Data Mining Lab, ***The Second (continued)***

Note

- **Preferably use Python and external libraries like Pandas.**

The (morphed) data supplied in the CSV represents the mark sheet of <beep> course taken at <beep> university. The roll numbers have been anonymized, just in case they contain data of some current industry leaders. Load the data and write programs to answer the following:

1. Normalize the data using μσ normalization.
2. Find the distance between all pairs of students. Print all details of student-pair that is furthest apart (use normalized values for calculations, however, print unnormalized values).
3. For every student find the average distance to the closest 10 students. Compute the mean (say μ) and standard deviation (say σ) of the distance for all students.
4. For every student find the average distance to the closest 10 students. Count the number of students (as a percentage of class population) for which the distance is greater than μ + 2σ, where μ and σ are obtained from the above question. Print all details of all students qualifying as per the mentioned criterion (use normalized values for calculations, however, print unnormalized values). Delete the students from the dataset (only for this question).
5. For every student find the average distance to the closest 10 students. Count the number of students (as a percentage of class population) for which the distance is greater than μ + *k* σ, where μ and σ are obtained from the above question. For different values of *k*, plot the graph of student count v/s *k*.
6. For every student find the average distance to the closest 10 students. Count the number of students (as a percentage of class population) for which the distance is less than μ - 2σ, where μ and σ are obtained from the above question. Delete all neighboring students appearing in the count WITHOUT deleting the student used to calculate the neighbors (only for this question). Solve repeatedly until the criterion cannot be satisfied for any student. Print all details of all students remaining in the dataset (use normalized values for calculations, however, print unnormalized values).
7. For the students remaining from the above question, plot the theory marks v/s lab marks as a scatter plot. Repeat for all students in the dataset.