STUDENT’S NAME: EBENEZER ESSILFIE WILSON

COURSE INSTRUCTOR: DR. KEIONA MIDDLETON

COURSE TITLE: DATA MINING II (TASK 1)

JANUARY 5, 2025

A1.

How can an organization segment its customer base into distinct groups based on their annual income and internet bandwidth usage (Bandwidth\_GB\_Year) to identify high-value customers and optimize service offerings?

This question is relevant to organizations in telecommunications or internet service providers (ISPs) looking to improve customer segmentation, understand customer usage patterns, and create targeted marketing strategies.

A2. Goal

The goal of this analysis is to cluster customers into distinct segments using the k-means clustering technique based on their **annual income** and **internet bandwidth usage**. The K-means clustering algorithm is known for its simplicity and is applied in clustering datasets from different domains. Despite this advantage, its performance is greatly hampered due to some of the problems inherent in its implementation (ScienceDirect, 2022). These segments will help the organization:

* **Identify high-value customers** who use significant bandwidth and have higher incomes, making them potential targets for premium service offerings.
* **Pinpoint low-value customers** who may benefit from cost-effective plans or promotions.
* **Understand usage patterns** to improve resource allocation and optimize bandwidth distribution.

B1. Explanation of Technique

The k-means clustering technique works by partitioning a dataset into k clusters based on the similarity of the data points. The original dataset is clustered into a relatively larger number of high‐density sub‐clusters, and the result is obtained by merging connected sub‐clusters respectively. The connectivity among sub‐clusters is evaluated by the sub‐clusters’ density and the nearest distance class between sub‐clusters (He, He, Wang, & Zhu, 2022). Here's how it works for the given dataset with Income and Bandwidth\_GB\_Year:

Step 1: Initialization

The algorithm starts by randomly selecting k centroids (the initial "center points" for clusters).

Each data point (a combination of Income and Bandwidth\_GB\_Year) is assigned to the nearest centroid based on the Euclidean distance.

Step 2: Iterative Optimization

After initial assignment, centroids are recalculated as the mean of all data points within each cluster. Data points are then reassigned to the closest recalculated centroid. This process repeats until cluster assignments no longer change (convergence).

Expected Outcomes:

Each customer will belong to one of k clusters, representing groups with similar income and bandwidth usage patterns.

Example outcomes:

Cluster 1: Low-income, low-bandwidth customers (e.g., budget-conscious users).

Cluster 2: High-income, high-bandwidth customers (e.g., gamers, heavy streamers).

Cluster 3: High-income, low-bandwidth customers (e.g., premium users with occasional internet needs).

This segmentation will enable actionable insights, such as identifying customers who may benefit from specific plans, promotions, or upgrades.

B2. Assumption  
The k-means algorithm assumes that clusters are spherical and separable in nature, meaning the data points within a cluster are closer to their centroid than to any other cluster's centroid.

* Implication:  
  k-means works best when the clusters in the dataset are relatively well-separated and of similar density. For this reason, normalizing or scaling the variables (e.g., Income and Bandwidth\_GB\_Year) is essential to ensure they contribute equally to the clustering process, as they may have different units and ranges.

B3. Packages and Libraries

1. pandas

Purpose: To manage and manipulate the dataset. An example is extracting dataset with the necessary columns and handling missing values.

2. numpy

Purpose: For numerical computations as it supports fast mathematical operations on arrays and matrices, which are critical for calculating distances in k-means clustering.

3. matplotlib and seaborn

Purpose: To enhance the clustering results visually

4. scikit-learn

Purpose: To perform the k-means clustering.

This includes an efficient implementation of the k-means algorithm (sklearn.cluster.KMeans), making it straightforward to apply clustering to the dataset and also provides functions for preprocessing, such as scaling data (StandardScaler), which is essential for ensuring that Income and Bandwidth\_GB\_Year contribute equally to distance calculations.

C1. Preprocessing Goal

The main data preprocessing goal for the k-means clustering technique is to normalize (scale) the continuous variables, ensuring that both Income and Bandwidth\_GB\_Year contribute equally to the distance calculations.

Relevance to k-Means:

k-means uses Euclidean distance to assign data points to clusters. Since Income and Bandwidth\_GB\_Year likely have different units (e.g., dollars vs. gigabytes per year) and ranges, the variable with the larger range will dominate the distance metric.

Scaling (normalization) ensures that both variables are on the same scale, preventing bias in cluster formation.

Steps in Preprocessing:

Handle missing values: Impute or remove any rows with missing data in the selected columns.

Normalize the continuous variables using StandardScaler (standardization with mean = 0 and variance = 1).

C2. Variables and their types

The variables to be used in the k-means clustering analysis are:

Income (Continuous) and Bandwidth\_GB\_Year (Continuous)

C3. Steps for analysis with codes

1. Load the dataset into a Python environment for processing.

A screenshot of a computer program

Description automatically generated

|  |  |
| --- | --- |
|  |  |

A screenshot of a graph

Description automatically generated

1. Identify and handle any missing values in the dataset to avoid errors during analysis.

A screenshot of a computer

Description automatically generated

1. Extract only the columns needed for clustering (Income and Bandwidth\_GB\_Year).



1. Normalize the data to ensure both variables contribute equally to the clustering process.

Use StandardScaler to standardize the values (mean = 0, standard deviation = 1).

A screenshot of a computer code

Description automatically generated

1. Verify that the data has been correctly scaled and is ready for clustering.

A screenshot of a computer

Description automatically generated

1. Provide a Copy of the Cleaned Data Set

A close-up of a sign

Description automatically generated

**Part IV: Analysis**

D1. Clusters

The Elbow Method is used to determine the optimal number of clusters. This method involves: Calculating the Within-Cluster Sum of Squares (WCSS) for a range of cluster numbers (e.g., k=1 to k=10).

Plotting the WCSS values against the number of clusters (k).

Identifying the "elbow point," where the rate of decrease in WCSS slows significantly, indicating the optimal number of clusters.

* The plot will show a sharp decrease in WCSS initially, followed by a slower rate of decrease.
* The "elbow point" (e.g., k=3) indicates the optimal number of clusters.

A screenshot of a computer program

Description automatically generated

A graph with blue dots

Description automatically generated

A graph of a number of clusters

Description automatically generated

D2. Code to perform clustering

Once the optimal number of clusters (k) is determined (e.g., k=3), the k-means clustering is performed.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

Visualizing the clusters:

To interpret and validate the clustering results using a scatterplot.

A screenshot of a computer code

Description automatically generated

A chart of a customer cluster

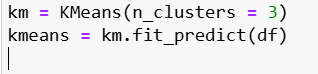
Description automatically generated with medium confidence

A yellow dots with black text

Description automatically generated

A screenshot of a computer code

Description automatically generated



A screenshot of a computer

Description automatically generated

A red and blue dots

Description automatically generated

A red and yellow dots

Description automatically generated

**A red dots on a white background

Description automatically generated**

**Part V: Data Summary and Implications**

E1. Clusters Quality

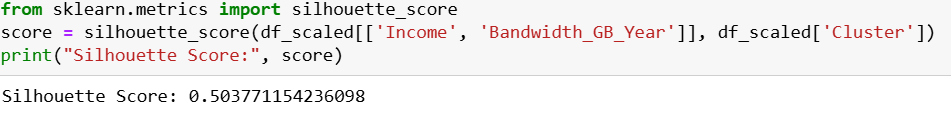
The quality of the clusters can be evaluated using the following metrics and observations:

Compactness (Intra-cluster Variance):

Clusters are compact if data points within each cluster are close to the cluster center. The Within-Cluster Sum of Squares (WCSS), minimized during the k-means process, ensures compact clusters.

Separation (Inter-cluster Distance): Clusters should be well-separated, meaning data points in one cluster are far from points in another cluster. Using visualizations (scatterplots) and silhouette scores can help confirm separation.

Silhouette Score: Measures how similar a point is to its cluster compared to other clusters. Scores range from -1 (poor clustering) to 1 (well-clustered).



A silhouette score above 0.5 generally indicates well-defined clusters.

E2. Results and Implications

The analysis segmented customers into distinct groups based on their Income and Bandwidth\_GB\_Year.

Each cluster represents a unique customer profile:

Cluster 1: Low-income, low bandwidth users (cost-sensitive customers).

Cluster 2: High-income, high bandwidth users (premium customers).

Cluster 3: Moderate-income, moderate bandwidth users (average users).

Implications:

The organization can design targeted marketing campaigns based on cluster profiles:

Cluster 1: Promote budget-friendly plans or discounts.

Cluster 2: Highlight premium services, high-speed internet, or additional features.

Cluster 3: Offer flexible plans that cater to both affordability and performance.

Optimize resource allocation by aligning bandwidth provisioning with customer usage patterns.

Identify opportunities for customer retention by focusing on specific clusters.

E3. Limitations

The main limitation of K-Means for its failure to account for non-spherical distribution is that it does not account for variance in data (Hasan, n.d.).

The analysis relies solely on two variables (Income and Bandwidth\_GB\_Year), which might oversimplify customer behavior. Other relevant variables (e.g., customer satisfaction, geographic location, or internet service type) were not included in the clustering, potentially reducing the depth of insights.

E4. Recommendations

Based on the clustering results, the organization should:

Implement Targeted Marketing Campaigns:

Tailor marketing strategies for each cluster. For instance:

Cluster 1: Focus on affordability and highlight promotions.

Cluster 2: Promote premium or add-on services.

Cluster 3: Offer mid-tier plans or customizable options.

Bandwidth Optimization:

Adjust infrastructure to meet the demand patterns identified in the clusters, ensuring high bandwidth availability for Cluster 2.

Data Enrichment for Future Analysis:

Incorporate additional variables (e.g., demographic or service satisfaction metrics) for a more comprehensive segmentation.

F. Panopto Video

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=06770c51-2c58-4903-8584-b264001cf5be>

G.

ScienceDirect. (2022). Title of the article. Retrieved December 27, 2024, from <https://www.sciencedirect.com/science/article/abs/pii/S0020025522014633>

Hasan, R. (n.d.). Clustering-K-Means-All-You-Care-About: Clustering K-Means 8 Limitations I. GitHub. Retrieved December 27, 2024, from <https://github.com/rhasanbd/Clustering-K-Means-All-You-Care-About/blob/master/Clustering-K-Means-8-Limitations%20I.ipynb>

H.

He, H., He, Y., Wang, F., & Zhu, W. (2022). Improved K‐means algorithm for clustering non‐spherical data. Expert Systems, 39(9), 1–23. <https://doi.org/10.1111/exsy.13062>