Lab Assignment Report

Assignment

CS503: Machine Learning

Name	Entry No.
Anshika	2021CSB1069

Instructor: Dr. Santosh Vipparthi

Teaching Assistant: Nishchala Thakur

Indian Institute of Technology Ropar
Punjab, India
April 13, 2025

Contents

1 | Data Analysis and Preparation

Loading and preprocessing the Mall Customer Segmentation data. The dataset contains demographic information such as age, gender, annual income, and spending score.
Steps:

- The dataset is read using `pandas`.
- `CustomerID` was dropped as it's not relevant to the analysis.
- Gender was encoded using one-hot encoding.
- Basic descriptive statistics were computed using `df.describe()`.
- Histograms were plotted to visualize distributions of features like Age, Annual Income, and Spending Score.

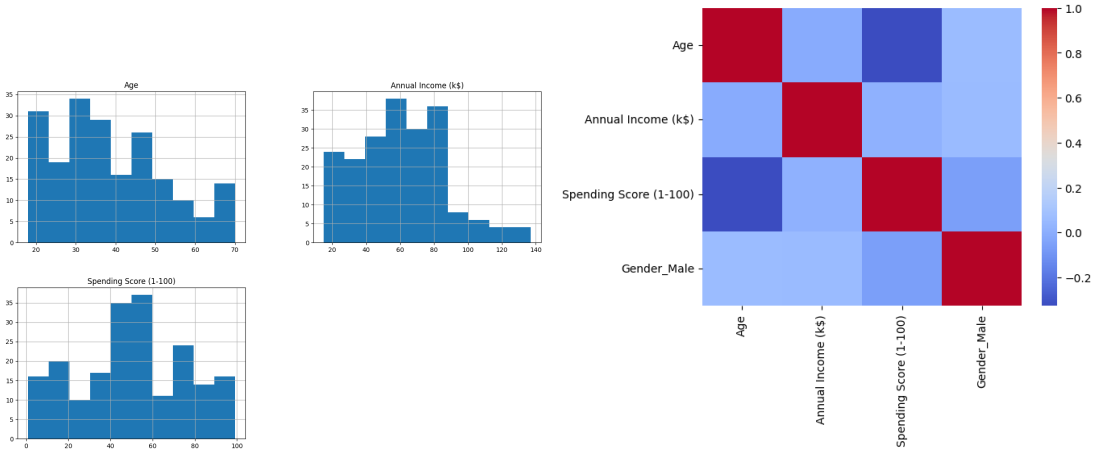


Figure 1.1: Dataset Visualization Mall Customer Segmentation data

Observations:

- The distributions show varying degrees of skewness.
- Spending Score seems to be uniformly distributed.
- Gender encoding (0 for Female, 1 for Male) helps with clustering later on.

2 | PCA for Feature Reduction

Principal Component Analysis (PCA) was used to reduce the number of features while retaining as much variance as possible.
Steps:

- StandardScaler was used to normalize the dataset.
- PCA was applied with a goal of preserving 95% of the variance.
- A Scree plot was used to determine the number of components.
- The dataset was then projected onto the first two principal components for visualization.

Observations:

- PCA helped reduce the dimensionality to 2 components while maintaining over 95% variance.
- The scatter plot of the two principal components shows some natural clustering patterns even before clustering algorithms are applied.

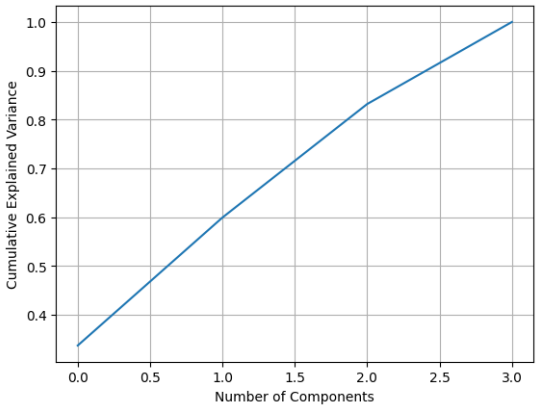


Figure 2.1: PCA

3 | Clustering Algorithms

This task involves applying multiple clustering algorithms to segment customers based on their features. The analysis is performed on PCA-reduced data for better visualization and efficiency.

3.1 | K-Means Clustering

Approach:

- The elbow method was used to determine the optimal number of clusters by plotting the within-cluster sum of squares (WCSS).
- The silhouette score was also used to evaluate the cohesion and separation of clusters.
- K-Means was applied with `k=5` as it gave a good balance between WCSS and silhouette score.

Observations:

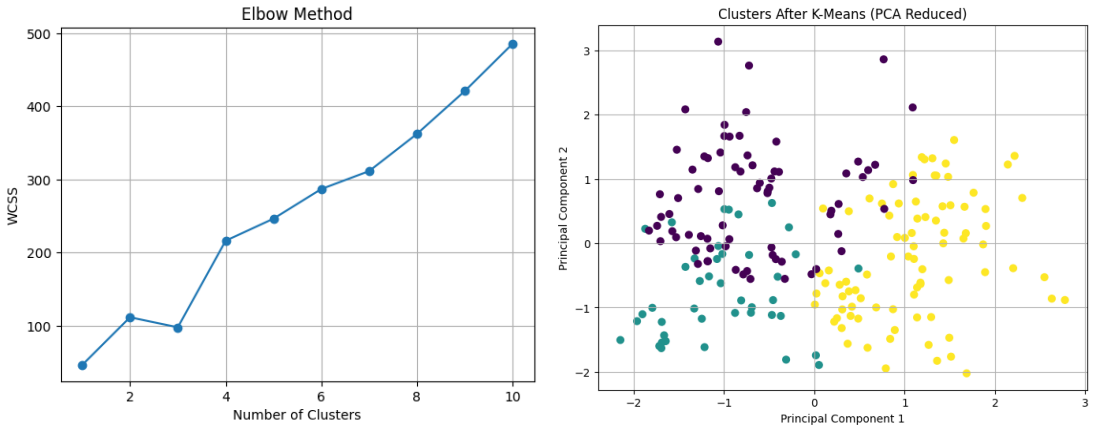


Figure 3.1: K-Means Clustering

Cluster	Age	Annual Income (k$)	Spending Score (1–100)	Gender_Male
0	30.92	82.59	60.19	0.41
1	25.83	32.33	64.19	0.38
2	52.09	55.59	34.71	0.49

Table 3.1: Cluster-wise mean values of key features

- The resulting clusters were well-separated and interpretable.
- PCA-reduced 2D visualizations showed distinct groups, indicating meaningful segmentation.
- Cluster centers reflected key customer behaviors such as high income–low spending vs. low income–high spending.

3.2 | DBSCAN Clustering- Sentiment Analysis of Tweets

Approach:

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was used to detect clusters of varying shapes and sizes.
- Parameters `eps=0.5` and `min_samples=5` were initially selected and tuned further based on clustering output.
- Unlike K-Means, DBSCAN does not require specifying the number of clusters.

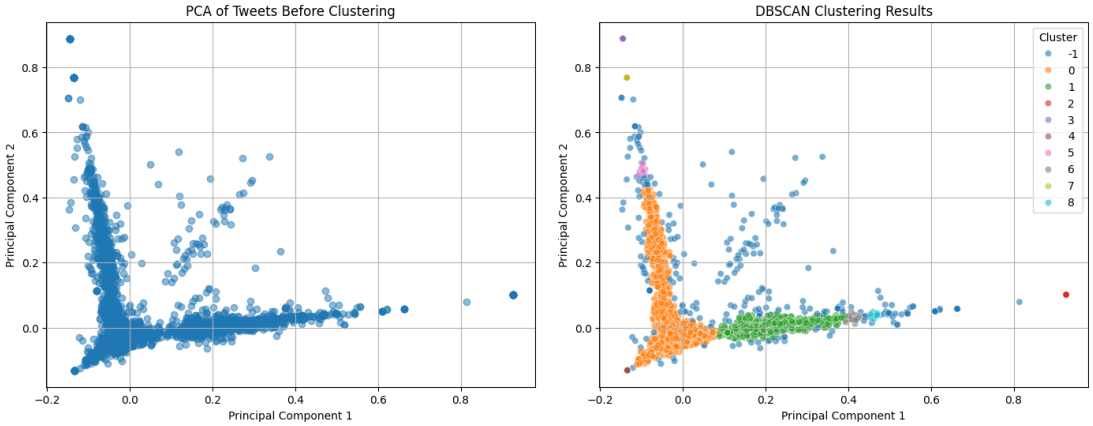


Figure 3.2: Before and After DBSCAN

Observations:

- DBSCAN identified a few dense clusters, but labeled many points as noise due to the parameter sensitivity.
- It struggled with this dataset, likely due to its assumption of equal density which doesn't hold well in the customer data.
- Despite some limitations, DBSCAN is valuable for identifying outliers or unusual customer groups.

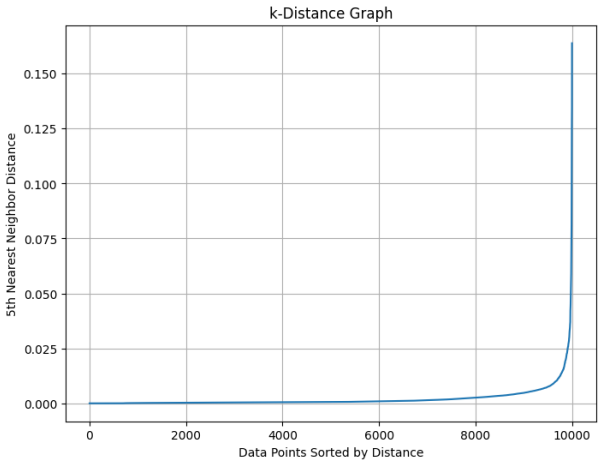


Figure 3.3: Determining epsilon

Cluster	Sample Tweets
0	@chrishasboobs AHHH I HOPE YOUR OK!!! @misstoriblack cool, i have no tweet apps for my razr 2 @TiannaChaos i know just family drama. its lame. hey next time u hang out with kim n u guys like have a sleepover or whatever, ill call u School email won't open and I have geography stuff on there to revise! *Stupid School* :'(upper airways problem
1	@PerezHilton Zach makes me pee sitting down! And I'm a grown gay man! i'm feeling quite sleepy today, wish i could stay in bed today...but OK! is my LAST YEAR, so let's go to school im turning 18 one week from now but i don't feel excited i really don't know why i've seen my friends they got excited but on my part @jasonhooha Going back to sleep is such a great idea, unfortunately I'm already at work notebook-less first night in myers. just not the same w/out lydia! but i'm actually excited about this summer!
2	I'm having an allergic reaction to silver earings oopss did that twice D: im such a silly moo. @dungodung I'm all ears im starving now, hix... @FashionKristin who won????? I'm very curious
3	@cjm55 Thanks! @BearTwinsMom thanks sweetie! @Mickie1 thanks @AustinIrl thanks Just downloaded twitterberry. Thanks for the tip @tiffanypr
4	@stephenfry Oooo, Sound's good @cosmicstuff Sounds good to me Picked up my SDS9 today! Pretty Good, Pretty Good! Good morning yall @AmyyVee good track isnt it!
5	good morning good morning! @TheRealJordin Good morning Jordin Good Morning Good morning!
6	@teamallen jealous! I'm yet to find a local chinese i'm satisfied with @aSaladADay I'm sorry...not trying to do that I'm back guys :3 wee. I'm ready to tweet again @thisisryanross I'm sick too @SeattleFutbol SCARVES UP!! I'm grabbing my crew and going to watch from Fuel, no tix.
-1	Going to miss Pastor's sermon on Faith... i don't feel good @ElleChanel want to shoot...but i'm not in charlotte .. @sephrenia1982 good work i miss him, i cant wait to see him, im happy

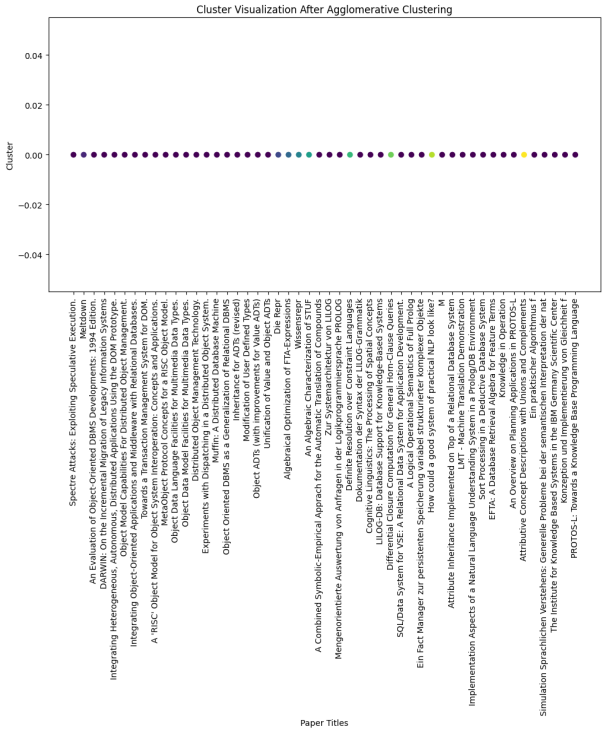
Table 3.2: Sample Tweets from Each Cluster

3.3 | Hierarchical Clustering

Approach: Dendogram is shown in figure 4.3

- Agglomerative Hierarchical Clustering was used with Euclidean distance and Ward's linkage.
- A dendrogram was plotted to decide the number of clusters.
- The final model was fit with the chosen number of clusters (e.g., 5), and the results were visualized.

Observations:



- Dendrograms provided useful visual intuition about how clusters form step by step.
- The resulting clusters were similar in distribution to those found by K-Means.
- Hierarchical clustering is more interpretable and does not require an initial guess for cluster centers.

Cluster	Top 5 Keywords
0	object, distributed, data, database, adts
1	meltdown, wissensrepr, vse, von, verstehen
2	repr, die, wissensrepr, von, vse
3	optimization, expressions, algebraical, fta, verstehen
4	wissensrepr, zur, vse, von, verstehen
5	stuf, algebraic, characterization, wissensrepr, vse
6	resolution, constraint, definite, languages, wissensrepr
7	queries, differential, clause, closure, computation
8	practical, like, look, nlp, good
9	unions, descriptions, attributive, complements, concept

Table 3.3: Top 5 Keywords for Each Cluster

4 | Association Rule Mining

This task explores frequent itemsets and association rules using the Apriori algorithm.
Steps:

- A transactional dataset (`Market_Basket_Optimisation.csv`) was loaded.
- The data was converted into a list of transactions.
- The Apriori algorithm was applied with a minimum support of 0.003 and minimum confidence of 0.2.

- The rules were sorted by lift and analyzed for insight.

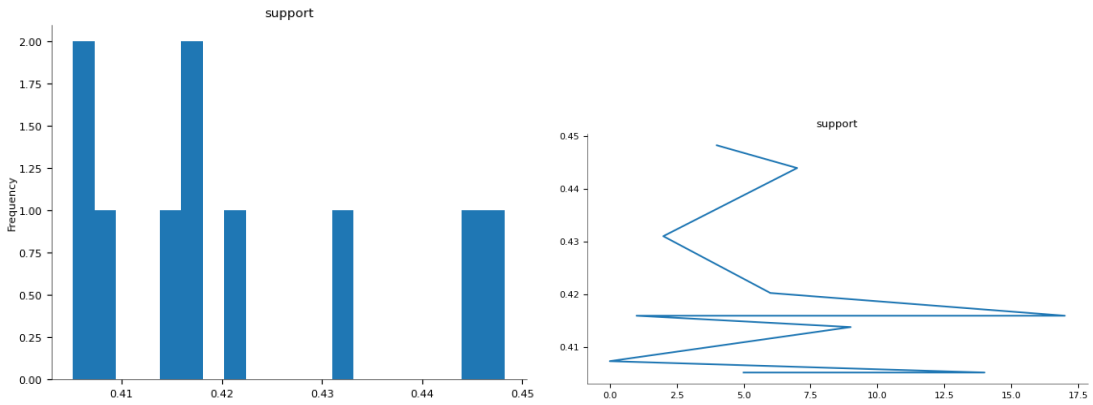


Figure 4.1: Data Visualization

Observations:

- Several strong association rules were found between commonly bought items like milk, bread, and eggs.
- High lift values (greater than 3) indicate strong associations beyond random chance.
- These insights can help with inventory planning and recommendation systems in retail settings.

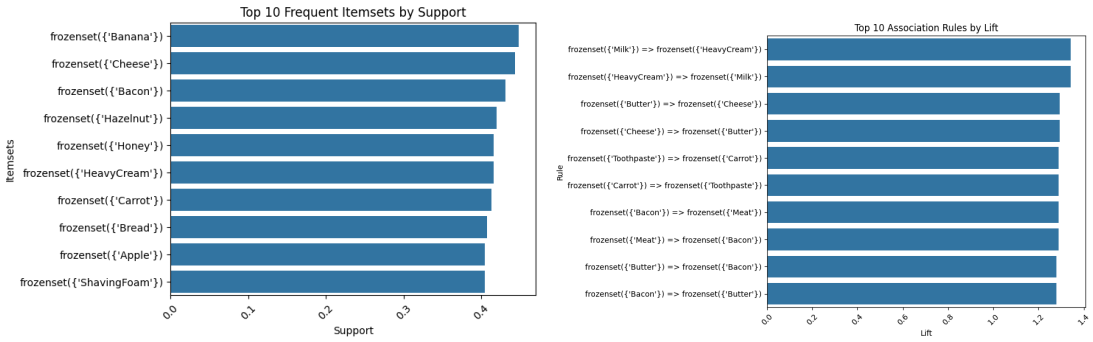


Figure 4.2: Association Rule Mining

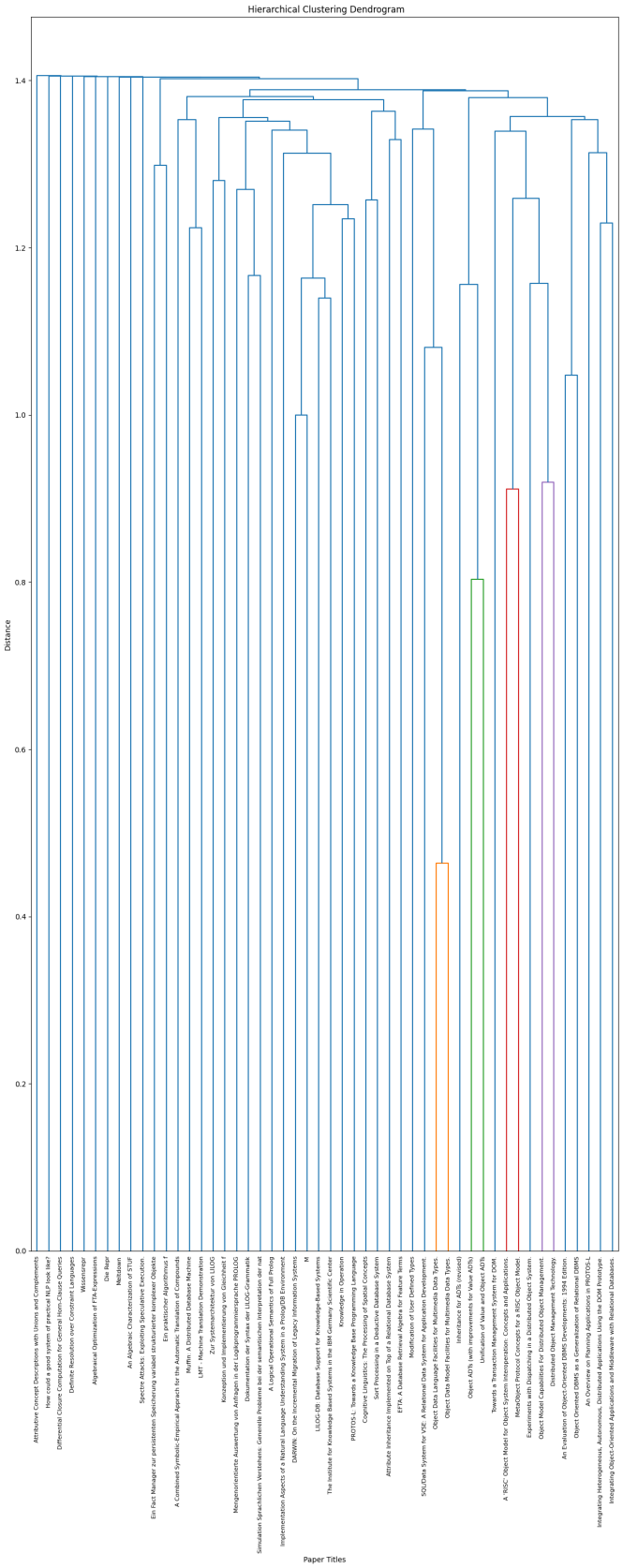


Figure 4.3: Hierarchical Clustering dendogram for DLBP dataset