DS605 Fundamentals of Machine Learning – Lab 6 Report

Aarsh Adhvaryu (202518031)[1]

[1] MSc Data Science

Date: September 26, 2025

Keywords: Mall Customers Clustering using K-Means, Hierarchical, and DBSCAN.

Abstract. This report analyzes the Mall Customers dataset using K-Means, Hierarchical Clustering, and DBSCAN. We determine the optimal number of clusters, compare algorithm performance, evaluate DBSCAN's handling of outliers, and discuss the most suitable algorithm for this dataset. Finally, we propose real-world marketing strategies based on the identified customer segments.

1 Introduction

Customer segmentation is essential for targeted marketing and business decision-making. In this lab, clustering algorithms were applied to the Mall Customers dataset to explore natural groupings of customers. The objective was to evaluate K-Means, Hierarchical Clustering, and DBSCAN, and to interpret the resulting segments for marketing applications.

2 Results and Discussion

2.1 Optimal Clusters

The optimal number of clusters for both K-Means and Hierarchical Clustering is five. This was determined using the Elbow Method, which shows a clear bend at $k = 5$, and the Dendrogram, which yields five main groups when cut at an appropriate distance.

2.2 Cluster Comparison

K-Means and Hierarchical Clustering produced nearly identical, well-defined, globular clusters. DBSCAN, however, identified only three core clusters and labeled many points as noise. This demonstrates that K-Means and Hierarchical methods assign every point to a cluster, while DBSCAN detects dense regions and marks outliers.

2.3 DBSCAN Performance

DBSCAN performed as expected by labeling non-conforming points as noise. This prevents outliers from distorting cluster structure, a key advantage over K-Means and Hierarchical Clustering. However, in this dataset, DBSCAN excluded some points that visually belonged to clusters, making it less suitable.

2.4 Algorithm Suitability

K-Means and Agglomerative Hierarchical Clustering are the most suitable algorithms for this dataset. The clusters are spherical and well-separated, aligning with the assumptions of these methods. DBSCAN was less effective due to excessive noise labeling.

2.5 Real-World Application

The five customer segments identified provide clear directions for marketing:

- **High Income, High Spending:** Offer exclusive VIP programs.

- **High Income, Low Spending:** Use personalized, high-value offers to encourage more spending.
- **Medium Income, Medium Spending:** Apply general promotions and loyalty programs.
- **Low Income, High Spending:** Send alerts for sales and discounts.
- **Low Income, Low Spending:** Attract with general advertisements and family-friendly events.

3 Conclusion

The Mall Customers dataset reveals five natural customer segments. K-Means and Hierarchical Clustering provided the most accurate representation, while DBSCAN highlighted outliers but misclassified some valid customers as noise. Clustering analysis thus enables businesses to design data-driven marketing strategies tailored to different customer groups.