SRN

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE20CS932

September 2021: END SEMESTER ASSESSMENT (ESA)
M TECH DATA SCIENCE AND MACHINE LEARNING_ SEMESTER II

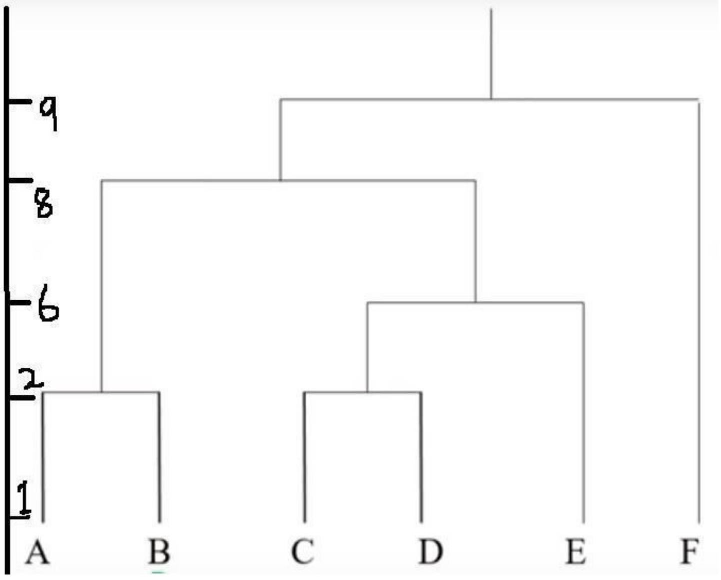
UE20CS932 - MACHINE LEARNING - III

Time: 3 Hrs	Answer All Questions	Max Marks: 100

Instructions

1. Answer all the questions.
2. Section A should be handwritten in the answer script provided and signed at the end of the same.
3. Section B and C are coding questions which have to be answered in the system and uploaded in Olympus Login.
4. Smartly use GridSearchCV as it might affect the system performance.

Section A (20 marks)

1	a)	Compare supervised and Unsupervised Machine Learning models with example	4
	b)	Explain the steps for cluster formation through K-means clustering	4
	c)	What is the need of linkage methods? Compare different linkage methods.	4
	d)	The points in the cluster1 at the final iteration are (2,4),(3,4),(1,3) and (2,5). Compute the cluster inertia.	4
	e)	 Compute the optimal number of clusters for the above dendogram ? What are all the observations (samples) present in each cluster ?	4

SRN

Section B (40 Marks)

2

Dataset Information: cluster_data.csv

The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, they have captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually.

This Dataset consist of

- A 561-feature vector with time and frequency domain variables.
- Its activity label.
- An identifier of the subject who carried out the experiment

Dataset Information: recommendation.csv

The recommendation.csv file consists of 85724 ratings given by 943 users on 1659 products.

It has the following 4 columns:

- UserID
- ItemID
- Rating (Integers 1 to 5)
- Timestamp (Unix time stamp).

Note:

1.Use cluster_data.csv for all the clustering and dimensionality reduction questions

2.Use recommendations.csv for recommendation system questions

3.'activity' column in the cluster_data.csv is the target column. Don't use this column for clustering purposes. You can use this for predictive model building.

(i) Perform required pre-processing and compute how many pairs of variables have the correlation more than 0.8 ? Apply PCA and compute the required number of principal components to capture the 90 percent variance of the original data. Print the Eigenvalues and Eigenvectors of top 5 PCs 10

(ii) Build the K-means clustering model with reduced PCA features (PCs which are explaining 90 percent variance) and compute the optimal value of clusters. Make the business inferences using the cluster groups. 14

(iii) Build/Plot the top 100 cluster dendogram using 4 different linkages and compare its performance. 10

SRN □□□□□□□□□□□□

	(iv)	Cluster the data into 5 groups using K-means and order the clusters in terms of the inertia(WCSS) of each cluster.	6
		Section C (40 marks)	
3	(i)	Build the following ML model and compare its performance: a. ML model with original inp_data and out b. ML model with pca_inp_data and output c. ML model with svd_inp_data and out d. ML model with lda_inp_data and out Note1: The ‘activity’ column in the dataset is the output column (out) Note2: inp_data ▯ All the columns in the original dataset (excluding ‘activity’) pca_inp_data ▯ number of PCA components which captures the 95 percent of variance svd_inp_data ▯ number of SVD components which captures the 95 percent of variance lda_inp_data ▯ required number of LDA components	12
	(ii)	Use the dataset: recommendation.csv Build the popularity based recommendation system and suggest top 5 items.	8
	(iii)	Use the dataset: recommendation.csv Build a collaborative recommendation engine to recommend the top 5 items to the specific user. Measure the model quality in terms of RMSE	20