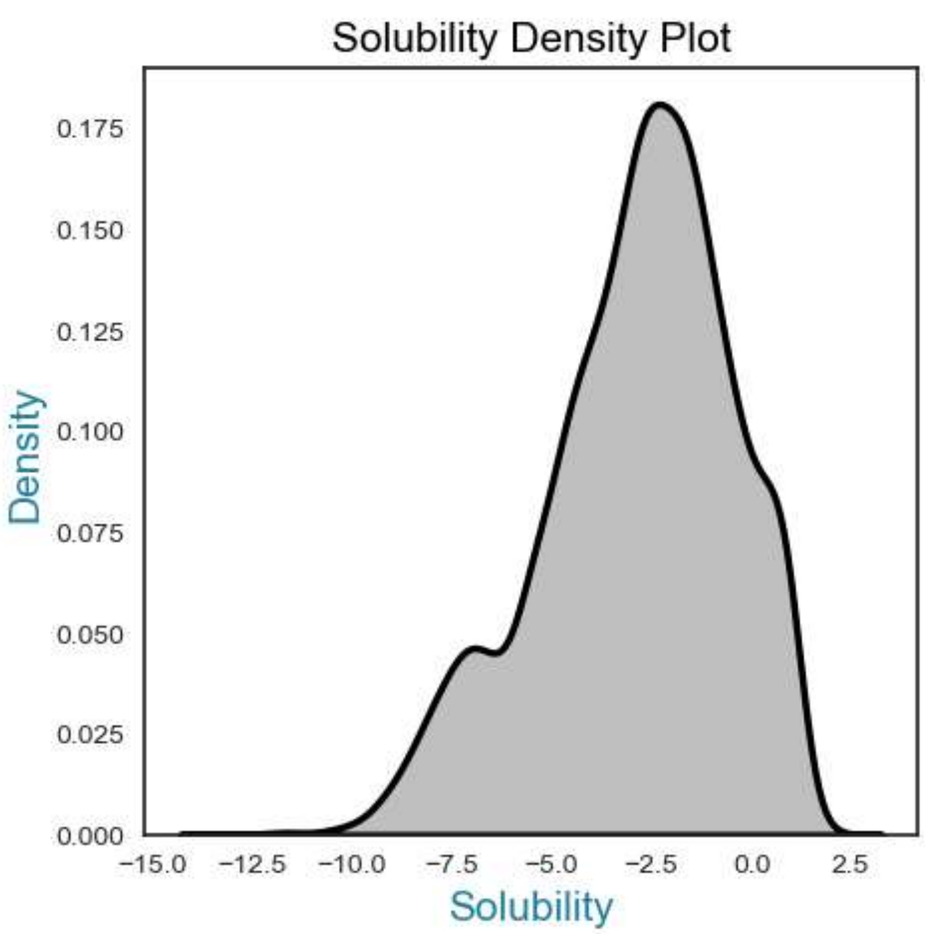
Chemical Industry AI-ML Case Study

The chemical industry is an important sector that produces a wide range of products, including pharmaceuticals, plastics, and fertilizers. Aqueous solubility is a critical property for chemicals used in a wide range of applications, including pharmaceuticals, agrochemicals, and other chemical industries. The ability to predict aqueous solubility accurately is crucial in the chemical industry as it can help to optimize the design of new compounds, improve manufacturing processes, and reduce costs.

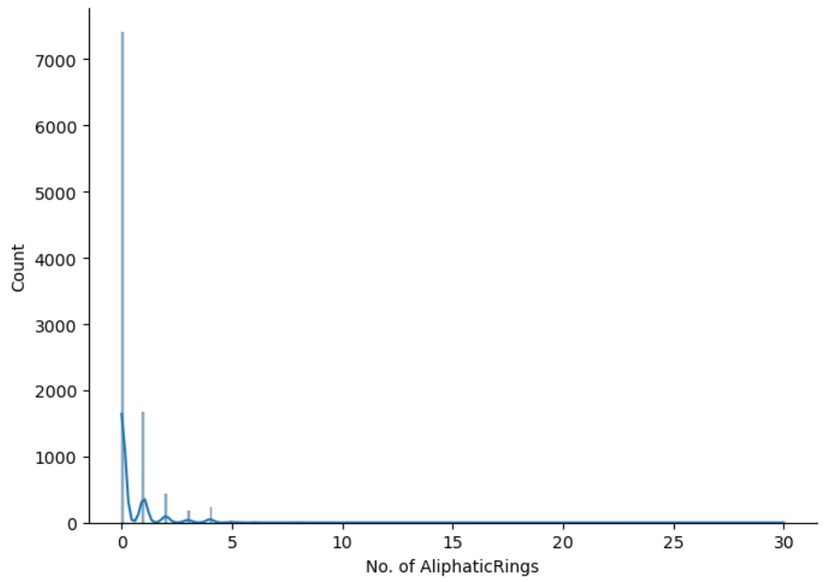
The objective is to develop a Auto-ML model that accurately predicts the solubility of new chemical compounds based on their molecular descriptors. The goal is to reduce the number of experiments required to determine solubility values, reduce costs associated with manufacturing and testing, and accelerate the development of new compounds for a wide range of applications.



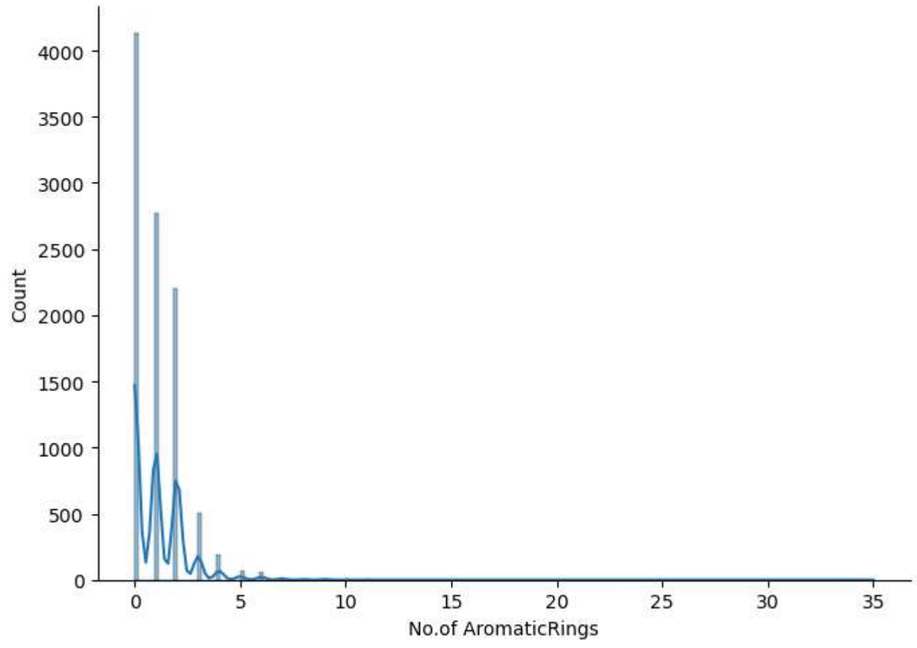
Aqueous solubility constitutes a crucial property of chemical substances that governs behavior of phenomena in several areas like geochemistry, climate predictions, biochemistry, drug-design, agrochemical design, and protein ligand binding.

The substances has limited solubility or is sparingly soluble, meaning that only a small amount of the substance is able to dissolve in the solvent at the given conditions.

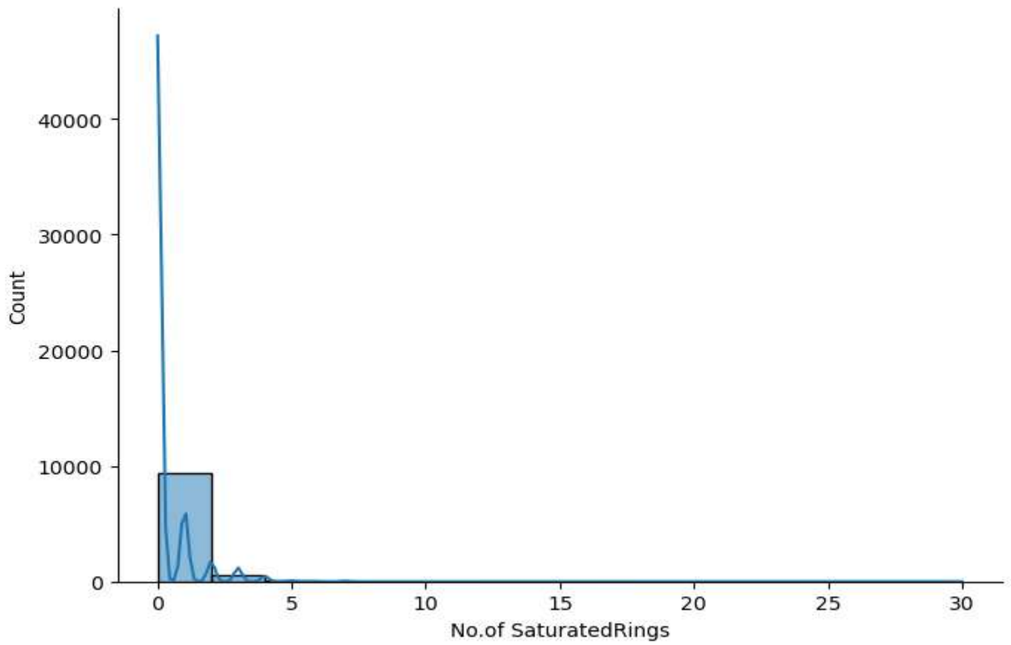
Features Responsible



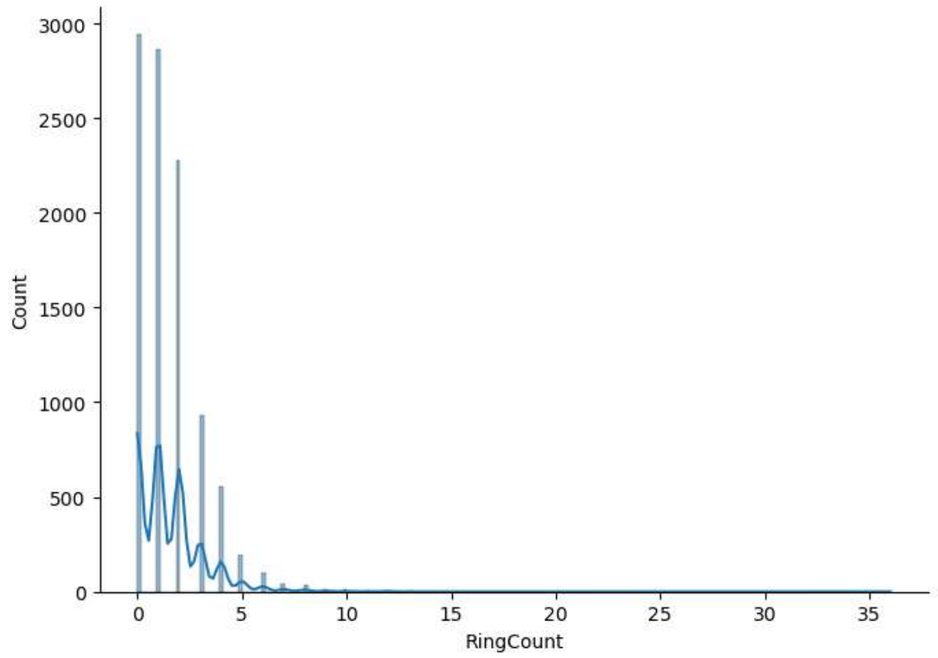
- In general, molecules with more **aliphatic rings** tend to be less soluble in water because they are often hydrophobic or water-repellent.



- In general, molecules with more **aromatic rings** are less soluble in water because of their hydrophobic nature.



- In general, **saturated rings** tend to be less polar than unsaturated rings, which can make them less soluble in polar solvents like water.



- In general, molecules with more **rings** tend to be more hydrophobic, meaning they are less soluble in water, which is a polar solvent.

Auto-ML Methodology Results

Case	Percentile	No. of Features	Random Forest	XGBoost	RNN	MLP	Lasso	Avg. Accuracy
Case 1	25	7	76.0	79.1	51.4	49.1	55.4	**62.2**
Case 2	50	13	81.5	79.9	50.50	32.5	56.1	**60.1**
Case 3	75	19	94.9	97.4	32.9	33.2	56.4	**62.96**
Case 4	90	23	94.8	97.4	50.4	43.1	56.4	**68.42**

- Based on our observation , XGBoost was the best performing algorithm with 97.4% accuracy in 75^{th} and 90^{th} percentile.
- 90th percentile is the best percentile with an average accuracy of 68.42%.

Conclusion

In conclusion, the development of a Auto-ML model for predicting aqueous solubility in the chemical industry is a crucial task that can significantly accelerate the discovery and development of new compounds. With the increasing demand for new and improved chemicals in a wide range of industries, such as pharmaceuticals, agriculture, and materials science, the ability to predict solubility values accurately and efficiently is becoming more important than ever. The dataset has 9982 records with 6 Categorical Features and 20 Numerical Features.

For regression, models were created with algorithms using Auto-ML techniques like Lasso, Recurrent Neural Network, Multilayer Perceptron, Random forest and XGBoost . With these models, performance measurement values were obtained for feature sets of 7, 13, 19 and 23. The Auto-ML algorithms were able to predict solubility with an average accuracy between 60% – 70% and helped to identify factors that determines the solubility. The major factors include No. of Aliphatic Rings, No. of Aromatic Rings, No. of Saturated Rings and Ring Count. The Random forest with 94.9 % accuracy in 75th percentile where tree showed a threshold of Octonal-water partition coefficient<= 1.74 units and topological complexity index<= 288.17 units which leads to highest solubility.

Overall, the application of Auto-ML in predicting aqueous solubility is a critical tool for the chemical industry. By leveraging Auto-ML algorithms to accurately predict solubility, the industry can become more efficient, reduce costs, and accelerate the development of new compounds for a wide range of applications.