CORIZO

MACHINE LEARNING

MAJOR PROJECT

TITLE: CUSTOMER SEGMENTATION USING UNSUPERVISED LEARNING

DATASET : MALL CUSTOMER DATA

GROUP MEMBERS:

1. DIVYA DHARSHINI
2. HIMANSHU
3. SHAIK ARSHIYA
4. CHINMAY HEGDE
5. PRATHIK R K
6. NIKHIL RAJ

INTRODUCTION

In this project, our team explores the domain of **Unsupervised Learning**, where the objective is to uncover hidden patterns or groupings in unlabeled data. We selected **K-Means Clustering** as our model and the **Mall Customers Dataset** as our dataset, focusing on segmenting customers based on their Annual Income and Spending Score.

ABSTRACT

This study analyzes customer segmentation using the Mall Customers dataset, which includes demographic and behavioral attributes such as Gender, Age, Annual Income, and Spending Score. After preprocessing by removing the non-predictive 'CustomerID' and encoding the 'Gender' column, clustering was performed using the two key features—Annual Income and Spending Score. The optimal number of clusters was determined through the Elbow Method and Silhouette Score, ensuring both visual and quantitative validation of the results. K-Means Clustering was then applied, effectively grouping customers into distinct segments, including high-income high-spenders, low-income minimal spenders, and moderate earners with diverse spending behaviors. The resulting visualizations, featuring scatter plots and centroids, provided clear insights into customer distribution and purchasing patterns.

RESULT

After performing clustering using the K-Means algorithm on the Mall Customers dataset, the following key results were obtained:

1. Optimal Number of Clusters Identified:
    - Using the Silhouette Score method, the best number of clusters (k) was found to be 5.
    - This value of k ensured that the clusters were well-separated and meaningful, with minimal overlap.
2. Cluster Formation and Visualization:
    - Each customer was assigned to one of the five clusters based on Annual Income and Spending Score.
    - A scatter plot visualized these clusters, clearly showing distinct groupings.
    - Cluster centroids were also plotted, indicating the central point of each customer group.
3. Silhouette Score:
    - The final Silhouette Score was approximately 0.272, indicating a moderate clustering structure.
    - A higher Silhouette Score (closer to 1) indicates well separated clusters. A score of 0.272 means the clusters are distinguishable, but with some overlap.
    - Customer Segment Insights:
By analyzing the mean values of Age, Annual Income, and Spending Score for each cluster, we observed:

- Cluster 0: High income, low spending – likely conservative spenders.
- Cluster 1: Low income, low spending – price-sensitive or less engaged customers.
- Cluster 2: Moderate income, high spending – potentially loyal or enthusiastic customers.
- Cluster 3: High income, high spending – premium customers or ideal marketing targets.
- Cluster 4: Low income, high spending – may require financial planning or targeted offers.

CONCLUSION

This project demonstrates how K-Means clustering can be applied to real-world business data to uncover hidden customer segments. The insights obtained can help businesses optimize marketing strategies, personalize customer experiences, and make data-driven decisions—all without any prior labels or supervision.

****