CLUSTERING ANALYSIS

PART A (30 Points)

1. Discuss the usability of clustering as an unsupervised learning technique. Include three applications (with examples) to support your discussion. **(6 points)**
2. How do we select the optimal number of clusters in K-Means Clustering? Discuss scenarios if we use a redundant or lesser number of clusters with an example. **(7 points)**
3. Discuss the advantages and disadvantages of Hierarchical Clustering and K-Means Clustering. Which clustering algorithm will you use to deal with a large data set? Discuss your rationales. **(7 points)**
4. Differentiate feature selection and feature extraction. Explain Principal Component Analysis to extract features from data. Discuss the pros and cons of Principal Component Analysis in the case of feature extraction. **(10 points)**

PART B (45 Points)

5. Consider the "USArrests" data. It is a built-in dataset you may directly get in RStudio. Perform hierarchical clustering on the observations (states) and answer the following questions.

   (a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. **(5 points)**

   (b) Cut the dendrogram at a height that results in three distinct clusters. Interpret the clusters. Which states belong to which clusters? **(5 points)**

   (c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. **(5 points)**

   (d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer. **(5 points)**

6. A data set named "marketing_campaign.csv" is uploaded to the Exam-1 folder. You will find a data dictionary in the same folder to understand the meaning of each of the columns. Now, using the given data, answer the following questions. Include your works in each of the steps.

a) Data Preprocessing: (10 points)

- Extract two new features (assume that the current date is 01-07-2021, dates are in European format):
  - Customer_Age based on the available information
  - MembershipDays (the length of membership in days)
- Treat Education column as ordinal categories and Encode
- Treat Marital_Status column as nominal categories and encode using dummy variables.
- Exclude the following columns: ID, Year_Birth, Dt_Customer, Marital_Status
- We have 21 columns in our data frame now. Standardize all the columns.

b) Principal Component Analysis: (10 points)

- Now, do Principal Component Analysis on the features available in the data frame and visualize the first 500 observations using PC1 and PC2.

c) Clustering: (10 points)

- Use K-Means Clustering to partition all the observations using PC1 and PC2.
- Discuss the clusters
  - From the aspect of Customer_Age using box plot
  - From the aspect of Education using histogram