

Task 3: Customer Segmentation / Clustering

I applied four clustering algorithms to know which algorithm best suited in this scenario—KMeans, Agglomerative, DBSCAN, and Gaussian Mixture Models (GMM)—to segment customers. Choosing a good algorithm for clustering is important because different algorithms work best for different types of data and problems. Using the wrong algorithm can create incorrect or meaningless clusters, leading to bad decisions. Now, among the four algorithms, Agglomerative performed the best with a DB Index of 0.808, forming compact and well-separated clusters. KMeans also produced 4 distinct clusters with a slightly higher DB Index of 0.865, making it a suitable alternative. DBSCAN struggled to form meaningful clusters, resulting in a higher DB Index of 1.044, while GMM performed the worst with overlapping clusters and a DB Index of 1.120.

Agglomerative created the most compact and well-separated clusters, making it the best algorithm for this dataset. KMeans also performed well but was slightly less compact than Agglomerative Clustering. DBSCAN struggled to form clear clusters and marked many points as noise. GMM had overlapping clusters that were not very distinct. Overall, Agglomerative Clustering worked the best.

In Agglomerative clustering each cluster was studied to understand its behavior. Cluster 0 had customers who spent a lot and bought often. Cluster 1 had customers who spent a moderate amount and bought sometimes. Cluster 2 had customers who spent less and bought rarely. Cluster 3 had customers who spent less but bought in large amounts.