Q1. Download the ”QSAR fish bioconcentration factor” from UCI repository and perform the following operations:

PART - I

1. Create a dataset with the content of the file
2. Drop columns titled Name, SMILES and KOW type.
3. Remove the values in column titled “LogKOW” corresponding to the values starting with V-Mey_NA in the column titled “CAS”.
4. Check for missing values. If available, fill it with zeros, ones and mean of column.
5. Remove the column titled “CAS”.
6. Perform linearity analysis on the resultant dataset.
7. Normalize the values using min-max normalization.
8. Construct a regression equation, y=mx+c with LogKOW as independent attribute(x) and logBCF as dependent attribute(y).
9. Manually check whether the results of m and c are correct using excel.
10. Identify MAE,MSE and R2 scores. What are you inferring from the scores?

PART - II

11. Repeat steps from 1 to 7.
12. Divide the dataset into having 750 and 308 rows. Randomization may be applied.
13. Store the data for 750 rows in train_x and train_y lists.
14. Store the data for 308 rows in test_x and actual_y.
15. Predict the value of y, dependent value using the calculated m and c values and store in predicted_y.
16. Compare the difference between actual_y and predicted_y.
17. Calculate MAE, MSE and r2. What do you infer from the scores?

PART - III

18. Perform cross – validation with 2, 3, 4 and till 14 folds with r2 as the metric.
19. Perform thorough analysis on this.
20. Draw appropriate graphs wherever necessary.