.

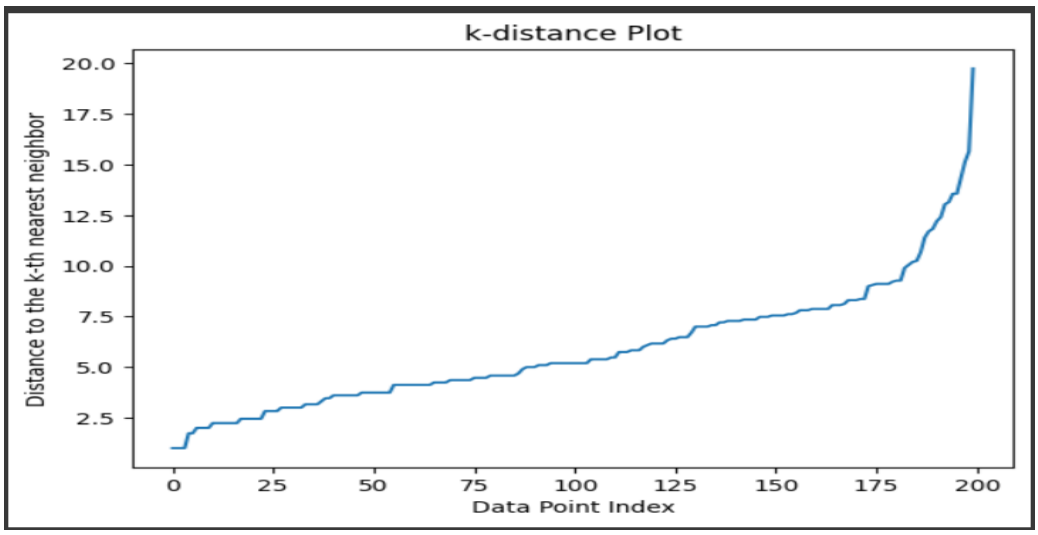


Assigns customers to clusters based on their Age and spending score, revealing distinct patterns in their behavior across different demographic dimensions

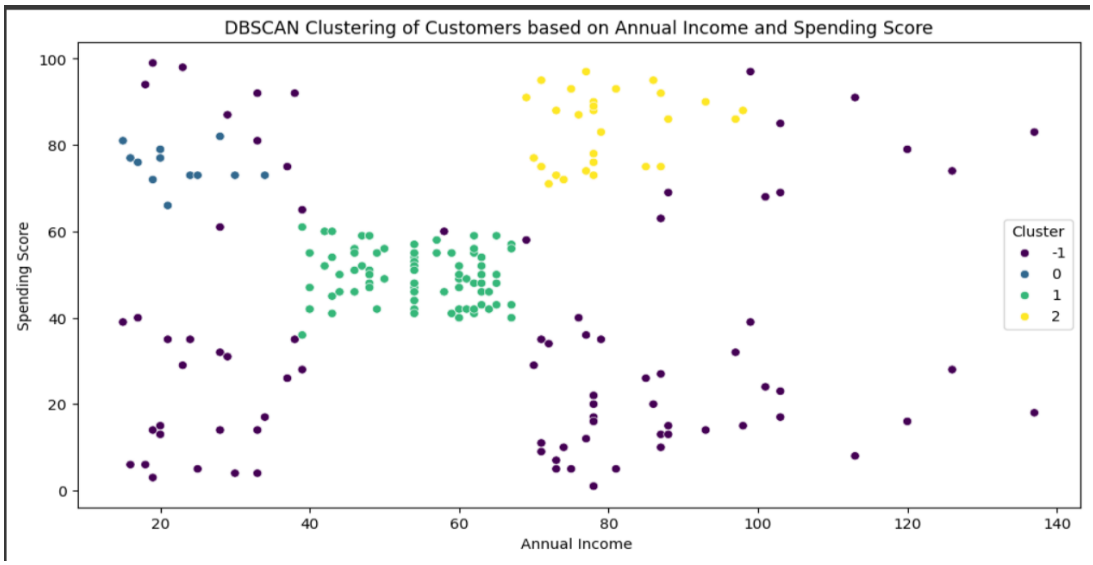


Density Based Spatial Clustering of Application with Noise (DBSCAN):

This is the "k-distance plot" and from the plot we observe that distance to Kth nearest neighbor is ranges from 5-10 so by taking eps value between these and choosing the reasonable value of min_samples based on density of the data and desired granularity of the clusters from 3-6 as parameters for DBSCAN.



With the application of Model DBSCAN to data given we got the results and visualized on graph



From the above graph it was shown that the clusters are (-1,0,1,2)

- where -1 indicates that the data points are outliners or Noise.
- 0 indicates that the data points are Belongs to no cluster.
- 1,2 indicates that the data points are divided into 2 clusters.

Silhouette score:

	GMM	**K-Means**	**Agglomerative Hierarchical**	**DBSCAN**
Annual Income vs Spending Score	**0.553689**	**0.554657**	**0.553808**	**0.197075**
Age vs Spending Score	**0.413429**	**0.447548**	**0.403452**	**0.340385**

- From these Silhouette scores we can infer that the quality of cluster formation for Annaul Income and Spending Score is good in GMM, K-Means and Agglomerative Hierarchical clustering models and quality of cluster formation for Age and Spending Score is good in K-means comparing to other models.

CONCLUSION AND FUTURE WORK:

In conclusion We have explored and applied four different clustering techniques like "**k-means**" "**Hierarchical Clustering**" "**GMM**" "**DBSCAN**" to the problem of Customer Segmentation based on demographic and behavioral attributes. Each technique offers unique advantages and considerations, contributing valuable insights to our analysis.
In Future work We Aim to enhance Customer Segmentation techniques by exploring additional features and transformations to capture nuanced customer behavior better. Additionally, We plan to investigate ensemble clustering methods to improve the robustness and stability of segmentation solutions. By integrating clustering with predictive modeling techniques will enable us to leverage customer segments for personalized marketing campaigns and recommendations.

REFERENCES

[1]. Jomark Pablo Noriega; Luis Antonio Rivera , Jose Alfredo Herrera. Machine Learning for Credit Risk Prediction: A Systematic Literature Review, 2023.

[2]. Norshakirah Aziz; Emelia Akashah Patah Akhir; Izzatdin Abdul Aziz. A Study on Gradient Boosting Algorithms for Development of AI Monitoring and Prediction Systems, 2020.

[3]. Aized Amin Soofi; Classification Techniques in Machine Learning: Applications and Issues, 2017.

[4]. Swastik Satpathy; SMOTE for Imbalanced Classification with Python, 2023