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: Sachin Kumar

Roll Number: 202401100300207

Section: C

KIET Group of Institutions, Ghaziabad

1. Introduction

In today's hyper-connected world, understanding how individuals use the internet is vital for businesses, service providers, and researchers alike. Internet usage patterns can vary widely among users, influenced by factors such as profession, lifestyle, age, and personal interests. These patterns are often reflected in how much time users spend online, the types of websites they visit, and how frequently they access the internet throughout the day.

This project aims to use clustering—a type of unsupervised machine learning—to segment users into distinct groups based on their internet usage behavior. The three key dimensions considered are:

Daily Usage Hours: Total hours spent online each day.

Site Categories Visited: Different types of websites accessed (encoded as numerical categories).

Sessions Per Day: Frequency of accessing the internet or visiting websites.

By applying clustering algorithms to this data, we can uncover meaningful patterns in how users interact with the web. These patterns can then be used for a wide range of applications, such as personalized recommendations, improved service planning, and more effective content delivery.

2. Methodology

The methodology used in this classification problem consists of the following steps:

a. **Data Loading:**

- The dataset is imported using Python's pandas library from a CSV file.

b. **Data Preprocessing:**

- Missing values are handled by dropping or imputing them using statistical methods.
- Standardization is applied using StandardScaler to ensure uniform scaling of features.
- Feature selection includes daily_usage_hours, site_categories_visited, and sessions_per_day.

c. Clustering Model Selection:

- The **K-Means algorithm** is chosen for clustering users based on their internet usage patterns.
- The **Elbow Method** is used to determine the optimal number of clusters by analyzing inertia values.

d. **Model Training:**

- K-Means clustering is applied with the optimal number of clusters (k=3 chosen based on Elbow Method results).
- Users are assigned to clusters based on their internet usage attributes.

e. Visualization:

- Several graphs are used to analyze clustering results:
    - **Elbow Method graph** to determine the best k value.
    - **Scatter plots** to visualize the clustering patterns.
    - **Histograms** to observe session frequency distribution among clusters.
    - **Box plots** for comparing daily usage across clusters.
    - **Pair plots** for feature correlation insights.

f. Model Evaluation:

- Cluster assignments are examined based on feature distribution and visual patterns.
- Insights from clustering are analyzed to provide recommendations on user behavior segmentation.

3. CODE :

```
# Importing necessary libraries

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

from sklearn.preprocessing import StandardScaler
```

```
# Load the dataset
# Replace the path with your local or hosted dataset path if needed
df = pd.read_csv("/content/internet_usage.csv")

# Display available column names for reference
print("Available columns:", df.columns)

# Selecting relevant features for clustering
X = df[['daily_usage_hours', 'site_categories_visited', 'sessions_per_day']]

# Removing rows with missing values to ensure clean input for clustering
X = X.dropna()

# Standardizing the features using StandardScaler
# This is important because K-Means is distance-based and sensitive to scale
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Finding the optimal number of clusters using the Elbow Method
inertia = []  # List to store the inertia values for different cluster counts

# Testing k values from 1 to 10
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
```

```
    kmeans.fit(X_scaled)
    inertia.append(kmeans.inertia_)  # Inertia measures how tightly the data is clustered

# Plotting the Elbow Curve
plt.figure(figsize=(8, 5))
plt.plot(range(1, 11), inertia, marker='o', linestyle='--')
plt.xlabel("Number of Clusters (k)")
plt.ylabel("Inertia")
plt.title("Elbow Method for Optimal Clusters")
plt.grid(True)
plt.show()

# Based on the elbow curve, we choose k=3 for clustering (as an example)
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
df['Cluster'] = kmeans.fit_predict(X_scaled)  # Assigning cluster labels to original dataframe

# Visualizing the clusters using the first two features
plt.figure(figsize=(8, 5))
sns.scatterplot(x=X_scaled[:, 0], y=X_scaled[:, 1], hue=df['Cluster'], palette="viridis")
plt.xlabel("Daily Usage Hours (standardized)")
plt.ylabel("Site Categories Visited (standardized)")
plt.title("User Clustering Based on Internet Usage")
plt.legend(title="Cluster")
plt.show()
```

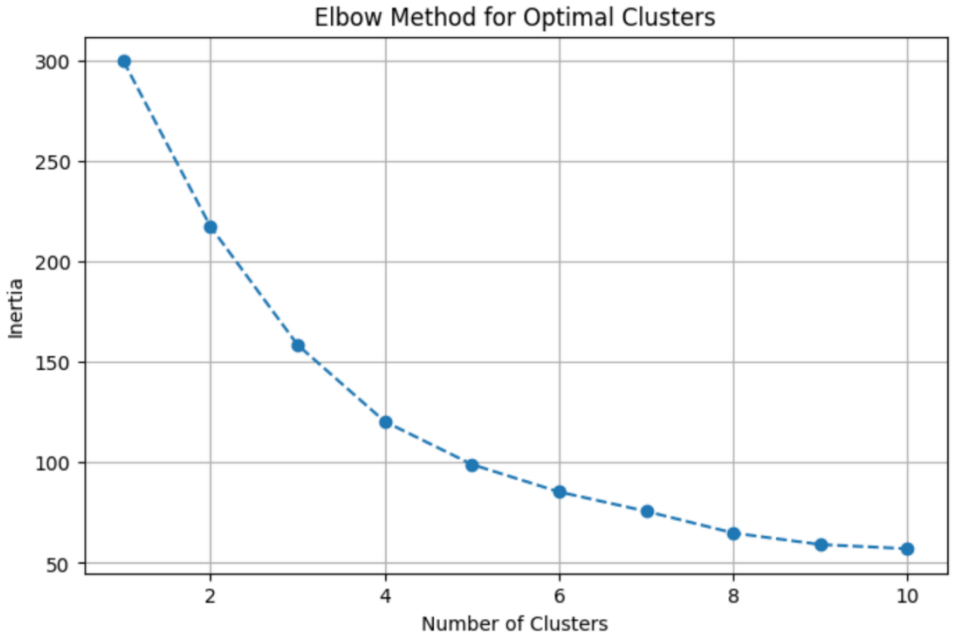
Displaying the first few records of the dataframe with cluster assignments

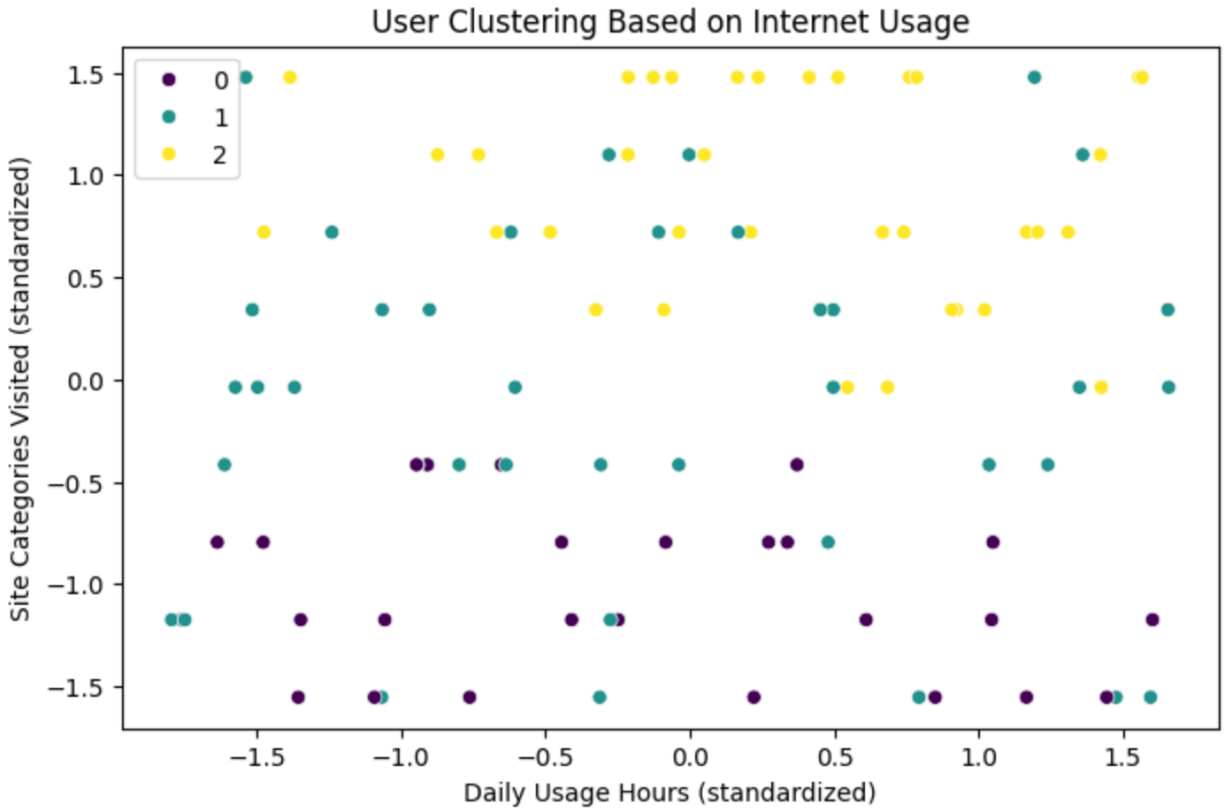
print(df.head())

---

4. Output

```
Available columns: Index(['daily_usage_hours', 'site_categories_visited', 'sessions_per_day'], dtype='object')
```





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

6. References

1. **Jain, A. K. (2010). Data Clustering:** 50 Years Beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011
2. **Seaborn Documentation:** Visualization library used to create clear and attractive data visualizations. https://seaborn.pydata.org/