Capstone Project

Assignment 2

Course code: CSA1643

Course: Data warehousing and data mining for data science

S. No: 1

Name: A. Mani Bhumika

Reg. No: 192211494

Slot: C

Title: Social Media User Segmentation for Targeted Advertising in Data Warehousing

Assignment Release Date:

Assignment Preliminary Stage (Assignment 1) submission Date:

Mentor Name: DR. Kanchana

Mentor Phone number and Department: department of industrial mathematics

```
# Load required libraries
library(dplyr)      # For data manipulation
library(ggplot2)    # For data visualization
library(cluster)    # For K-means clustering

# 1. Data Preparation
# Read the data (replace "social_media_data.csv" with your dataset)
social_media_data <- read.csv("social_media_data.csv")

# Check if the dataset is successfully loaded
if (is.null(social_media_data)) {
  stop("Error: Unable to load the dataset. Please check the file path.")
}

# Perform any necessary data cleaning and preprocessing steps here
# Add data cleaning and preprocessing steps if required

# 2. Feature Selection
# Select relevant features for segmentation
selected_features <- social_media_data [, c("age",
"engagement_score")]

# Check if selected features contain any missing values
if (anyNA(selected_features)) {
  stop ("Error: Missing values detected in selected features. Please handle missing data.")
}

# 3. Data Standardization (if necessary)
# Standardize numerical features to have mean = 0 and standard deviation = 1
scaled_features <- scale(selected_features)

# 4. Segmentation (K-means Clustering)
# Determine the number of clusters (K)
k <- 5  # Number of clusters

# Apply K-means clustering algorithm
kmeans_result <- kmeans(scaled_features, centers = k)

# Get cluster labels for each data point
cluster_labels <- kmeans_result$cluster

# 5. Visualization
# Visualize the clusters in a scatter plot
# (Note: You may need to adjust the plotting variables based on your dataset)
ggplot(data = social_media_data, aes(x = age, y = engagement_score, color = factor(cluster_labels))) +
  geom_point() +
  labs(title = "Social Media User Segmentation",
       x = "Age",
       y = "Engagement Score") +
  scale_color_discrete(name = "Cluster") +
  theme_minimal()
```

OUT PUT :

```
                          Reference
Prediction                Fraudulent        Non-Fraudulent
Fraudulent                  TP                  FP
Non-Fraudulent              FN                  TN
```