Assessment Report

on

“Internet Usage Clustering”

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE(AI)

By

Name : Anushka Shakya

Roll Number : 202401100300060

Section: A

Under the supervision of

“MR. BIKKI KUMAR”

KIET Group of Institutions, Ghaziabad

April, 2025

---

1. Introduction

In today's digital age, understanding user internet behaviour is crucial for service providers, marketers, and app developers. Clustering internet usage patterns helps identify user types and customize services accordingly. This project focuses on segmenting users based on their device usage time, categories of websites visited, and the number of sessions per day using clustering techniques.

---

2. Problem Statement

Users consume internet in diverse ways, leading to inconsistent patterns. Without grouping these behaviors, organizations cannot effectively tailor their services. The goal is to analyze and group users into meaningful clusters based on their internet usage metrics.

---

3. Objectives

- To analyze user internet usage behavior
- To preprocess and normalize the data for effective clustering
- To use the K-Means algorithm to group similar users
- To visualize the clustering results through a heatmap

---

4. Methodology

1. Data Collection
   The dataset used contains three features: daily_usage_hours, site_categories_visited, and sessions_per_day. These features represent how much time users spend online, the variety of sites they access, and how frequently they engage with the internet daily.

2. Data Preprocessing
   Data was scaled using StandardScaler to normalize different ranges of values, ensuring fair clustering.
3. Model Building
   The K-Means algorithm was chosen due to its simplicity and effectiveness for unsupervised clustering tasks. The Elbow method was used to determine the optimal number of clusters.
4. Model Evaluation
   Inertia (sum of squared distances) was evaluated for different values of k to choose the best number of clusters. A heatmap was generated to analyze the average behavior per cluster.

---

5. Data Preprocessing

- Selected key features: daily_usage_hours, site_categories_visited, sessions_per_day
- Standardized the features using StandardScaler
- Checked for null values and inconsistencies (none were found in this case)

---

6. Model Implementation

- Used Python with pandas, scikit-learn, matplotlib, and seaborn
- Created an elbow plot to decide the optimal cluster count
- Applied K-Means clustering (k=3)
- Labeled each user with a cluster ID

---

7. Evaluation Metrics

- **Inertia**: Measured within-cluster variance to determine the best value of k

- **Visual Evaluation**: Used a heatmap to compare feature averages across clusters

---

8. Results and Analysis

- The elbow method showed the optimal number of clusters to be 3. After applying K-Means:
- Cluster 0 represented light users
- Cluster 1 represented moderate users
- Cluster 2 represented heavy users with high frequency and diverse site access
- The heatmap clearly showed behavioral distinctions among clusters, helping in future personalization and targeting strategies.

---

9. Conclusion

K-Means clustering effectively grouped users based on internet behavior. This segmentation can help in marketing strategies, resource allocation, and UX personalization. Future improvements can include more features like device type, time-of-day usage, or session duration.

---

10. References

- scikit-learn documentation: https://scikit-learn.org/
- seaborn heatmap documentation: https://seaborn.pydata.org/generated/seaborn.heatmap.html
- "An Introduction to Statistical Learning" by Gareth James et al.

```
[8] import pandas as pd
    from sklearn.preprocessing import StandardScaler
    from sklearn.cluster import KMeans
    import matplotlib.pyplot as plt
    import seaborn as sns
```

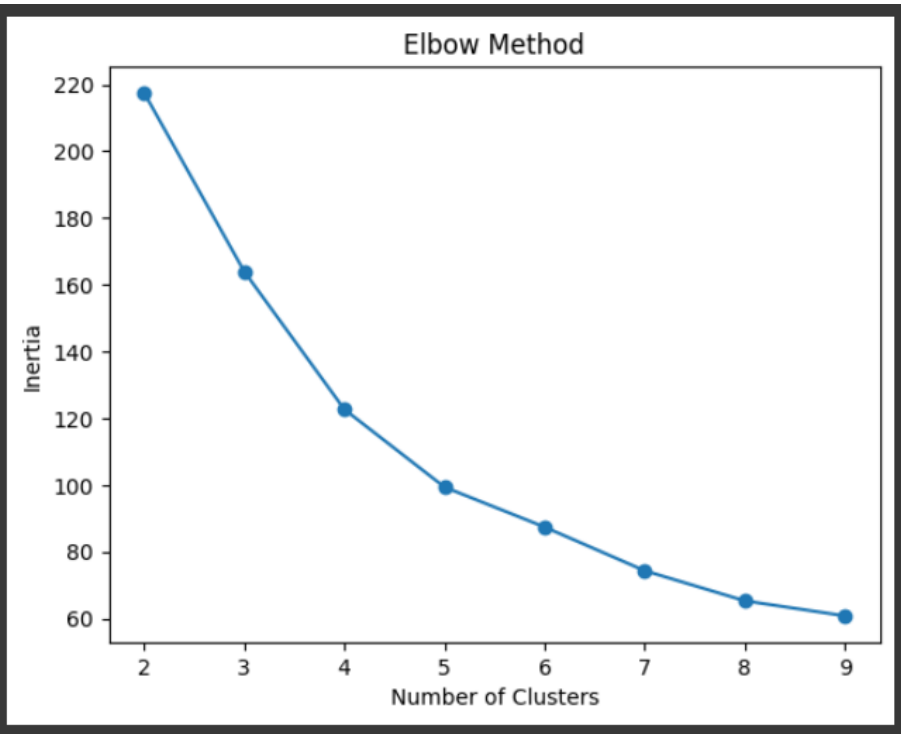
```
0s  df = pd.read_csv("internet_usage.csv")
    print(df.head())
```

```
   daily_usage_hours  site_categories_visited  sessions_per_day
0           9.884957                        2                13
1           1.023220                        9                 1
2          10.394205                        9                 3
3           5.990237                        6                16
4           3.558451                        4                 4
```

```
[11] X = df[['daily_usage_hours', 'site_categories_visited', 'sessions_per_day']]
     scaler = StandardScaler()
     X_scaled = scaler.fit_transform(X)
```

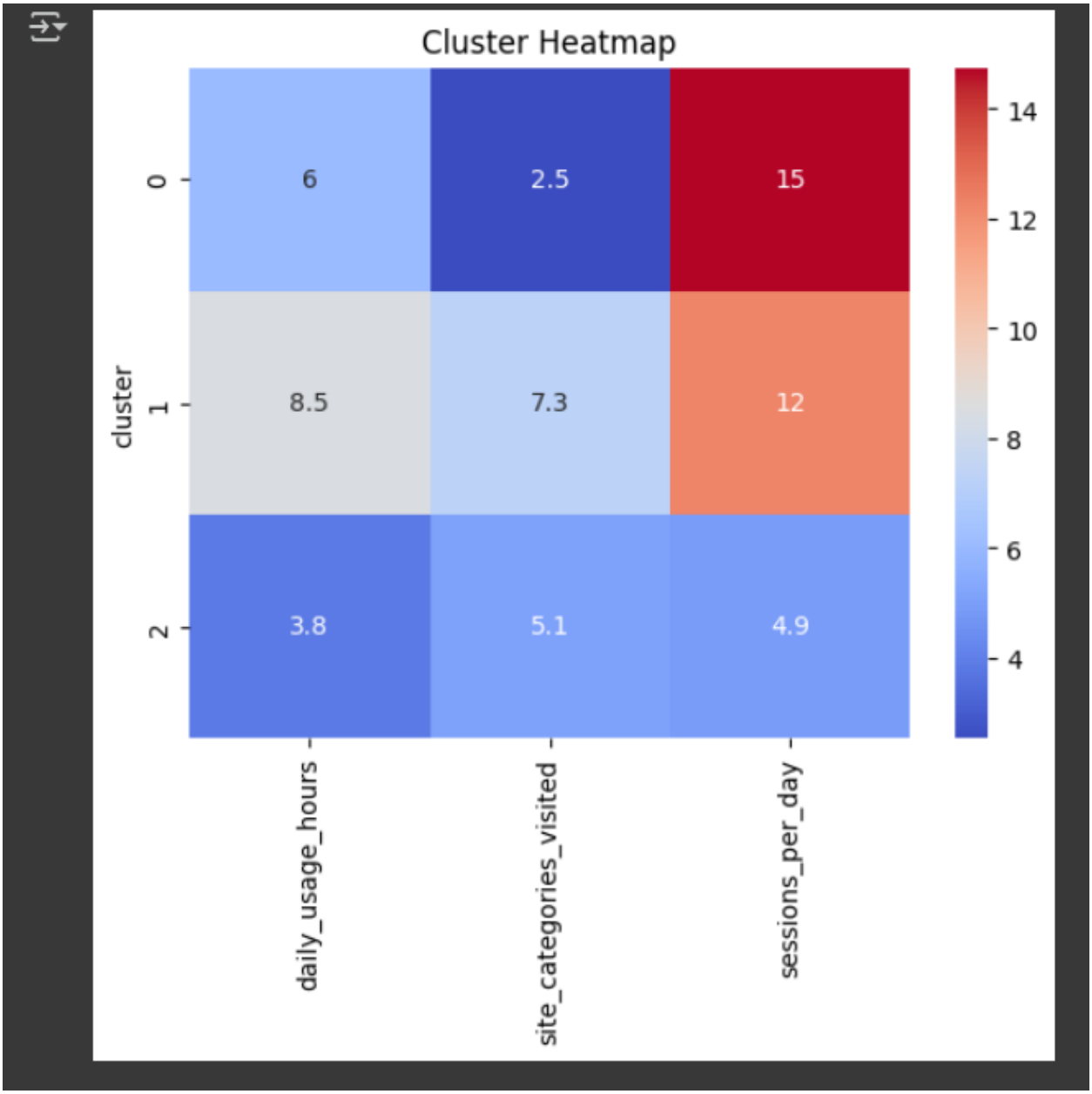
```
[12] inertias = []
     for k in range(2, 10):
         kmeans = KMeans(n_clusters=k, random_state=0)
         kmeans.fit(X_scaled)
         inertias.append(kmeans.inertia_)
```

```
0s [13] plt.plot(range(2, 10), inertias, marker='o')
     plt.title("Elbow Method")
     plt.xlabel("Number of Clusters")
     plt.ylabel("Inertia")
     plt.show()
```




```
[14] kmeans = KMeans(n_clusters=3, random_state=0)
     df['cluster'] = kmeans.fit_predict(X_scaled)
```

```
[15] cluster_means = df.groupby('cluster').mean()
     sns.heatmap(cluster_means, annot=True, cmap='coolwarm')
     plt.title("Cluster Heatmap")
     plt.show()
```


Cluster Heatmap
cluster
0
1
2
6
2.5
15
8.5
7.3
12
3.8
5.1
4.9
daily_usage_hours
site_categories_visited
sessions_per_day
14
12
10
8
6
4