JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST-3 EXAMINATION- 2022

B. Tech (CSE/IT) Semester VI

COURSE CODE: 19BIWCI637   **MAX. MARKS:** 35

COURSE NAME: Statistics and Exploratory Data Analytics

COURSE CREDITS: 2   **MAX. TIME:** 2 Hours

Note: Attempt all the questions

---

1. You are given an image file as "image.png". Write a Python program and use OpenCV to perform the following:
   - Load and display the original image
   - Split the image into Red, Green, and Blue arrays
   - Initialize PCA with top 50 principal components
   - Apply PCA to red channel to perform the transformation. Next, Apply inverse transformation to the transformed array.
   - Apply PCA to Green channel and then apply the inverse transform to transformed array.
   - Finally, apply PCA to *blue* channel and then apply the inverse transform to transformed array.
   - Display the compressed image with first 50 principal components.   [10 marks]

2. The ratings of various online books are given in a data set 'bookrating.csv'. The fields include **user_id, book_id**, and **rating** (the rating given by the user for a particular book from 0 to 5).
   Write a Python program to perform the following
   - Display the active users, i.e., the users who reviewed more than ten books
   - Display the popular books, i.e., the books that were reviewed by more than ten users
   - Keep and display only the popular books and active users   [6 marks]

3. You are given a matrix *M*. Write a Python program to decompose the matrix into three components *U*, *S*, and *V* using Singular Value Decomposition (SVD) method. Next, use *U*, *S* and *V* to convert into original matrix *M*.   [4 marks]

4. The details of the students' marks in the last year are given in a file *student_marks.csv* with the following two fields.

Score (out of 100)	Number (of students)
40	10
45	14
....	
100	12

Write a Python program to draw the histogram (score vs. number).
Write a Python program to split the histogram into two equal portions. [4 marks]

5. Differentiate between the following terms.
   - Hard Clustering and Soft Clustering
   - PCA and SVD
   - K-Mean Clustering and Gaussian Mixture Model Clustering [2×3=6 marks]

6. What is the need of data normalization and data transformation in a dataset? Explain any four commonly used data transformation techniques. [5 marks]