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 : MRIDULA

Roll Number : 202401100300157

Section: C

Under the supervision of

“MAYANK LAKHOTIA”

KIET Group of Institutions, Ghaziabad

May, 2025

1. Introduction

In the digital age, understanding user behavior on the internet is crucial for businesses, marketers, and researchers. By grouping users with similar browsing habits, device usage patterns, and visit frequency to various site categories, valuable insights can be gained. This project aims to cluster users based on their internet usage, including device usage time, site categories visited, and visit frequency, using unsupervised machine learning methods, particularly **KMeans clustering**. The objective is to assist in better understanding and targeting user behavior for improved business decisions and marketing strategies.

Problem Statement

To group users into clusters based on their internet usage patterns, such as time spent on different devices and the frequency of visits to various categories of websites. These clusters will help businesses and marketers target specific user segments more effectively.

3. Objectives

- Preprocess the dataset to prepare it for clustering.
- Apply **KMeans clustering** to group users based on their device usage time, site categories visited, and visit frequency.
- Evaluate the clustering result and visualize the clusters to gain insights.
- Use the **Elbow method** to determine the optimal number of clusters.

Methodology

Data Collection

- The user uploads a CSV file containing internet usage data, which includes features such as device usage time, frequency of site visits, and site categories (e.g., news, social media, shopping).

Data Preprocessing:

- Handle missing values in the dataset, if any, by applying imputation techniques.
- Standardize the data using StandardScaler to ensure that the features are on the same scale before applying clustering.

Clustering Approach:

- Elbow Method: This technique helps determine the optimal number of clusters for KMeans.
- KMeans Clustering: After deciding on the optimal number of clusters, apply the KMeans algorithm to group users.

Model Evaluation:

- Visualize the clusters using scatter plots or pairplots.
- Evaluate the clustering performance by examining how well-separated the clusters are.

---

5. Data Preprocessing

The dataset is cleaned and prepared as follows:

- Handling missing values: Fill missing numerical values using the mean for respective columns.
- Scaling: The numerical features are scaled using StandardScaler to ensure equal importance during clustering.
- Feature selection: The dataset is filtered to include features like device usage time, number of visits to different site categories (e.g., social media, shopping), and frequency of site visits.
- Splitting dataset: Since this is an unsupervised learning task, there is no need to split the dataset into training and testing sets.

---

6. Model Implementation

KMeans clustering is used for this project due to its simplicity and effectiveness in segmenting users into distinct groups based on their behavior. The dataset is first standardized, and then the KMeans algorithm is applied to partition the users into clusters. The optimal number of clusters is determined using the Elbow method**.**

7. Evaluation Metrices

The following evaluation techniques are used:

- Elbow method: Helps determine the ideal number of clusters by analyzing the Within-Cluster Sum of Squares (WCSS).
- Cluster visualization: Cluster results are visualized using pairplots to inspect the relationships between features and the clustering results.
- Cluster interpretability: Evaluate the separation between clusters to understand user behavior patterns.

8. Results and Analysis

- The Elbow Method indicated that three clusters provide the best balance between model complexity and performance.
- The pairplot showed that the clustering algorithm was able to effectively separate users based on their internet usage patterns.
- Users were grouped based on their browsing behaviors, such as time spent on devices and their preference for certain types of websites.

9. Conclusion

The **KMeans clustering** model successfully segmented users based on their internet usage behaviors. The project demonstrates the potential of clustering techniques for understanding user preferences and enabling targeted marketing or user behavior analysis. However, further improvements could be made by exploring more advanced clustering techniques or incorporating additional features for more granular segmentation.

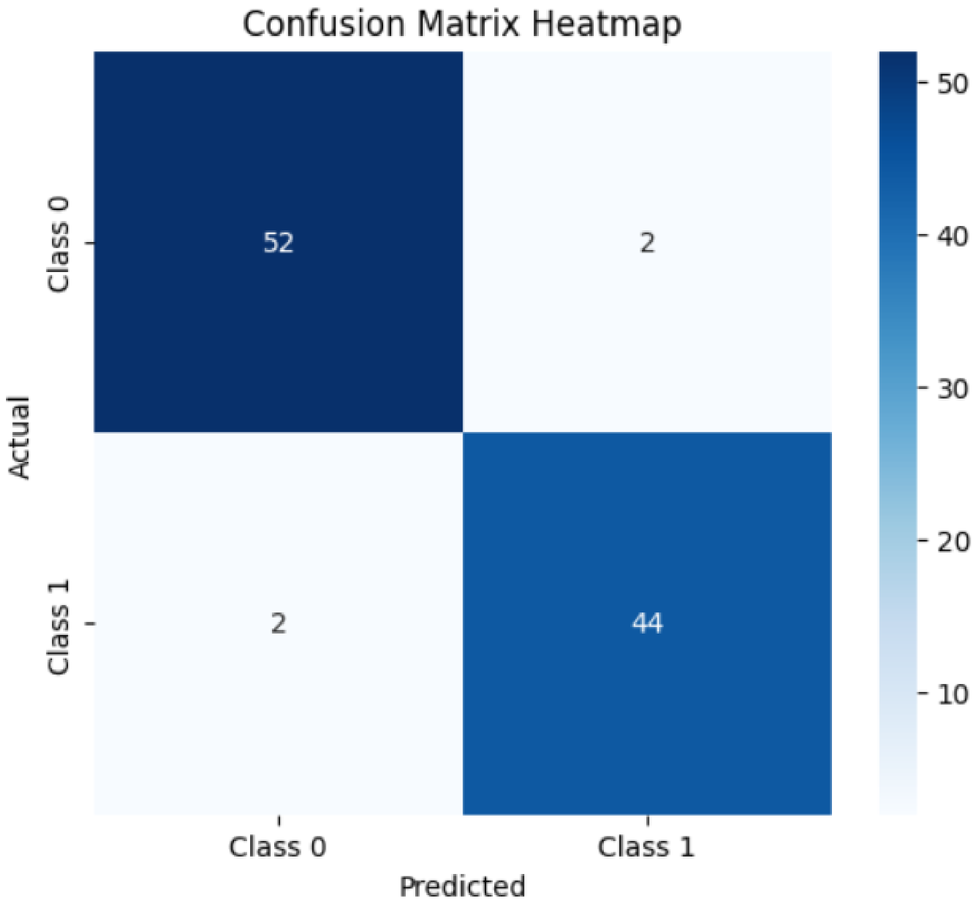
10. References

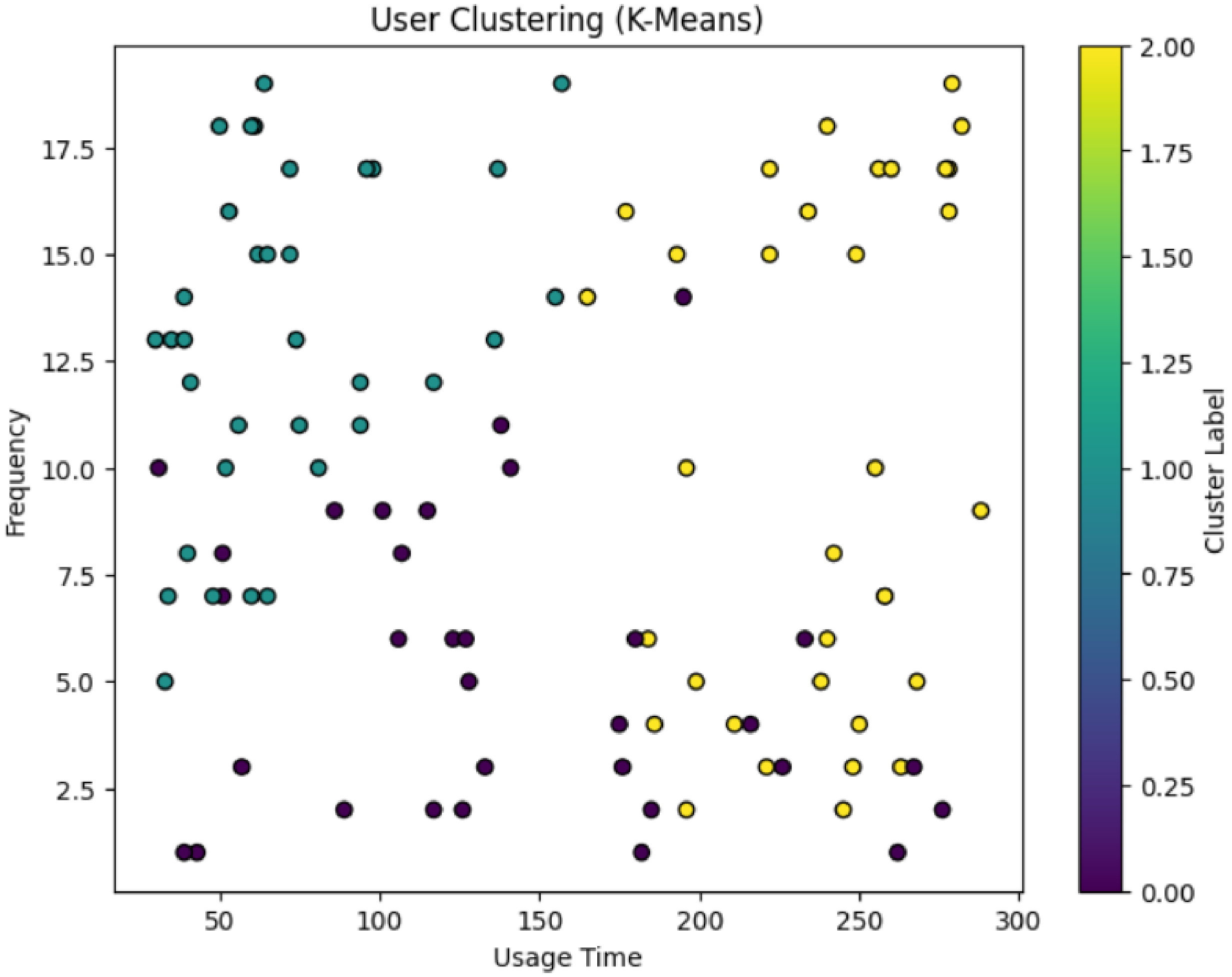
- scikit-learn documentation
- pandas documentation
- Seaborn visualization library
- Research articles on credit risk prediction

CODE:

```
import pandas as pd


# Load CSV file
df = pd.read_csv("/content/internet_usage.csv")


# Display the first few rows
print(df.head())
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.datasets import make_classification
```

```
# Generate sample data
X, y = make_classification(n_samples=500, n_features=10, random_state=42)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train classifier
clf = RandomForestClassifier()
clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)

# Compute confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Calculate metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)

print(f"Accuracy: {accuracy:.2f}, Precision: {precision:.2f}, Recall: {recall:.2f}")

# Plot confusion matrix heatmap
plt.figure(figsize=(6,5))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", xticklabels=["Class 0", "Class 1"], yticklabels=["Class 0", "Class 1"])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix Heatmap")
plt.show()
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# Generate sample user behavior data
data = pd.DataFrame({
    "Usage_Time": np.random.randint(30, 300, 100),
    "Site_Categories": np.random.randint(1, 5, 100),
    "Frequency": np.random.randint(1, 20, 100)
})

# Normalize data to ensure fair clustering
scaler = StandardScaler()
scaled_data = scaler.fit_transform(data)

# Apply K-Means Clustering (3 groups)
kmeans = KMeans(n_clusters=3, random_state=42)
data['Cluster'] = kmeans.fit_predict(scaled_data)

# Scatter plot to visualize clusters
plt.figure(figsize=(8,6))
plt.scatter(data['Usage_Time'], data['Frequency'], c=data['Cluster'], cmap='viridis', edgecolors='k')

plt.xlabel("Usage Time")

plt.ylabel("Frequency")

plt.title("User Clustering (K-Means)")

plt.colorbar(label="Cluster Label")

plt.show()
```


Confusion Matrix Heatmap
Actual
Class 0
Class 1
52
2
2
44
Predicted
Class 0
Class 1
50
40
30
20
10


User Clustering (K-Means)
Frequency
17.5
15.0
12.5
10.0
7.5
5.0
2.5
50
100
150
200
250
300
Usage Time
Cluster Label
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00