Project Guideline: Marketing Campaign Analysis

Objective:

In this project, students will use Machine Learning techniques to analyze customer behavior based on marketing campaign data. The key tasks involve performing exploratory data analysis, clustering the data, and applying classification techniques to understand customer segments and predict behavior.

Step 1: Perform Exploratory Data Analysis (EDA)

Goals:

• Understand the dataset, identify important features, and examine patterns in customer behavior.
• Explore the relationship between key variables like Income, Spending Score, Age, and Product Category Preference.

Tasks:

1. Summary Statistics: Calculate basic summary statistics such as mean, median, standard deviation, min, and max for continuous variables like Age, Income, and Spending Score.
2. Visualizations:
   • Histograms for Age, Income, and Spending Score to understand distributions.
   • Box plots to check for outliers in the data.
   • Bar plots to visualize categorical variables like Gender, Married, Children, and Product Category Preference.
   • Correlation Matrix: Visualize the correlation between numerical variables.

Insights:

You should come up with your own insights. What is written in this document is to help you understand the process. However, we expect you to explore the data further.
For example:
• What is the general distribution of Income and Spending Score?
• Are there any correlations between Age and Spending Score? Or between Income and Product Category Preference?

Step 2: Cluster Analysis Using K-Means

Goals:

Group customers into distinct clusters based on their spending behavior, income, and age to uncover patterns in customer segmentation.

Tasks:

1. Prepare Data for Clustering:
   - Normalize the data to ensure that all features contribute equally to the clustering algorithm.
2. Determine the Best Number of Clusters:
   - Use K-Means clustering to group the data.
   - Elbow Method: Plot the sum of squared distances (inertia) for different values of k (number of clusters) and identify the "elbow" point where the curve bends.
   - Try values of k ranging from 2 to 10 and observe the plot.
3. Interpret the Clusters:
   - For each cluster, calculate the average Age, Income, and Spending Score.
   - Describe the characteristics of customers in each cluster (e.g., "Cluster 1 consists of young, high-spending customers with moderate income").
4. Visualization:
   - Visualize the clusters using scatter plots (e.g., Spending Score vs Income, colored by cluster).

Step 3: Classification Using Naive Bayes and K-Nearest Neighbors (KNN)

Goals:

For each customer cluster, predict whether a customer will Buy or Not Buy using classification algorithms.

Tasks:

1. Naive Bayes Classification:
   - Train a Naive Bayes classifier on the data.
   - Predict the outcome (Buy/Not Buy) for the test data.
   - Evaluate the performance using a Confusion Matrix.
2. K-Nearest Neighbors (KNN) Classification:
   - Train a K-Nearest Neighbors classifier using the same data.
   - Test the model and evaluate the results using a Confusion Matrix.
3. Comparison of Models:
   - Compare the performance of Naive Bayes and KNN for each cluster.
   - Discuss which model performs better and under what conditions.

Submission Requirements:

1. Python Script:
   • Submit a well-documented Python file containing the code for data analysis, clustering, and classification.
   • Ensure the script includes comments to explain key steps.
   • The script should output important visualizations (Elbow Plot, Cluster Scatter Plots) and classification results (Confusion Matrix).
2. PowerPoint Presentation:
   • Create a PowerPoint presentation that summarizes the findings for a client.
   • The presentation should include:
      • Introduction: Briefly explain the goal of the analysis.
      • EDA Results: Key insights from the exploratory data analysis.
      • Clustering Results: Visualizations and explanations of the customer segments found via K-Means clustering.
      • Classification Results: A summary of the classification models, their performance, and which model performed best.
      • Business Implications: Explain how the clustering and classification results can be used by the marketing team to better target promotions and improve customer engagement.

Final Notes:

• Data Cleaning: Ensure the data is cleaned and preprocessed (handle missing values, normalize features where necessary).
• Visualization: Use meaningful visualizations to support your conclusions.
• Presentation: The PowerPoint should be visually appealing, concise, and tailored for a non-technical audience (e.g., the marketing department).

Deadline:

The final submission (Python script and PowerPoint) is due by 12th October 2024.