SRN | | | | | | | | | | | |

PES University, Bengaluru

(Established under Karnataka Act No. 16 of 2013)

UE20CS932

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

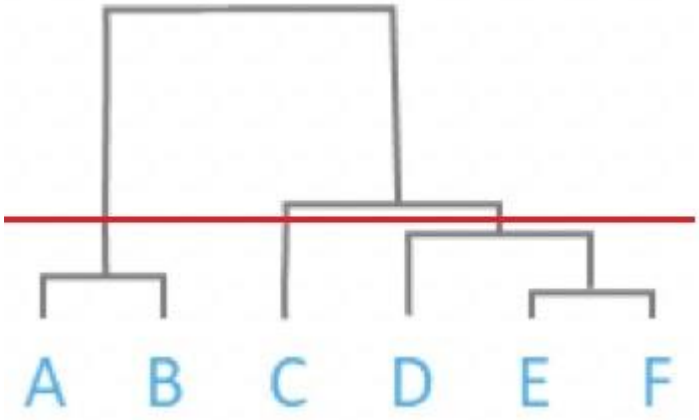
UE20CS932 - MACHINE LEARNING - III

Time: 3 Hrs	Answer All Questions	Max Marks: 80

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.

Section A (20 marks)

1	a)	Explain the role of Silhouette score in measuring the quality of the clusters.	4
	b)	Say the distance between cluster B and A is 4, and the distance between C and A is 6. What is the distance between the cluster (BC) and A using complete and single link method	4
	c)	Write any two characteristics of Principle components derived from the data	4
	d)	At the end of 3rd iterations the two clusters formed with the following observation C1 : (2,3),(4,5),(1,4) C2:(4,6),(3,5) Compute the Euclidian distance between the updated centroids and the point (4,6) during the next iteration.	4
	e)	How many clusters can be formed for the following dedogram with respect to the distance marked as a redline. Also write the samples name exist in each cluster.  	4

Section B (30 Marks)

2		**Dataset Information:** **food_final.csv** This dataset derived from the USDA National Nutrient Database. The data set refers to the various food groups, having nutrient compositions. This Dataset consist of the following features: ID = ID of food group FoodGroup = Name of food group Energy_kcal = Energy content in kilo_calories Protein_g = Protein content in grams Fat_g = Fat content in grams VitC_mg = Vitamin C content in mili_grams Folate_mcg =Folate content in micro_grams Niacin_mg = Niacin content in mili_grams Riboflavin_mg = Riboflavin content in mili_grams Thiamin_mg = Thiamin content in mili_grams Calcium_mg = Calcium content in mili_grams Iron_mg =Iron content in mili_grams Magnesium_mg = Magnesium content in mili_grams Note: Drop the 'FoodGroup' column while computing PCA and forming clusters. **amazon_ratings_Musical_Instruments.csv** This dataset contains the UserID, ItemID, Rating and timestamp information collected for musical instrument products from Amazon. **Recommendation_mini.csv** This dataset contains the UserID, ItemID, Rating and timestamp information collected from a movie rating dataset. Note: Drop the timestamp column while building recommendation models Use food_data.csv for the questions 2(i),2(ii) , 2(iii) ,2(iv) and 3(i) Use amazon_ratings_Musical_instruments.csv for 3(ii) Use recommendation_mini.csv for 3(iii)	
	(i)	Perform the pre-processing techniques required for PCA and clustering. Will PCA lead to dimensionality reduction for this data ? Compute how many number of principle components are capturing the 90 percent variance in this dataset. Print the top 5 Eigen values and Eigen vectors. [Use food_final.csv]	8
	(ii)	Find the optimal number of clusters for the K-means clustering model [Note: Use the PCs	10

		which are explaining the 90% variance]. Make the business inferences using the characteristics of each cluster group.	
	(iii)	Explore the optimal number of cluster using hierarchical clustering through its dendogram. [Plot the top 100 clusters only in the dendogram. Use the best linkage technique]	6
	(iv)	Cluster the data in to 4 groups using K-means and order the clusters in terms of the inertia(WCSS) of each cluster.	6
Section C (30 marks)			
3	(i)	Build the following ML model to predict 'FoodGroup' 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 Note1: The 'FoodGroup' column in the dataset is the output column (out). This column has 25 levels. Note2: inp_data → All the columns in the original dataset (excluding 'FoodGroup') 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	15
	(ii)	Use the dataset: amazon_ratings_Musical_instruments.csv Build the popularity based recommendation system and suggest top 5 items.	7
	(iii)	Use the dataset: recommendation_mini.csv Build collaborative recommendation engine to recommend a top 5 items to the specific user. Measure the model quality in terms of RMSE	8