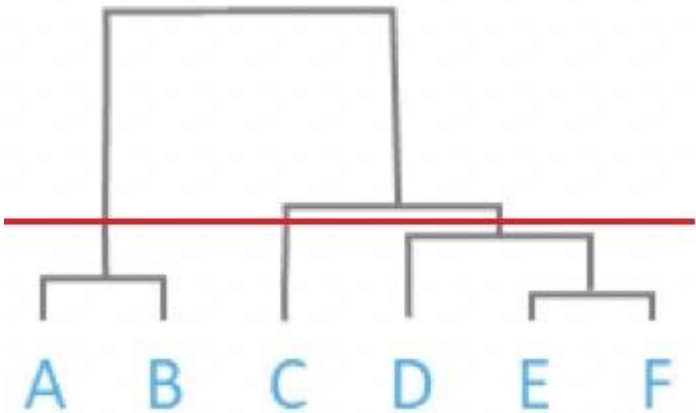
	<b>PES University, Bengaluru</b> (Established under Karnataka Act No. 16 of 2013)		<b>UE20CS932</b>
	<b>April 2022: END SEMESTER ASSESSMENT (ESA)</b> <b>M TECH DATA SCIENCE AND MACHINE LEARNING_ SEMESTER II</b> <b>UE20CS932 - MACHINE LEARNING - III</b>		
	Time: 3 Hrs	Answer All Questions	Max Marks: 80
	<b>Instructions</b>  1. Answer all the questions. 2. Section A should be handwritten in the answer script provided. 3. Section B and C are coding questions which have to be answered in the system and uploaded.		

Section A (20 marks)			
1	a)	List any three distance measures used for clustering with its mathematical expression	4
	b)	Explain the role of Silhouette score in measuring the quality of the clusters.	4
	c)	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
	d)	Compare PCA and SVD.	4
	e)	How many clusters can be formed for the following dendrogram 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:	

		<p>1. Use clust_data.csv for all the clustering and dimensionality reduction questions</p> <p>The simulated sensor data is provided for the analysis (clust_data.csv). Sensor values from 20 different sensors are collected for 4 different operating conditions (0,1,2,3). The sensor information is labeled as F1, F2, ..., F20. Target column contains the operating condition level.</p> <p>2. Use the "Target" column as an output column. Don't use this for dimensionality reduction and clustering. This column you can use to build the ML models using the dimensionality reduced data</p> <p>3. Use electronics_rating.csv for recommendation system questions. This is a subsample version of ratings provided for the electronics items in the Amazon website.</p>	
	(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.	8
	(ii)	Find the optimal number of clusters for the K-means clustering model [Note: Use the PCs which are explaining the 95% variance]. Make the business inferences using the characteristics of each cluster group.	10
	(iii)	Plot and compare the dendrogram using different linkage methods. [Plot the top 100 clusters only in the dendrogram. Use the best linkage technique]	6
	(iv)	Explore the optimal number of cluster using hierarchical clustering through its dendrogram.	6
<b>Section C (30 Marks)</b>			
3	(i)	<p>3 (i) Build the following ML model to predict the 'target' and compare its performance:</p> <p>a. ML model with original inp_data and out</p> <p>b. ML model with pca_inp_data and out</p> <p>c. ML model with svd_inp_data and out</p> <p>Note1: The 'target' column in the dataset is the output column (out). This column has 4 levels 0,1,2,3.</p> <p>Note2:</p> <p>inp_data --&gt; All the columns in the original dataset (excluding 'target')</p> <p>pca_inp_data --&gt; number of PCA components which captures the 95 percent of variance</p> <p>svd_inp_data --&gt; number of SVD components which captures the 95 percent of variance</p>	15
	(ii)	<p>Use the dataset: electronics_rating.csv</p> <p>Build the popularity based recommendation system and suggest top 5 items.</p>	7
	(iii)	<p>Use the dataset: electronics_rating.csv</p> <p>Build collaborative recommendation engine to recommend a top product/item to the specific user. Measure the model quality in terms of RMSE</p>	8