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 : Aditya Singh

Roll Number : 202401100300020

Section: A

Under the supervision of

"Bikki Kumar"

KIET Group of Institutions, Ghaziabad

22 May, 2025

---

1. Introduction

In the digital age, understanding user behavior on the internet has become essential for optimizing user experience, content delivery, and network management. Clustering users based on their usage patterns helps identify distinct behavior groups, allowing businesses and service providers to better cater to each segment.

---

2. Problem Statement

The goal is to group internet users based on their daily device usage time, frequency of internet sessions, and types of site categories visited. This clustering allows for identifying different types of users such as heavy users, casual users, or niche content users.

---

3. Objectives

- To explore and analyze internet usage patterns.
- To apply clustering techniques to group similar users.
- To visualize the clusters and interpret usage trends.
- To evaluate the clustering model performance and derive insights.

---

4. Methodology

- **Data Collection**: The user uploads a CSV file containing the dataset.

- **Model Building**:

  - **Algorithm Used**: K-Means Clustering.
  - **Number of Clusters**: 3 (chosen manually, can be optimized using elbow method).
  - **Clustering Goal**: Minimize intra-cluster variance and maximize inter-cluster difference.

- **Model Evaluation**:

- **Visual Evaluation**: 2D scatter plot (daily_usage_hours vs. sessions_per_day) with color-coded clusters.
- **Interpretability**: Cluster centers were transformed back to original scale for interpretation

---

5. Data Preprocessing

The dataset is cleaned and prepared as follows:

- **Standardization**: Used `StandardScaler` to normalize the features (`daily_usage_hours`, `site_categories_visited`, and `sessions_per_day`) to ensure equal contribution to clustering.
- **Feature Selection**: Selected all three features relevant to usage behavior.

---

6. Model Implementation

The model was implemented using the **K-Means clustering algorithm**, a popular unsupervised machine learning technique that partitions the data into *K* distinct, non-overlapping clusters based on feature similarity. The goal of K-Means is to minimize the **intra-cluster variance** while maximizing **inter-cluster separation**.

---

7. Evaluation Metrics

For clustering (unsupervised learning), traditional classification metrics (accuracy, precision, recall) are not applicable. Instead, the following are used:

- **Inertia**: Sum of squared distances from each point to its assigned cluster center.

- **Silhouette Score** (optional): Measures how similar a point is to its own cluster vs. other clusters.
- **Cluster Interpretability**: Analyzing centers and distributions of each feature.

---

8. Results and Analysis

- **Cluster 0**: Likely represents **heavy users** – high daily usage and frequent sessions.
- **Cluster 1**: Likely **casual users** with low usage and fewer sessions.
- **Cluster 2**: Possibly **moderate users** with balanced usage patterns.

The scatter plot clearly shows separation between the groups, indicating that K-Means clustering effectively grouped users based on behavior.

---

9. Conclusion

This project successfully applied K-Means clustering to segment internet users into meaningful groups based on their behavior. These insights can be used for targeted marketing, personalized recommendations, or bandwidth optimization.

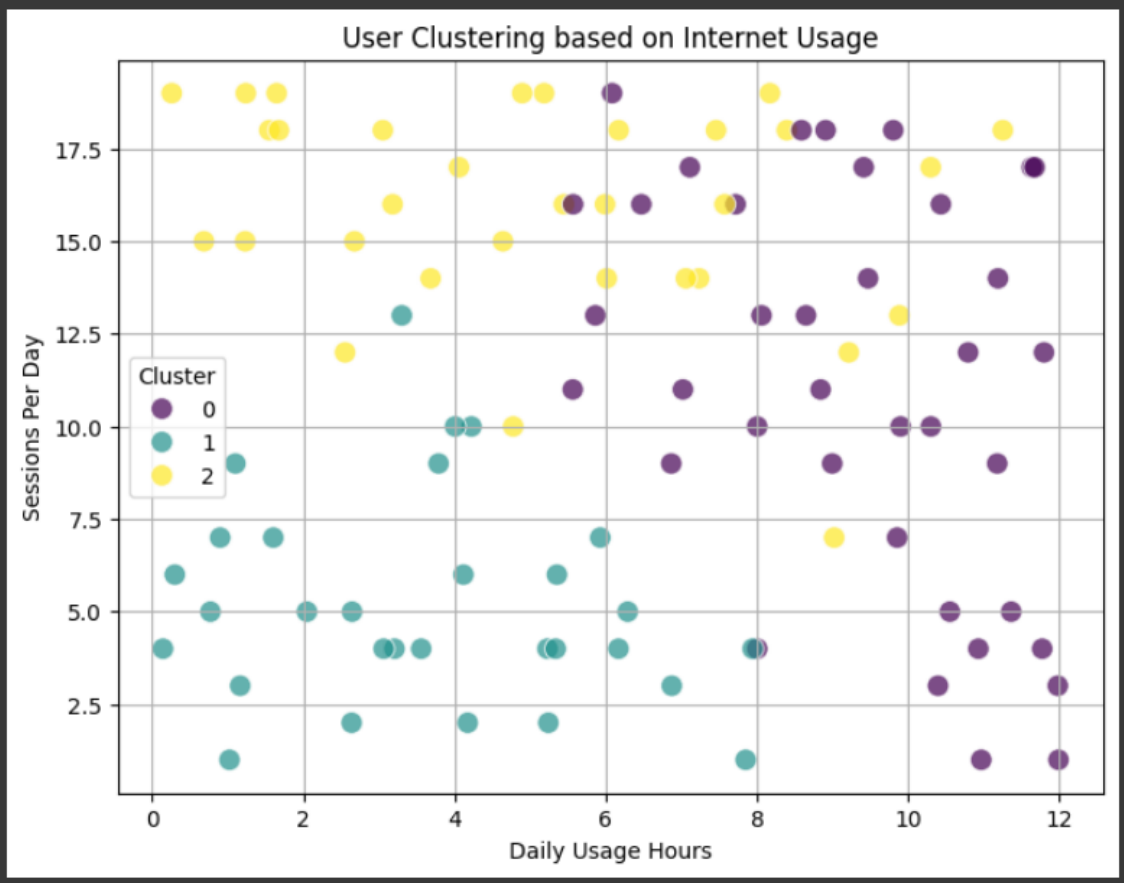
---

10. References

- scikit-learn documentation
- pandas documentation
- Seaborn documentation

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

# Load the dataset
df = pd.read_csv('internet_usage.csv')

# Standardize the features
scaler = StandardScaler()
scaled_data = scaler.fit_transform(df)

# Apply KMeans with 3 clusters
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(scaled_data)

# Adding cluster label to dataframe
df['Cluster'] = clusters

# 2D scatter plot: usage hours vs sessions per day
plt.figure(figsize=(8, 6))
sns.scatterplot(
    x='daily_usage_hours',
    y='sessions_per_day',
    hue='Cluster',
    palette='viridis',
    data=df,
    s=100,
    alpha=0.7
)
```

```python
)
plt.title('User Clustering based on Internet Usage')
plt.xlabel('Daily Usage Hours')
plt.ylabel('Sessions Per Day')
plt.legend(title='Cluster')
plt.grid(True)
plt.show()

# Optional: Show cluster centers in original scale
cluster_centers = scaler.inverse_transform(kmeans.cluster_centers_)
cluster_centers_df = pd.DataFrame(cluster_centers, columns=df.columns[:-1])
print("Cluster Centers (Original Scale):")
print(cluster_centers_df)
```



Cluster Centers (Original Scale):

	daily_usage_hours	site_categories_visited	sessions_per_day
0	9.307566	6.631579	11.131579
1	3.666101	5.166667	5.200000
2	5.192716	3.218750	15.875000