ESTIMATING AQUEOUS SOLUBILITY DIRECTLY FROM MOLECULAR STRUCTURE

MINOR RESEARCH PROJECT (CH355)

Submitted by:

Grandhe Sreesai Vaishnavi (186119)

Mareddy Sri Lokesh Reddy (186229)

Soumyajit Das(186246)

Renu Ravindhar (186143)

Supervisor:

Dr. Anand Kishore Kola (Professor)

Department of Chemical Engineering

National Institute of Technology, Warangal

(2020-2021)

BONAFIDE CERTIFICATE

This is to certify that the project report titled "Estimating aqueous solubility directly from molecular structure" by Grandhe Sreesai Vaishnavi (186119), Mareddy Sri Lokesh Reddy (186229), Soumyajit Das (186246) and Renu Ravindhar (186143) is a bonafide record of their own work done under my supervision and guidance.

Dr. Anand Kishore Kola
Professor
Department of Chemical Engineering
National Institute of Technology, Warangal

DECLARATION

I declare that this written submission represents my ideas in my own words and where others ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/ data/ fact/ source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1. Grandhe Sreesai Vaishnavi (186119)
2. Mareddy Sri Lokesh Reddy (186229)
3. Soumyajit Das (186246)
4. Renu Ravindhar (186143)

Date: 10-05-2021

ABSTRACT

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. In this study we explore how machine learning is being used for drug discovery particularly by building regression models in Python for predicting the solubility of molecules (i.e. LogS values). The model was derived from a set of 1144 measured solubilities using regression against four molecular properties. Three regression algorithms were run employing Linear, Ridge and Adaboost techniques. Adaboost regression provided the best results followed by linear and then ridge regression models.

Contents

S.No.	Topic name	page number
1.	INTRODUCTION	1
2.	LITERATURE REVIEW	2-4
3.	DATA PREPARATION AND ANALYSIS	5-10
4.	REGRESSION INTRODUCTION	11-13
5.	LINEAR REGRESSION	14
6.	RIDGE REGRESSION	15-16
7.	ADABOOST REGRESSION	17-18
8.	RESULTS AND COMPARISON	19-20
9.	SCOPE FOR IMPROVEMENT AND CONCLUSION	21
10.	REFERENCES	22
11.	ACKNOWLEDGEMENT	23

List of Figures

S.No.	Figure name	Page Number
1.	Picture of a portion of the data used.	6
2.	Portion of the dataframe version of the excel file.	6
3.	Portion of final dataframe after processing.	7
4.	Correlation Heatmap of the descriptors	8
5.	Measured log (solubility: mol/L) vs Molecular Weight	8
6.	Measured log (solubility: mol/L) vs Number of Rotatable Bonds	9
7.	Measured log (solubility: mol/L) vs MolLogP	9
8.	Measured log (solubility: mol/L) vs Aromatic Proportion	10
9.	Linear Regression	12
10.	Multivariate Linear Regression Scatter Plot	19
11.	Ridge Regression Scatter Plot	20
12.	Adaboost Regression Scatter Plot	20

1. INTRODUCTION

Solubility of a compound is defined as the amount of solute dissolved in a saturated solution under equilibrium conditions. Dissolution is the process of approaching the equilibrium solubility. Solubility is a property of interest to many areas of research, such as pharmaceutical, material, physical, and environmental sciences. It is particularly important to the field of formulation of pharmaceuticals because solubility is relevant to pharmacokinetic properties (absorption, distribution, metabolism, and excretion) and toxicity. For example, a drug must be soluble so that it can be absorbed across the biological membrane to reach the target organ or issue. Solubility of a compound must be accurately determined to assess the concentrations that the drug will achieve in the target area, to establish the therapeutic level, and to prevent toxicity. Lower solubility can hinder the biological activity (for example, absorption and distribution) of a compound, and often a special formulation or modification is required to enhance the solubility. Drug modification can be complex, time-consuming, and sometimes lead to unexpected results.

As aqueous solubility is probably one of the most fundamental property to be studied and deserves attention in the early phases of drug discovery. Therefore, aqueous solubility has been extensively studied, and a large number of computational methods for the estimation of this highly important property have been reported.

Predictive models for aqueous solubility are generally based on a diverse set of descriptors such as experimentally based descriptors, molecular properties, and collection of relevant structural features, that are correlated to activity by means of various statistical techniques including Multiple Linear Regression Model and neural networks.

2. LITERATURE REVIEW

Hansch pioneered the extensive use of CLogP parameter in developing variables for quantitative structure activity relationship (QSAR) equations. Jain and Yalkowsky reported a general solubility equation requiring the water-octanol partition coefficient, logP, and melting point. Huuskonen published a QSPR for solubility prediction based on 30 topological descriptors. The author developed a multilinear regression model and artificial neural network that gave impressive results. Liu and So4 developed a QSPR based on seven 1D and 2D descriptors and an artificial neural network. The preeminent method is probably the "General Solubility Equation" (GSE16), which has just two variables-logP and melting point (Tm). These parameters handle the partition between liquid compound and water (logP) and correct for the transition from solid to liquid (Tm). Octanol partition can be calculated with reasonable accuracy from a compound's structure, but estimating melting point is far harder. Where a measured melting point is available, GSE becomes the method of choice, while other methods, based solely on structure, have to be used in situations where Tm is not available. So here we are trying to compensate for Tm with other properties. There are many parameters that affect solubility like H-bond donor count, H-bond acceptor count, Molecular weight, Aromatic proportion, Non- carbon proportion, Rotatable bonds, clogP, Polar surface area. So, from absolute t-statistic which tells about significance of each parameter and the regression analysis by Delaney four parameters were most significant.

1. Molecular Weight
2. clogP
3. Rotatable bonds
4. Aromatic Proportion

Molecular weight is the sum of the atomic masses of all atoms in a molecule, based on a scale in which the atomic masses of hydrogen, carbon, nitrogen, and oxygen are 1, 12, 14, and 16, respectively. This can be calculated using SMILES.

The logarithm of the 1-octanol/water partition coefficient (log P) is a well-known measure of molecular hydrophobicity and is a very important factor in determining Aqueous Solubility. It is used to assess biological properties relevant to drug action, such as lipid solubility, tissue distribution, receptor binding, cellular uptake, metabolism, and bioavailability. 1-Octanol is a natural choice as the hydrophobic solvent in this respect because of its physicochemical similarity to lipids, its ready availability, and its ease of use. ALOGP and CLOGP are two of the most widely used methods for the estimation of log P. The CLOGP method has improved considerably over the years to cover most neutral organic compounds. It can be calculated using Daylight clogP version 4.72

A rotatable bond is defined as any single non-ring bond, attached to a non-terminal, non-hydrogen atom. Amide C-N bonds are not counted because of their high barrier to rotation. It can be calculated by using an in house program from SMILES.

Aromatic Proportion= No of aromatic atoms in a molecule/ No of heavy atoms in a molecule.

Heavy atom refers to any atom that is not hydrogen. The heavy atoms in proteins are carbon, oxygen, nitrogen, and sulfur. The atoms of an aromatic molecule which satisfies Hückel's rule are called aromatic atoms. No of aromatic atoms and heavy atoms in a molecule can be calculated separately in a molecule. It can be calculated using an in-house program from SMILES. Uses the Daylight SMARTS definition of aromatic ([a]) to count “aromatic atoms”.

SMILES (Simplified Molecular Input Line Entry System) is a chemical notation system designed for modern chemical information processing. Based on principles of molecular graph theory, SMILES allows rigorous structure specification by use of a very small and natural grammar. The SMILES notation system is also well suited for high-speed machine processing. The resulting ease of usage by the chemist and machine compatibility allow many highly efficient chemical computer applications to be designed including

generation of a unique notation, constant-speed (zeroth order) database retrieval, flexible substructure searching, and property prediction models. SMILES is simple to write because rules and hierarchical procedures, which are inherently difficult for the chemist, are relegated to computer algorithms.

3. DATA PREPARATION AND EXPLORATORY ANALYSIS

Here we perform primarily 3 operations:

1. Data Sourcing
2. Data Preparation
3. Exploratory Analysis

Data Sourcing

Data used in this project has been sourced from Open Access Database from ACS publications from the work by John S Delaney, Syngenta, Jealott's Hill International Research Centre, Bracknell, UK[1].

Data Preparation

We receive the data in the form of a Comma Separated File(csv).

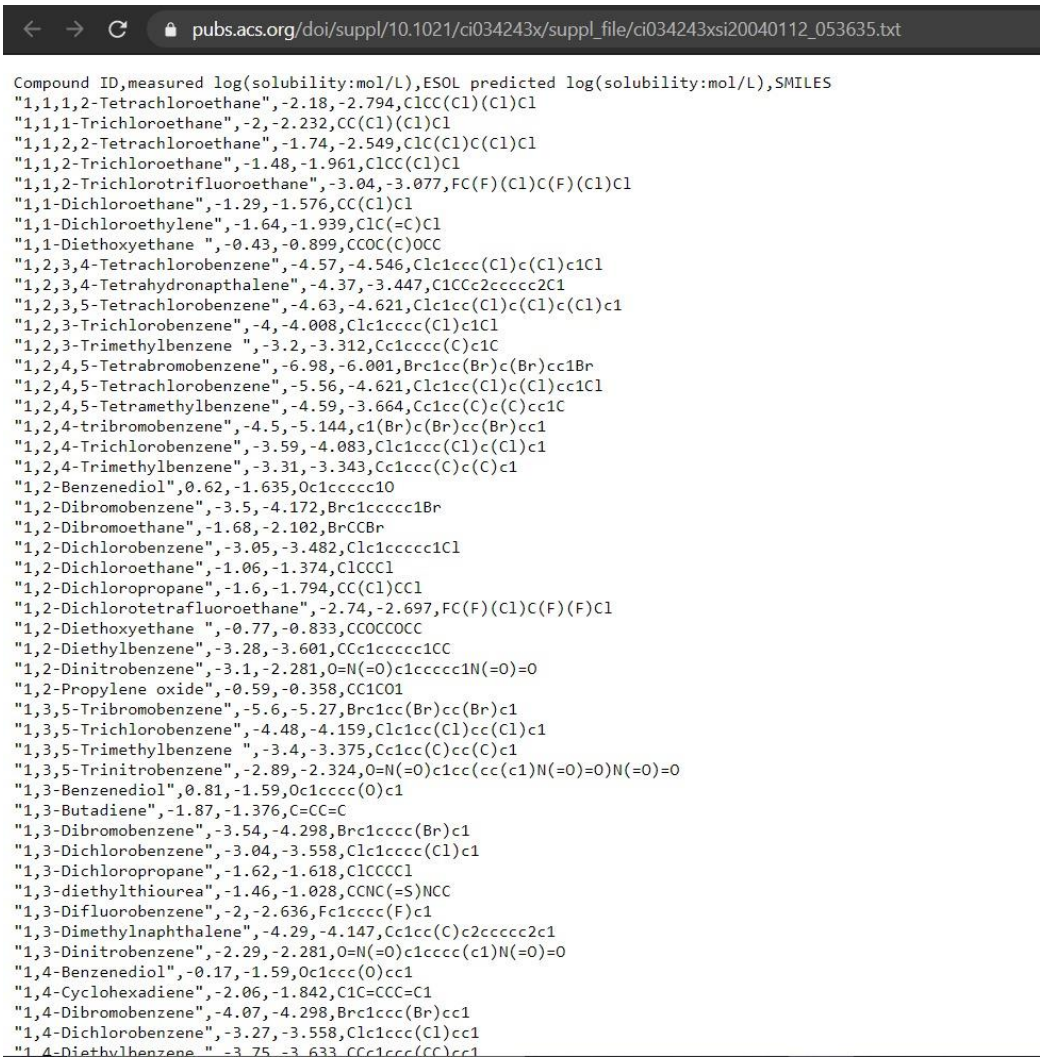

Fig.1. Picture of a portion of the data used.

We convert this data to a more readable form, an excel file.

	Compound ID	measured log(solubility:mol/L)	ESOL predicted log(solubility:mol/L)	SMILES
0	1,1,1,2-Tetrachloroethane	-2.18	-2.794	ClCC(Cl)(Cl)Cl
1	1,1,1-Trichloroethane	-2.00	-2.232	CC(Cl)(Cl)Cl
2	1,1,2,2-Tetrachloroethane	-1.74	-2.549	ClC(Cl)C(Cl)Cl
3	1,1,2-Trichloroethane	-1.48	-1.961	ClCC(Cl)Cl
4	1,1,2-Trichlorotrifluoroethane	-3.04	-3.077	FC(F)(Cl)C(F)(Cl)Cl

Fig.2. Portion of the dataframe version of the excel file.

We remove the empty columns and rows. Chemical structures are encoded by a string of text known as the SMILES notation which is an acronym for

Simplified Molecular-Input Line-Entry System. Then we use the rdkit[2] Python library to decode the SMILES data. We convert the SMILES string to rdkit object, by making use of the for loop to iterate through the list of SMILES strings.

To predict LogS (log of the aqueous solubility), the study by Delaney makes use of 4 molecular descriptors: cLogP (Octanol-water partition coefficient), Molecular weight, Number of rotatable bonds, Aromatic proportion. While rdkit class readily provided the first three parameters, we need to calculate AP by manually computing the ratio of the number of aromatic atoms to the total number of heavy atoms. Then we combine all computed descriptors from the 2 dataframes into 1 dataframe, which gives us the X matrix. Doing this, we get 1144 rows, corresponding to 1144 compounds along with 4 columns containing the descriptors.

	MolLogP	MolWt	NumRotatableBonds	AromaticProportion
0	2.59540	167.850	0.0	0.000000
1	2.37650	133.405	0.0	0.000000
2	2.59380	167.850	1.0	0.000000
3	2.02890	133.405	1.0	0.000000
4	2.91890	187.375	1.0	0.000000
...	...	...	...	...
1139	1.98820	287.343	8.0	0.000000
1140	3.42130	286.114	2.0	0.333333
1141	3.60960	308.333	4.0	0.695652
1142	2.56214	354.815	3.0	0.521739
1143	2.02164	179.219	1.0	0.461538

1144 rows × 4 columns

Fig.3.Portion of final dataframe after processing.

Exploratory Analysis

For the exploratory analysis we plot some graphs to see correlations among certain descriptors:

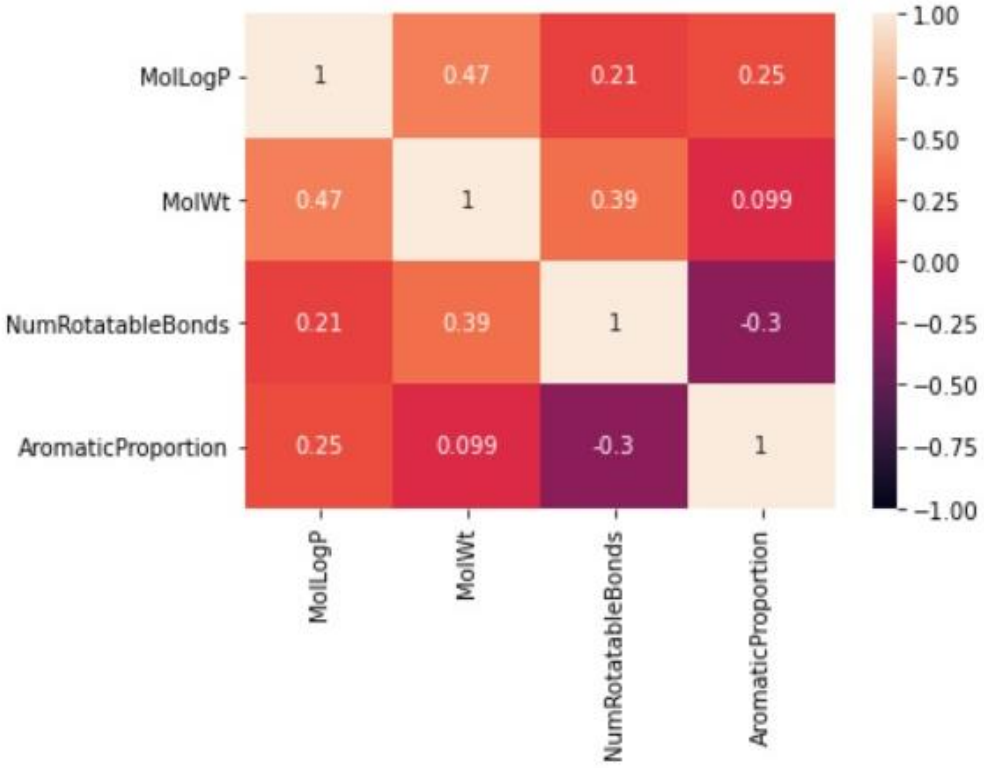

Fig.4.Correlation Heatmap of the descriptors.

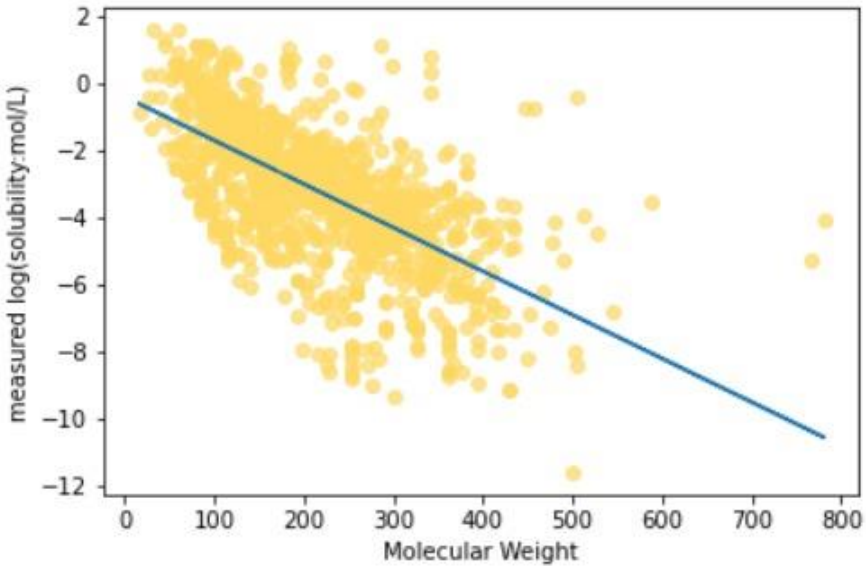

Fig.5. Measured log (solubility: mol/L) vs Molecular Weight

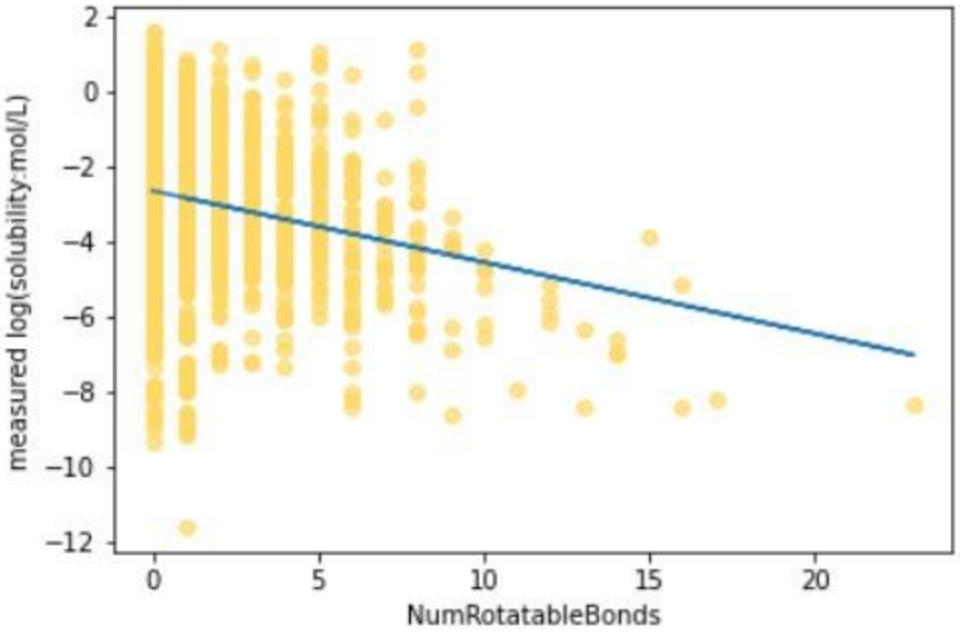

Fig.6.Measured log (solubility: mol/L) vs Number of Rotatable Bonds

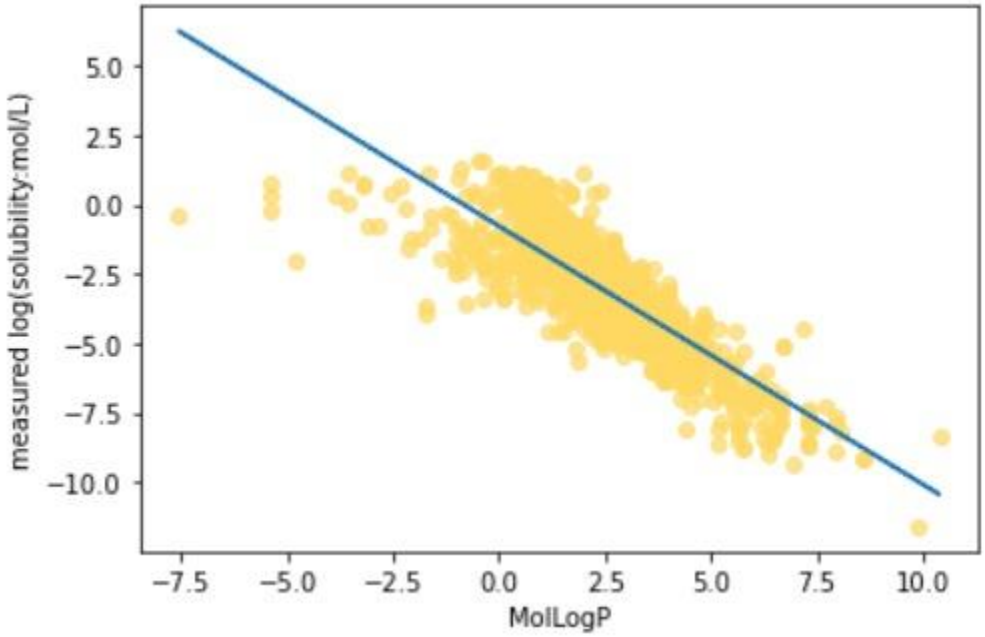

Fig.7.Measured log (solubility: mol/L) vs MolLogP

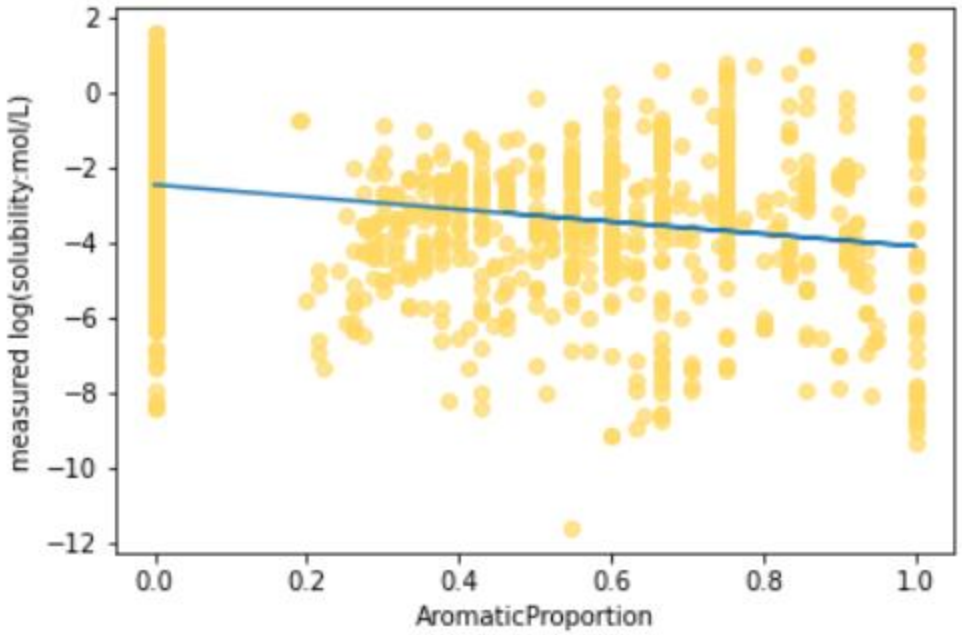

Fig.8.Measured log (solubility: mol/L) vs Aromatic Proportion

4. REGRESSION: INTRODUCTION

Regression is a statistical method used that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables). Regression analysis is primarily used for two conceptually distinct purposes. First, regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Second, in some situations regression analysis can be used to infer causal relationships between the independent and dependent variables. To use regressions for prediction or to infer causal relationships for example, to find a relation between home environment and education, a researcher must carefully justify why existing relationships have predictive power for a new context.
For our study, we use Python language with all the code written on Jupyter IDE. We can simplify our model as:

$$yi= f(p,q,r,s)= w1p+w2q+w3r+w4s+c+\epsilon$$

Where:
yi is the dependent variable.
The variables p,q,r,s represent the attributes, or distinct pieces of information, we have about each observation. For our study, p,q,r,s are Octanol-water partition coefficient , Molecular weight, Number of rotatable bonds, and aromatic proportion.
wi represent the respective coefficients.
c is the y-intercept.
ϵ is the error term.

Two parameters we use in this study to measure the working of our models are stated below:

Mean Squared Error(MSE)

Let us take a simple function $y_i = f(x)$ as shown below

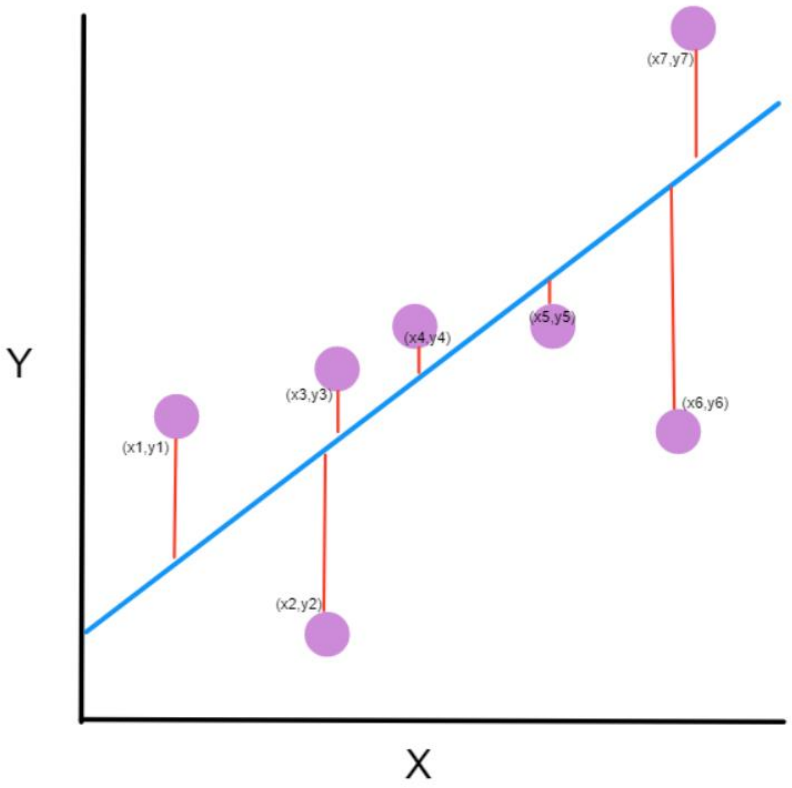

Fig.9. Linear Regression

Here, the purple dots are the points on the graph. Each point has an x-coordinate and a y-coordinate.

The blue line is our prediction line. This is a line that passes through all the points and fits them in the best way. This line contains the predicted points.

The red line between each purple point and the prediction line are the errors. Each error is the distance from the point to its predicted point.

$$MSE = (1/n)\sum(y_i - y'_i)^2$$

For each point, we take the y-coordinate of the point, and the y'-coordinate. The y-coordinate is our purple dot. The y' point sits on the line we created. We subtract the y-coordinate value from the y'-coordinate value, and calculate the square of the result. Then we take the sum of all the $(y-y')^2$ values, and divide it by n, which will give the mean.
Therefore, for our experiment

$$MSE= (1/n)\sum(y_i-(w_1p+w_2q+w_3r+w_4s))^2$$

Coefficient of determination (R^2)

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared ranges from 0 to 1, where 0 represents a model that does not explain any of the variation in the response variable around its mean and 1 represents a model that explains all the variation in the response variable around its mean.
Again, let's consider a dataset with n values with y_i as the observed data and y' as the predicted data.
If y_m is the mean of the observed data:

$$y_m=(1/n)\sum y_i$$

The total sum of squares (proportional to the variance of the data):

$$SS_T= \sum(y_i-y_m)^2$$

Now if we consider residuals as the residuals as $e_i = y_i - y'_i$
The sum of squares of residuals, also called the residual sum of squares:

$$SS_R= \sum e_I^2$$

Thus, we get

$$R^2=1-(SS_R/SS_T)$$

5. LINEAR REGRESSION

Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range.

For our model, we make a train-test split. The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset. The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.

After training the model we test it on the out-of-sample test set and observe the R^2 and MSE values. The values we get are:

R^2	MSE
0.75	1.13

To inspect further we try and derive the coefficients (w_i) for the Linear Regression Equation from the test data as:

logS = 0.235096 -0.76774926 logP -0.00640718 MW +0.01230929 RB -0.38766774 AP

6. RIDGE REGRESSION

Ridge Regression is a popular type of regularized linear regression that includes an L2 penalty. This has the effect of shrinking the coefficients for those input variables that do not contribute much to the prediction task. L2 penalty minimizes the size of all coefficients, although it prevents any coefficients from being removed from the model by allowing their value to become zero. Here too, we used a train-test split as before. Ridge Regression consists of various hyperparameters that determine the computational work that the model will follow. They are described in the table below.

Hyperparameter	**Function**	**Value used**
alpha	Regularization strength. Regularization improves the conditioning of the problem and reduces the variance of the estimates	0.99
fit_intercept	Whether to fit the intercept for this model.	True
normalize	If True, the regressors X will be normalized before regression by subtracting the mean and dividing	False
copy_X	X will be copied	True
max_iter	Maximum number of iterations for conjugate gradient solver	None
tol	Precision of the solution.	1e-3
solver	Solver to use in the computational routines	auto

After training the model we test it on the out-of-sample test set and observe the R^2 and MSE values. The values we get are:

R^2	**MSE**
0.74	1.17

Thus, here we do not find any improvement. In fact, we find a marginal reduction in R^2 and an increase in MSE values, both of which are undesirable.

However, we proceed to find the coefficients and get the Regression Equation for our model:

logS= 0.234446 -0.76768984 logP -0.00640935 MW + 0.01240782 RB -0.38568153 AP

7. ADABOOST REGRESSION

Boosting refers to a class of machine learning ensemble algorithms where models are added sequentially and later models in the sequence correct the predictions made by earlier models in the sequence.

Adaboost, expanded to Adaptive Boosting, is called so as it responds adaptively to the errors of the weak hypotheses[4]. The weak learners in AdaBoost are decision trees with a single split, called decision stumps.Adaboosting works by putting more weight on difficult to classify instances and less on those already handled well.The training algorithm involves starting with one decision tree, finding those examples in the training dataset that were misclassified, and adding more weight to those examples. Another tree is trained on the same data, although now weighted by the misclassification errors. This process is repeated until a desired number of trees are added.[5]

Here too we have some hyperparameters to dictate computation of the adaboost model.

Hyperparameter	**Function**	**Value**
base_estimator	The base estimator from which the boosted ensemble is built. If None, then the base estimator is DecisionTreeRegressor	None
n_estimators	The maximum number of estimators at which boosting is terminated.	50
learning_rate	Weight applied to each classifier at each boosting iteration. A higher learning rate increases the contribution of each classifier. There is a trade-off between the learning_rate and n_estimators parameters.	1
loss	The loss function to use when updating the weights after each boosting iteration.	‘linear’

For the model, here we use Hyperparameter Optimisation to determine the optimal values we need for a better R^2 score. We use k-fold cross-validation, where we split the input data into k(here, 10) subsets of data (also known as folds). Then we train an ML model on all but one (k-1) of the subsets, and then evaluate the model on the subset that was not used for training. This process is repeated k times, with a different subset reserved for evaluation (and excluded from training) each time.
We use the same train-test split to implement our Adaboost model, and we get the following results:

R^2	**MSE**
0.79	0.95

We get significantly better results with the Adaboosting model with improvements in both out-of-sample R^2 and MSE.

8. RESULTS AND COMPARISON

After implementing Linear, Ridge and Adaboost Regressions, we can compare the $R^{2.}$ and MSE values as shown.

Regression Type	R^2	MSE
Multivariate Linear	0.75	1.13
Ridge	0.74	1.17
Adaboosting	0.79	0.95

Thus, we see significant improvements in Adaboosting when compared to the other models. To better understand the comparison, we can visualize the correlation of the Experimental LogS values with those of the Predicted LogS values in the training sets by means of the scatter plots for each Regression Model

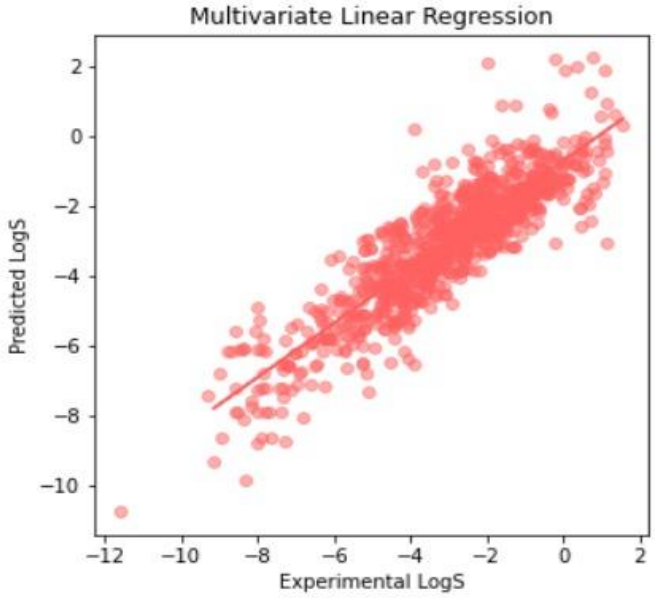

Fig.10.Multivariate Linear Regression Scatter Plot

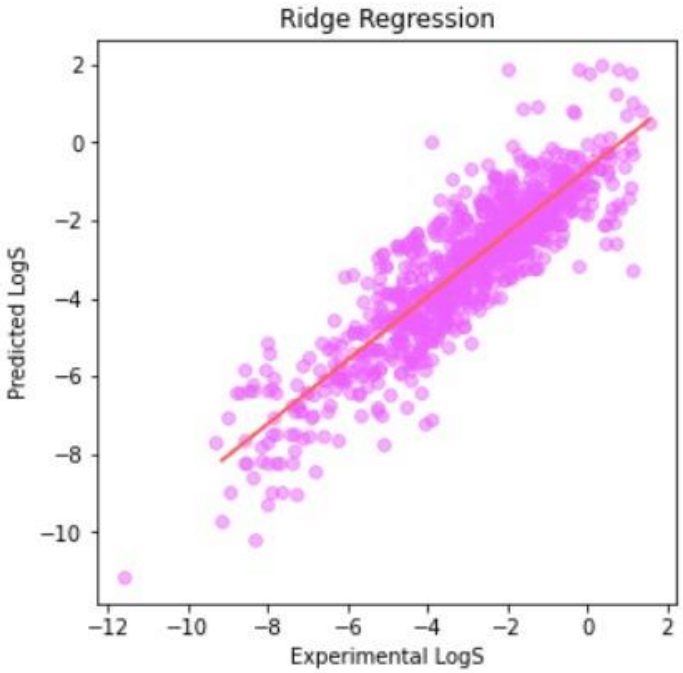

Fig.11.Ridge Regression Scatter Plot

Even visually, the Adaboost model is significantly better, with all the plot points much closer to the regression line than in other models.

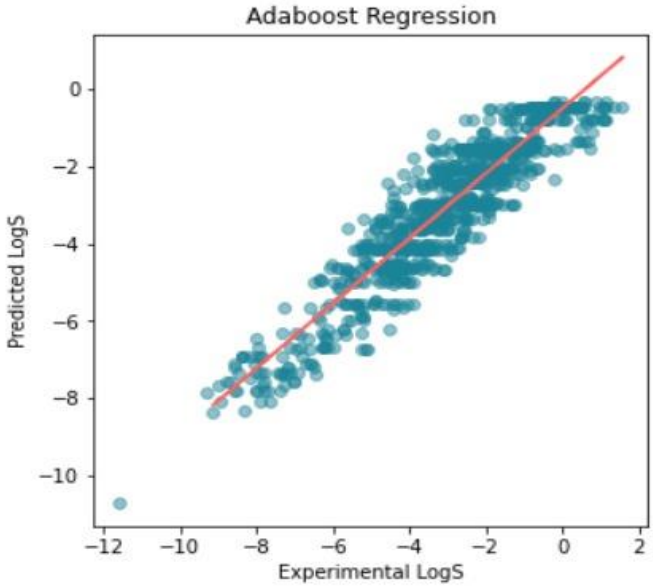

Fig.12.Adaboost Regression Scatter Plot

9. SCOPE FOR IMPROVEMENT AND CONCLUSION

- While we receive satisfactory results from the Adaboost model, we can look to further tune the hyperparameters and in fact look to develop models based on other Machine Learning algorithms like XG Boosting.

- Further research may also be done into looking at more molecular properties that may influence the solubility and incorporate them in models.

- All models must be developed with the importance of accuracy in mind especially when we consider it's importance in crucial industries like drug-delivery.

- In conclusion Machine Learning can be implemented to create robust methods for estimating the solubility of compounds without having to resort to physical measurements.

10. References

[1]Cheng, A.; Merz, K.M. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure-Property Relationships. J. Med. Chem. 2003, 46, 3572-3580.

[2]Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. J. Chem.Inf. Comput. Sci. 2003, 43, 837-841.

[3]Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of LOGP and CLOGP Methods. J. Phys. Chem. 1998, 102, 3762-3772.

[4]John S. Delaney ESOL: Estimating Aqueous Solubility Directly from Molecular Structure J. Chem. Inf. Comput. Sci. 2004, 44, 1000-1005

[5] DAVID WEININGER SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules J. Chem. In$ Comput. Sci., Vol. 28, No. I , I988

[6]DAVID WEININGER, ARTHUR WEININGER, and JOSEPH L. WEININGER SMILES. 2. Algorithm for Generation of Unique SMILES Notation J. Chem. Inf. Comput. Sci., Vol. 29, No. 2, 1989

[7]https://www.rdkit.org/docs/Overview.html

[8]Gibbs Y. Kanyongo, Janine Certo, Brown I. Launcelot, Using regression analysis to establish the e environment and reading achievement: A case of Zimbabwe, International Education Journal, 2006, 7(5), 632-641

[9]Yoav Freund, Robert E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, In: Vitányi P. (eds) Computational Learning Theory. EuroCOLT 1995. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 904. Springer, Berlin, Heidelberg.

[10]Trevor Hastie (Department of Statistics, Stanford University, Stanford, Calif., U.S.A.), Saharon Rosset (Department of Statistics, Tel Aviv University, Tel Aviv, Israel), Ji Zhu (Department of Statistics, University of Michigan, Ann Arbor, Mich., U.S.A.), Hui Zou (School of Statistics, University of Minnesota, Minneapolis, Minn., U.S.A.), Multi-class AdaBoost, Statistics and Its Interface Volume 2 (2009) Number 3, 349-360

11. ACKNOWLEDGEMENT

First of all, we would like to express our deep and sincere gratitude to our project guide, Dr Anand Kishore Kola, for being our supervisor for this minor research project. Thank you, sir, for your guidance and support which helped us understand about a research project and for completion of our project successfully. We are very grateful to have the chance to learn from you to be rigorous, serious and reliable for research.

We would also like to give our special thanks to Dr Kishant Kumar. We are very grateful for your general support and cooperation.

Deep appreciation to the Dept. of Chemical Engineering, NITW for providing us with this opportunity in this semester.

Last but not the least, our heartfelt gratitude is extended to our parents and friends for their ongoing understanding and unconditional support throughout the semester for this project.