Data Set Information:

LC50 data, which is the concentration that causes death in 50% of test fish over a test duration of 96 hours, was used as a model response.

The data comprised 6 molecular descriptors:

MLOGP (molecular properties),

CIC0 (information indices),

GATS1i (2D autocorrelations),

NdssC (atom-type counts),

NdsCH ((atom-type counts),

SM1\_Dz(Z) (2D matrix-based descriptors).

Details can be found in the quoted reference: M. Cassotti, D. Ballabio, R. Todeschini, V. Consonni.

Attribute Information:

6 molecular descriptors and 1 quantitative experimental response:

1) CIC0

2) SM1\_Dz(Z) : 0 means missing value

3) GATS1i

4) NdsCH

5) NdssC

6) MLOGP

7) quantitative response, LC50 [-LOG(mol/L)]

**Objective:**

Prepare the dataset for analysis, using methods learnt and researched independently, including but not limited to Missing Value Imputation, Outlier Detection etc.

Find out your observations on different variables using descriptive statistics, Visualization etc. Report if there is any pattern present in the data.

Take LC50 [-LOG(mol/L)] as the target variable and fit different models for regression using other variables present in the data. Optimize the model parameters and find the best performing model. Compare all models used, using various performance metrics.

Provide inferences to your findings.

Use any other algorithms if necessary to solve any problem the data expects you to.

Look for gaps and problems to solve and report the approach you applied to attempt to solve it.

Submit a **detailed report** on the entire analysis with your findings and insights **as a PDF file** by 19th July ’24.