INTERENT USAGE CLUSTERING

INTRODUCTION TO AI

MSE 2

Name – Prateek Baliyan

Roll number – 202401100300177

Branch & sec – CSE(Ai) – C

KIET GROUP OF INSTITUTION , GHAZIABAD

22 – April - 2025

INTRODUCTION

With the increasing reliance on the internet across all demographics, understanding user behavior has become essential for service providers, marketers, and data analysts. This project focuses on clustering internet users based on their online activity patterns, including:

- **Daily usage hours**
- **Types of sites visited (site categories)**
- **Browsing frequency (sessions per day)**

By identifying user segments with similar behaviors, stakeholders can tailor experiences, allocate resources efficiently, and design personalized services. The project uses **unsupervised learning techniques** to uncover hidden user groups without any predefined labels.

METHODOLOGY

Dataset

The dataset used contains 100 records with the following features:

- daily_usage_hours – total hours a user spends online daily
- site_categories_visited – number of distinct categories of sites visited
- sessions_per_day – number of internet browsing sessions initiated in a day

Steps Involved

1. **Data Preprocessing**
    - Loaded the dataset from CSV.
    - Standardized all numerical features using **StandardScaler** to bring them to a common scale.
2. **Clustering (KMeans)**
    - Implemented the **KMeans** algorithm to divide users into distinct groups based on their usage patterns.
    - Initially selected **k=3** clusters.
    - PCA (Principal Component Analysis) was used to reduce dimensions for visualization.
3. **Visualization**
    - Applied **PCA** to reduce the 3D user behavior data into 2D.
    - Visualized clusters using a scatter plot with each cluster color-coded.
4. **Evaluation**

- Calculated clustering performance metrics:
  - **Silhouette Score**: 0.30 – Indicates moderate separation between clusters.
  - **Davies-Bouldin Index**: 1.16 – A lower score signifies better clustering.
- Used the **Elbow Method** to help determine the optimal number of clusters by plotting inertia (within-cluster sum of squares) against k values from 1 to 10.

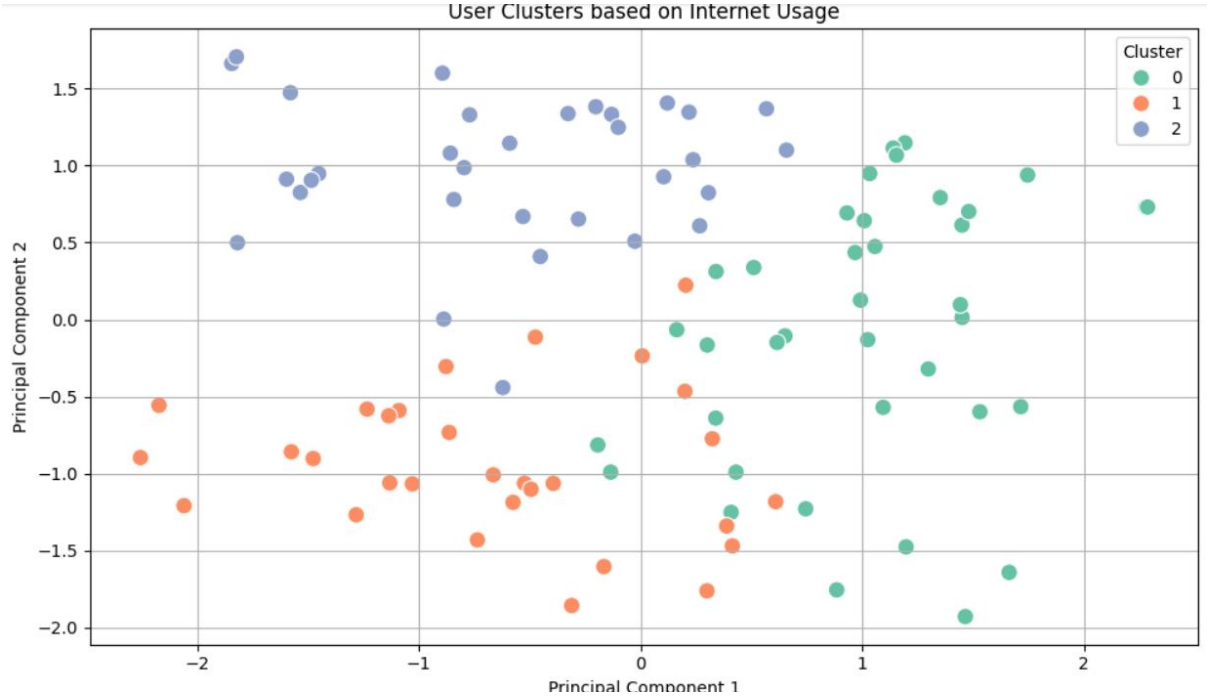
Code

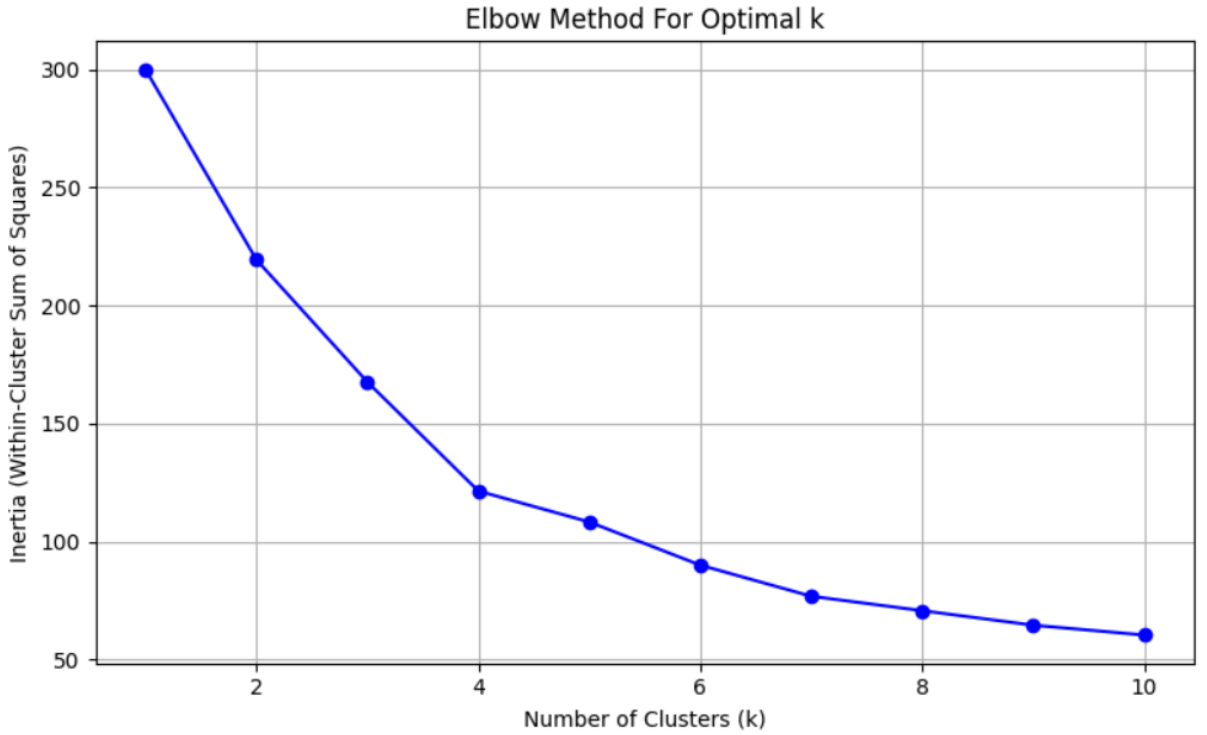
```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score, davies_bouldin_score
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
file_path = "/content/internet_usage.csv"
df = pd.read_csv(file_path)

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

# Apply KMeans clustering (k=3)
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(scaled_data)
df['cluster'] = clusters

# PCA for 2D visualization
```

```
pca = PCA(n_components=2)
reduced_data = pca.fit_transform(scaled_data)
df['PC1'] = reduced_data[:, 0]
df['PC2'] = reduced_data[:, 1]

# Plotting the clusters
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='PC1', y='PC2', hue='cluster', palette='Set2', s=100)
plt.title("User Clusters based on Internet Usage")
plt.xlabel("Principal Component 1")
plt.ylabel("Principal Component 2")
plt.legend(title="Cluster")
plt.grid(True)
plt.tight_layout()
plt.show()

# Evaluation Metrics
sil_score = silhouette_score(scaled_data, clusters)
db_index = davies_bouldin_score(scaled_data, clusters)
print(f"Silhouette Score: {sil_score:.3f}")
print(f"Davies-Bouldin Index: {db_index:.3f}")

# Elbow Method to determine optimal number of clusters
```

```
inertia_values = []
k_range = range(1, 11)
for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_data)
    inertia_values.append(kmeans.inertia_)

# Plotting the Elbow Curve
plt.figure(figsize=(8, 5))
plt.plot(k_range, inertia_values, 'bo-')
plt.xlabel('Number of Clusters (k)')
plt.ylabel('Inertia (Within-Cluster Sum of Squares)')
plt.title('Elbow Method For Optimal k')
plt.grid(True)
plt.tight_layout()
plt.show()
```

Output


User Clusters based on Internet Usage
Cluster
0
1
2
Principal Component 2
Principal Component 1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2
−1
0
1
2


Elbow Method For Optimal k
Inertia (Within-Cluster Sum of Squares)
Number of Clusters (k)
300
250
200
150
100
50
2
4
6
8
10

References

- All the data is taken from the dataset named internet_usage.csv