|  |  |  |  |
| --- | --- | --- | --- |
|  | **PES University, Bengaluru**  (Established under Karnataka Act No. 16 of 2013) | | **UE20CS932** |
| **March 2024: 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. 3. Section B and C are coding questions which have to be answered in the system and upload in the given drive link. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Section A (20 marks)** | | | |
| 1 | a) | Write down the steps for the K-means clustering algorithm. | 4 |
| b) | How do you define a group of clusters as a good quality cluster? | 4 |
| c) | What are dimensionality reduction techniques? Write down the steps of computing principal components. | 4 |
| d) | Write a note on any four of the linkage methods. | 4 |
| e) | What is market basket data? Explain briefly the following terms from Association rule mining with an example: support, confidence and lift. | 4 |
|  | | | |
| **Section B (40 Marks)** | | | |
| 2 |  | Use Credit-Card-Dataset-for-Clustering.csv for all the clustering and dimensionality reduction questions.  Use Book\_ratings.csv for popularity-driven recommendation system questions and collaborative recommendation engine.  Use Online Retail.csv for association rule mining using, apriori algorithm questions. |  |
| a) | Perform EDA and pre-processing techniques (remove unnecessary variables, check the defects in the data) required for PCA and clustering. (10 marks)  Print the top 5 Eigenvalues and Eigenvectors. (4 marks) | 14 |
| b) | Find the optimal number of clusters for the K-means clustering model [Note: Use the PCs, which are explaining the 90% variance]. | 6 |
| c) | Plot the dendrograms using 4 linkage methods for the PCA transformed data and identify which one is the best. [Note: Use the PCs, which are explaining the 90% variance]. | 6 |
| d) | Cluster the data into 4 groups and order the cluster quality in terms of the inertia (WCSS) of each cluster. [Use ward linkage metric] | 6 |
|  | e) | Compare the quality of clusters for K-means clustering algorithm on original data and PCA transformed data. | 8 |
| **Section C (40 marks)** | | | |
| 3 | a) | Develop a popularity-driven recommendation system, print Total no of ratings, Total No of Users,Total No of products and recommend the top 5 items.  Use the dataset: Book\_ratings.csv | 10 |
| b) | Build a collaborative recommendation engine. Use KNNBasic library with cosine similarity and measure the model quality by performing cross validation in terms of RMSE.  Use the dataset: Book\_ratings.csv | 20 |
| c) | Create association rule mining using apriori algorithm.Perform basic pre-processing operations required by algorithm ( drop missing values, drop unnecessary columns). (7 marks)  Create the basket only for France. Run algorithm with minimum support 0.07, tune with lift. (3 marks)    Use the dataset: Online Retail.csv | 10 |