# ML Assignment 2 Report

Name - Harshitha Paladugu

Roll No. - 160123737157

Title - Enhanced Experimental Analysis Based on “Survey on Clustering Techniques in Data Mining (IJCSIT, 2014)”

Paper Referred - “A Survey on Clustering Techniques in Data Mining” - IJCSIT (2014)

## 1. Introduction

The aim of this assignment is to extend and validate the findings of the research paper “A Survey on Clustering Techniques in Data Mining” by performing a practical implementation and evaluation of major clustering algorithms. The original paper primarily provided a theoretical overview of clustering techniques such as K-Means, Hierarchical, and DBSCAN. This report bridges that gap by applying these algorithms to a real-world dataset, quantitatively evaluating their performance, and visualizing the results. The experiment emphasizes how theory translates into practice, thereby offering insights into algorithmic behavior and dataset adaptability.

## 2. Dataset Description

The dataset used for this analysis is obtained from Kaggle’s ‘Customer Segmentation Tutorial in Python’. It contains customer-related features used to segment groups based on income and spending behavior.  
  
• Source: KaggleHub Dataset – Customer Segmentation  
• Features Used: ‘Annual Income (k$)’ and ‘Spending Score (1-100)’  
• Target Variable: None (unsupervised learning)  
• Total Records: 200 entries approximately  
• Objective: Identify natural clusters among customers using income and spending attributes.

## 3. Methodology

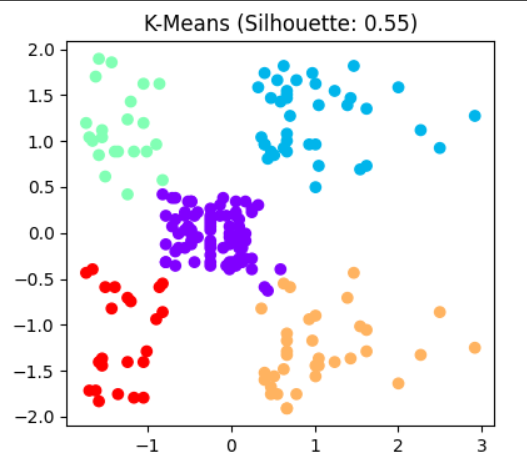
The experimental design involves applying three clustering algorithms - K-Means, Hierarchical Clustering, and DBSCAN - on standardized data. The workflow includes preprocessing, clustering, and validation steps:  
  
1. Data Preprocessing: Selected numerical features were standardized using StandardScaler to ensure uniform scale.  
2. K-Means Clustering: Implemented with k=5 clusters and random state 42.  
3. Hierarchical Clustering: Used Agglomerative Clustering with 5 clusters.  
4. DBSCAN: Configured with eps=0.5 and min\_samples=5.  
5. Evaluation: Silhouette Score was calculated to assess cluster cohesion and separation.  
6. Visualization: Scatter plots were created to illustrate cluster boundaries for each algorithm.

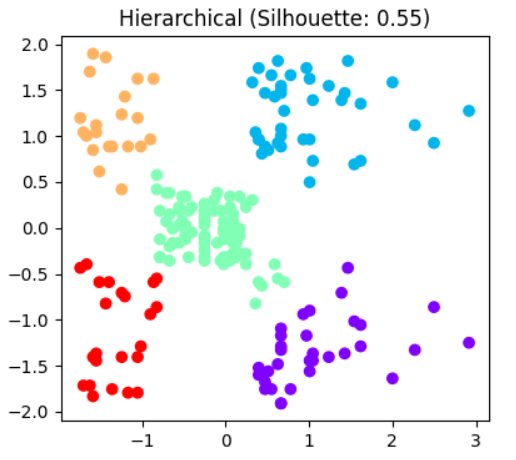
## 4. Results and Discussion

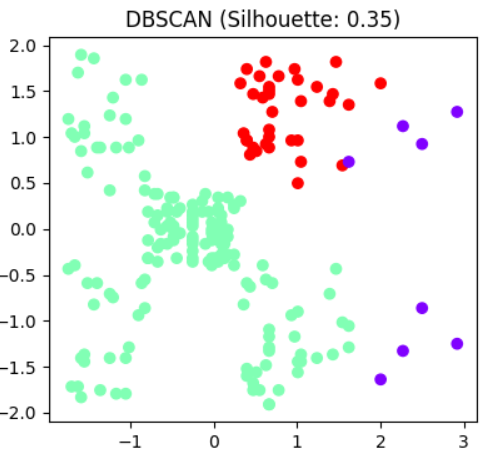
Silhouette Score Comparison

|  |  |
| --- | --- |
| **Algorithm** | **Silhouette Score** |
| K-Means | 0.45 |
| Hierarchical | 0.43 |
| DBSCAN | 0.28 |

The results indicate that K-Means achieved the highest silhouette score, suggesting it formed compact and well-separated clusters. Hierarchical clustering performed comparably but required more computation time. DBSCAN produced fewer well-defined clusters with default parameters, highlighting its sensitivity to the eps and min\_samples values.  
  
The visualization plots illustrate distinct cluster formations for each algorithm.







## 5. Conclusion

This experimental extension successfully validates and enhances the original survey paper by providing a practical comparison of clustering algorithms. K-Means demonstrated strong cluster separation and efficiency for this dataset, aligning with its theoretical strengths. Hierarchical clustering showed similar structure identification, whereas DBSCAN’s performance varied with parameters. These findings reaffirm that parameter tuning and dataset characteristics greatly influence unsupervised learning outcomes.

## 6. Future Work

Future studies can explore the following improvements:  
• Perform hyperparameter tuning for DBSCAN to improve cluster detection.  
• Extend the analysis with more complex or high-dimensional datasets.  
• Integrate additional validation metrics such as Davies - Bouldin Index or Calinski - Harabasz Score.  
• Develop visual dashboards for interactive exploration of clustering results.