WiDS PROJECT

PROJECT UID 7

Prediction of Solubility of Chemical Compounds using Machine Learning techniques

By
Soumya Saha
Roll Number 23B2721

CONTENTS

ABSTRACT

Solubility, the phenomenon of dissolution of solute in solvent to give a homogenous system, is one of the important parameters to achieve desired concentration of drug in systemic circulation for desired (anticipated) pharmacological response. Low aqueous solubility is the major problem encountered with formulation development of new chemical entities as well as for the generic development. More than 40% NCEs (new chemical entities) developed in pharmaceutical industry are practically insoluble in water. Solubility is a major challenge for formulation scientist. Any drug to be absorbed must be present in the form of solution at the site of absorption. Various techniques are used for the enhancement of the solubility of poorly soluble drugs which include physical and chemical modifications of drug and other methods like particle size reduction, crystal engineering, salt formation, solid dispersion, use of surfactant, complexation, and so forth. Selection of solubility improving method depends on drug property, site of absorption, and required dosage form characteristics. So, knowing the properties of compounds and predicting the solubility of the compounds using machine learning is very necessary.

That is why, I have made a machine learning project to predict the solubility of certain compounds using certain molecular descriptors. Here , I have used two different machine learning models to see which one is better for my project. It will also provide some values, Mean Absolute Error, Mean Squared Error, R2 score, Root Mean Squared Error, Accuracy Percentage which will tell how accurate the model is. Finally, it will plot a graph between the predicted solubility and actual solubility.

ABOUT PYTHON AND MACHINE LEARNING

Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions. Recently, generative artificial neural networks have been able to surpass many previous approaches in performance.

Machine learning approaches have been applied to many fields including large language models, computer vision, speech recognition, email filtering, agriculture, and medicine, where it is too costly to develop algorithms to perform the needed tasks.

One important thing to note is that machine learning and statistics are very similar in terms of methods but very different in their principal goal; statistics draws population inferences from a sample, while machine learning finds generalizable predictive patterns.

For this, Python is the most suitable and widely recognizable programming language having all the required libraries like NumPy, Pandas, Matplotlib, Scikitlearn, etc.

I have done this whole project using Python in Google Colab.

ABOUT THE DATASET AND DESCRIPTORS

The dataset which I have used in making the project is an organised collection of many properties of the chemical compounds. It contains data of the following properties:

- NAME OF THE COMPOUNDS
- THEIR SMILES FORMAT
- MOLECULAR WEIGHT
- Log P VALUE
- TOPOLOGICAL POLAR SURFACE AREA(TPSA)
- NUMBER OF HYDROGEN ACCEPTORS
- NUMBER OF HYDROGEN DONORS
- NUMBER OF ROTATABLE BONDS
- NUMBER OF VALENCE ELECTRONS
- NUMBER OF AROMATIC RINGS
- Log OF SOLUBILITY

In this project, my aim is to train the machine learning model to predict the solubility of compounds by feeding the model their experimental solubility values.

A key point to be noted here is that the chemical compounds are represented in SMILE (Simplified Molecular Input Line Entry System) string format in Python. By importing RDKIT module from Python, I have used MOLECULAR WEIGHT, Log P VALUE, TPSA, NUMBER OF HYDROGEN ACCEPTORS, NUMBER OF HYDROGEN DONORS, NUMBER OF AROMATIC RINGS, NUMBER OF VALENCE ELECTRONS as molecular descriptors to train the model.

The Models

In this project, I have used two machine learning models and compared them to see which model suits best:

- **RANDOM FOREST REGRESSION**
- **LINEAR REGRESSION**

Random Forest Regression is an ensemble learning algorithm based on the principle of constructing a multitude of decision trees during training and outputting the average prediction of the individual trees for regression tasks. It is a versatile and powerful algorithm known for its high accuracy and robustness.

Linear Regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the input features and the output variable. The goal of linear regression is to find the best-fit line that minimizes the difference between the predicted and actual values of the dependent variable.

SECTIONS OF THE CODE

I have written the whole code for this project in Google Colab and uploaded it in my Github repository “wids_project-repo”, the link of which I have given towards the end. So, the major sections of the code are:

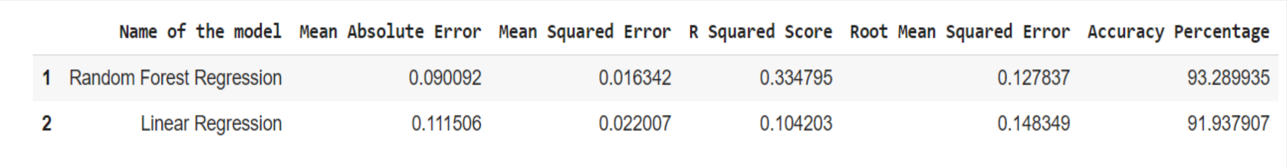
- **Importing of various libraries and modules:** Python provides us the opportunity to use many libraries we want. I have also used some of them to create my project like NumPy, Pandas, Matplotlib, SciKitLearn, RDKit etc.

  I have also imported several modules and classes from these libraries. I have written some lines of code to ignore all the warnings and to not show them in the output console. From sklearn.linear_model I have imported LinearRegression model and from sklearn.ensemble I imported RandomForestRegressor model.
- **Reading of the csv file:** Next step is reading the csv file of the Solubilty dataset using Pandas and displaying it as output.
- **Pre-processing of the dataset:** Using Min Max Scalar, I normalize the dataset to ensure all the values and data contribute equally to the model.
- **Adding Molecular Descriptors:** By using RDKit module, I added some molecular descriptors to build a predictive model that relates the molecular characteristics of a compound to its solubility.
- **Splitting the dataset:** Before training of the model, we have to split the whole dataset into two parts: train dataset and test dataset. I am keeping 20 percent of the dataset as test dataset and rest 80 percent as train dataset. It is important to keep the test data size large so that the model will be more accurate.
- **Training the model:** At first, we will train the model using Random Forest Regressor on the train dataset. Then we will make the model predict the values of log of Solubility on the test dataset. After this, we will do the same thing using Linear Regression model.
- **Evaluating the model:** Now it’s time for checking how accurate the models are. For this, we measure different parameters like Mean Absolute Error, Mean Squared Error, R squared value, Root Mean Squared Error, Accuracy percentage.
- **Data Visualization:** Using Matplotlib we can see how the predicted solubility and actual solubility is varying. I have used many plots to understand the varying of the data in a better way like Scatter Plot, Histogram, Box Plot, Pie charts.
- **Comparison between the models:** Lastly, by creating a dataframe, I compare between the two models, Random Forest Regressor and Linear Regression to see which one is more accurate.

OBSERVATION

So, here is the comparison between the two models:

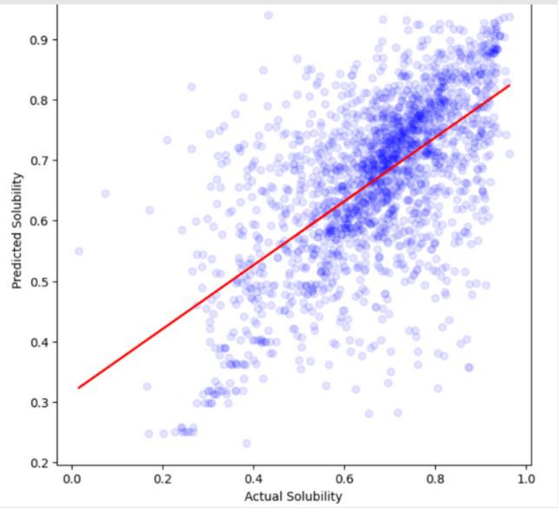
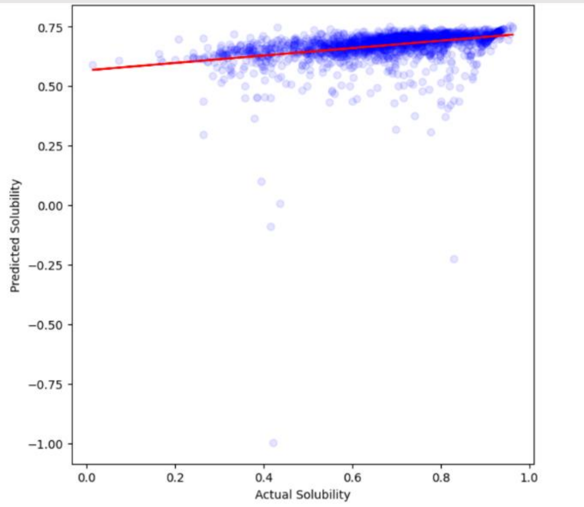
METRICS:



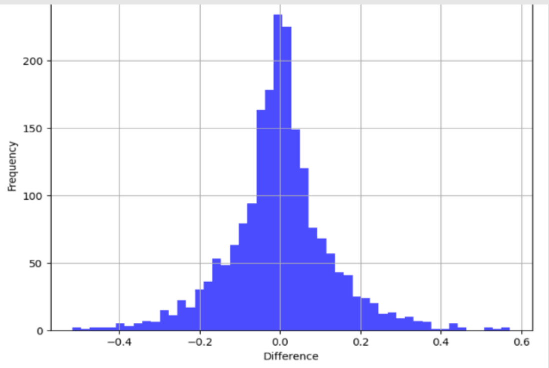
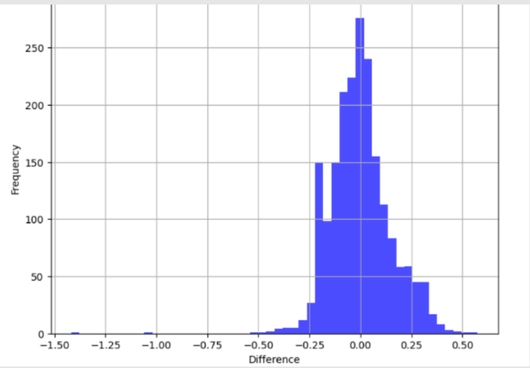
	Name of the model	Mean Absolute Error	Mean Squared Error	R Squared Score	Root Mean Squared Error	Accuracy Percentage
1	Random Forest Regression	0.090092	0.016342	0.334795	0.127837	93.289935
2	Linear Regression	0.111506	0.022007	0.104203	0.148349	91.937907

THE PLOTS:

Scatter Plots

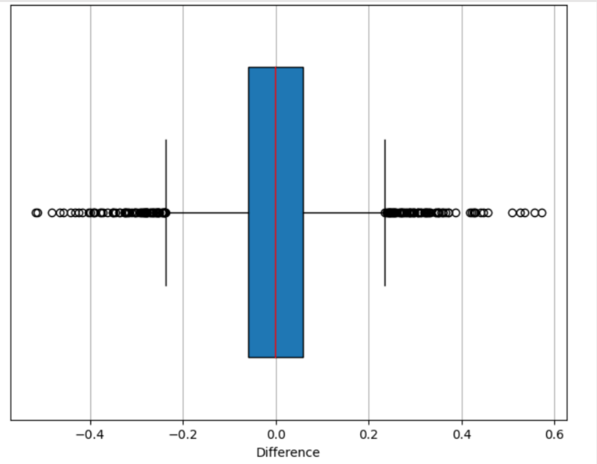
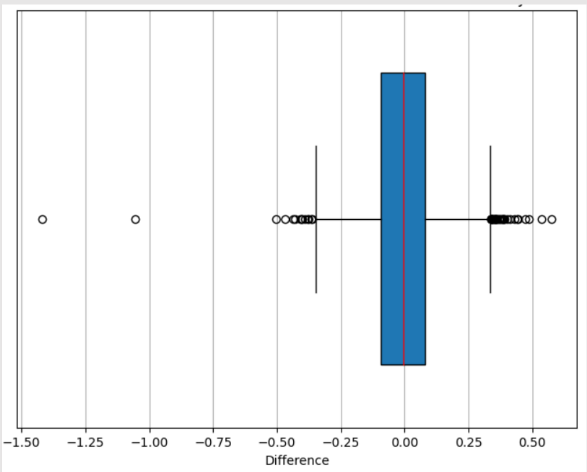
Random Forest Regressor	Linear Regression
	

Histogram

Random Forest Regressor	Linear Regression
	

***difference = Predicted values – Actual values**

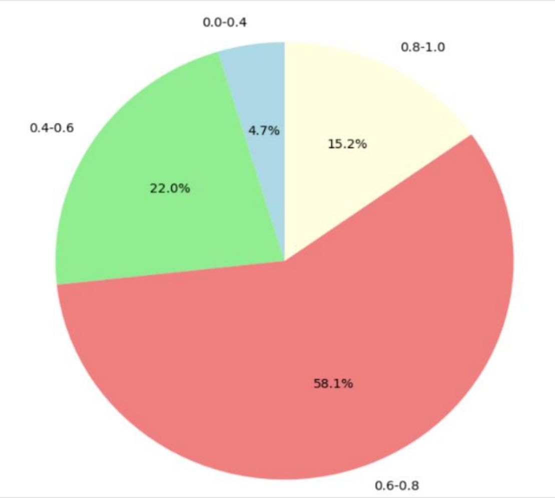
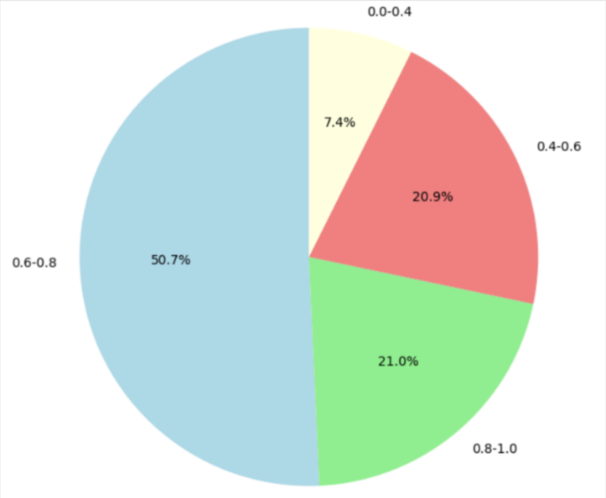
Box Plots

Random Forest Regressor	Linear Regression
	

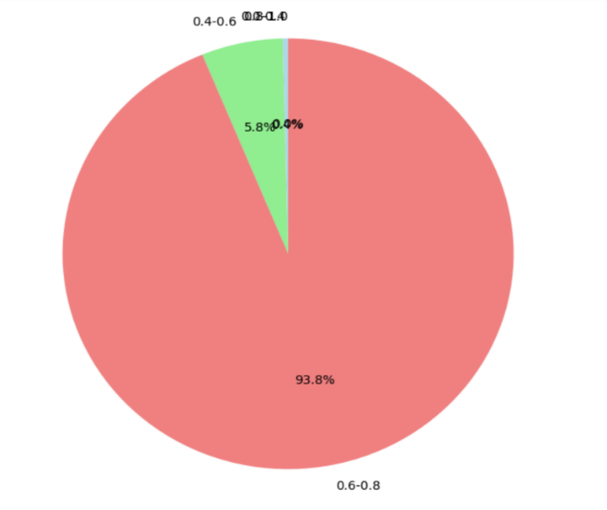
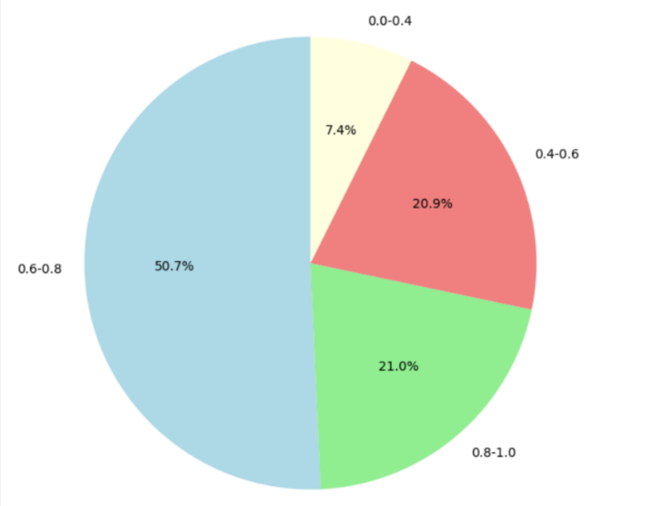
***difference = Predicted values – Actual values**

Pie Charts

Random Forest Regressor:

Predicted Solubility data	Actual Solubility data
	

Linear Regression:

Predicted Solubility Data	Actual Solubility Data
	

INTERPRETATION

Before interpreting, we should know what the metrics mean.

Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted and actual values.

Mean Squared Error (MSE): MSE measures the average squared difference between predicted and actual values.

R2 Score: The R2 score (coefficient of determination) represents the proportion of the variance in the dependent variable (solubility) that is predictable from the independent variable (predicted solubility).

Root Mean Squared Error: It's nothing but the square root of MSE.

Accuracy Percentage: I want to consider predictions within ±0.25 of the actual values as accurate. So, I set the tolerance value as 0.25 and measure how accurate the model is in percentage.

If we look at the metric, we will see that the MAE and MSE values for Random Forest Regressor are lesser than Linear Regression suggesting that Random Forest Regression made more accurate predictions. Also, the R2 score is more in case of Random Forest Regressor. Overall, the accuracy percentage of Random Forest Regressor is almost 93.3 percent whereas it is 92 percent for Linear Regression.

Also, the plots for the Random Forest Regressor are more appealing than Linear Regression.

So, we can conclude that Random Forest Regressor has done more accurate solubility predictions than Linear Regression suggesting **Random Forest Regressor is more suited for my project.**

While the model is providing reasonable predictions, there might still be room for improvement, depending on the specific requirements of application. I will consider further optimization, feature engineering, or trying different algorithms to enhance the model's performance.

GITHUB LINK

For this project, I opened a GITHUB account of mine. I have uploaded all the codes and dataset in the repository “wids_project-repo”.

I am providing the link here:

https://github.com/soumyasahaiitb/wids_project-repo

REFERENCES

In making this project, I have taken most of the information from ChatGPT regarding the coding part.

Besides this, I have also learned about machine learning from certain Youtube videos:

https://youtu.be/29ZQ3TDGgRQ?si=1stU1TNWffYfi56P

https://youtu.be/7eh4d6sabA0?si=41wiPDwYcIbUdS5u

https://youtu.be/kqtD5dpn9C8?si=f43vN6OJr55HA50o

https://youtu.be/8Y0qQEh7dJg?si=zuhJkXgGvQD8F3nm

https://youtu.be/wB9C0Mz9gSo?si=sEj9N-XEZKyZf2vx

https://youtu.be/PcvsOaixUh8?si=cqXupGuisYTcbI5-

The dataset I have taken from this website:

https://www.kaggle.com/datasets/sorkun/aqsoldb-a-curated-aqueous-solubility-dataset?resource=download

And how can I forget about the resources provided by my mentors!!!

https://github.com/Iris-Agape/WiDS_23

ACKNOWLEDGEMENT

I want to extend my gratitude to Samuel Raju Bethala and Vanshika Singh Sindhu who mentored me on this project. I would like to thank for their willingness to share their valuable inputs throughout the project. Their expertise and contributions significantly enhanced the quality of this project.

Finally, I would like to thank the Analytics Club of IIT Bombay who provided me the opportunity to do this project.