**Predicting Interfacial Tension with Artificial Intelligence**

**Introduction**

The models compared are going to be an ensemble method called Extreme Gradient Boosting and A neural network with RMSProp or Adam optimizer. The data consists of 2,242 data points with 9 features which are subdivided into 7 continuous(numerical) and 2 discrete (categorical features).

**Data Preparation.**

The data is going to be explored statistically and technically using Python libraries like Pandas for data exploration and Matplotlib for visualization. About 1,118 data points had null values which had to be dropped to preserve the statistical attributes of the data. The data statistics after dropping null values is as follows:

|  | **Temperature** | **Pressure** | **Methane** | **Nitrogen** | **M1\_conc** | **M2\_conc** | **D\_change** | **IFT** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 1606.000000 | 1606.000000 | 1606.000000 | 1606.000000 | 1606.000000 | 1606.000000 | 1606.000000 | 1606.000000 |
| **mean** | 75.275965 | 16.492740 | 3.735990 | 6.741936 | 0.739823 | 0.590028 | 0.653853 | 40.711513 |
| **std** | 39.820374 | 13.050292 | 14.502853 | 18.480021 | 1.263320 | 1.278646 | 0.296067 | 11.978374 |
| **min** | 5.250000 | 0.100000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.012192 | 12.400000 |
| **25%** | 40.000000 | 6.182500 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.388625 | 31.150000 |
| **50%** | 70.000000 | 14.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.731555 | 38.175000 |
| **75%** | 100.000000 | 20.965000 | 0.000000 | 0.000000 | 0.980000 | 0.045000 | 0.906625 | 48.645000 |
| **max** | 196.250000 | 69.510000 | 80.000000 | 76.360000 | 4.950000 | 5.000000 | 1.273900 | 76.100000 |

The correlation of features is also as follows:  


The data was normalized and split randomly, into a section of 70% for training and 30% for testing as the standard practice. The base model was a linear regression model and the average score for testing was 80.7%. The highest test score for the linear regression model is 83.57%, which was achieved on the 5th random state of data splitting. The best mean squared error was achieved on the 5th random state is 23.967. The average mean squared error for the test data is 27.630. The average training score is 81.08%. The highest training score 82.29% which was achieved on the 27th random state of data splitting.