

Application of Machine Learning Methods in Predicting Chemical Solubility

Shengye Shi

Luwan Senior High School, Shanghai, China

Kimishishengye@outlook.com

Abstract. Solubility is an important property in chemistry, especially in drug development, as it affects how well a drug can work in the human body. Traditionally, scientists predicted solubility by studying the structure of molecules, but these methods are not always accurate. In recent years, machine learning has become popular in many fields, and this study explores how it can help improve solubility prediction. A dataset from Kaggle with four features including MolLogP, MolWt, NumRotatableBonds, and AromaticProportion was used. The study tested three models: Random Forest, Decision Tree, and Linear Regression. The data was split into training and test sets. Among the models, Random Forest performed best, with an accuracy of 0.90 and AUC of 0.89. It also showed that MolLogP was the most important factor affecting solubility. Visual tools like heatmaps and boxplots helped explain the results clearly. This research shows that machine learning is a useful tool in chemistry and can support better drug design in the future.

Keywords: Solubility; machine learning; artificial intelligence.

1. Introduction

Solubility refers to the ability of a solid, liquid, or gaseous chemical substance to dissolve in solvent and form a solution [1, 2]. It plays a crucial role in chemistry research, influencing thousands of kinds of chemical solutions. Especially in biopharmaceuticals, the solubility of the drug can strongly affect the effect of the drug, which also determines whether it can save people's lives. To harness the importance of solubility, chemists have developed a lot of methods to predict solubility. The general method is to predict solubility by analyzing the molecular structure. However, these traditional methods can sometimes be inaccurate. Because of this, machine learning algorithms can be considered into chemical research to predict solubility more effectively and accurately due to their excellent performance in learning high-level complicated correlation between features and target variable.

Artificial intelligence was first introduced in 1956 and has been widely applied across various areas [3-5]. In recreation areas, it can be used as chat bots based on Large Language Models (LLMs). In statics areas, it can be used to analyze whole branches of data. In addition, there are more exciting implements still under development, like robots and automatic drive. In computer science, Artificial Intelligence (AI) is often associated with machine learning. With representative models including decision tree Classifier, random forest classifier and logistic regression. Machine learning algorithms based on the implementation of Python has been used in multiple areas such as medical treatment and statistics [6-8]. However, there is limited research on its use in predicting chemical solubility. Because of this, this research decides to explore the application of machine learning in predicting solubility.

In order to address the problem, this research collected and utilized a dataset from Kaggle, which contains five variables: LogS, MolLogP, MolWt, NumRotatableBonds, and AromaticProportion. Three machine learning models were employed in this study: Random forest classifier, decision tree classifier, and linear regression. The dataset was first preprocessed to ensure data quality and consistency. Then, the data was divided into training and test sets to evaluate model performance effectively. Each model was trained on the training set to learn patterns between molecular features and solubility. The trained models were then used to predict solubility values on the test set.

2. Method

2.1. Dataset Preparation

This study used the dataset from Kaggle [9]. The dataset is composed of five variables: "MolLogP", "MolWt", "NumRotatableBonds", "AromaticProportion", and "logS". More detailed information about them is as follows: 1) logS: solubility, expressed in molar concentration logarithm. 2) MolLogP: The lipid-water partition coefficient of a compound, reflecting the lipophilicity of the molecule. 3) MolWt: Molecular weight (g/mol), related to molecular volume. NumRotatableBonds: The number of rotatable bonds, indicating molecular flexibility. 4) AromaticProportion: atomic proportion of aromatic ring, affecting molecular polarity.

The research mainly focusses on the relationship between "logS" and "MollogP" or the relationship between "logS" and "MolWt". The author used the median value of logs (-3.5) was used as threshold and, the solubility is divided into two categories: high solubility and Low solubility. This resolves the potential class imbalances to ensure model fairness.

In addition. This study used check missing values and found that. The output is data integrity is good (i.e. no any missing value). The study selected all 4 descriptors (MolLogP, Molwt, NumRotatableBonds, AromaticProportion) as the input for feature, and binary classification label (Solubility Class) as output. The training set (915 samples) and test set (229 samples) were randomly divided by 8:2 ratio, and the fixed random seed (randomstate=42) ensured that the results were reproducible.

2.1.1. Random forest classifier.

Random Forest is an ensemble method that combines multiple decision trees, reducing overfitting through majority voting or averaging [10]. It handles high-dimensional data well, provides feature importance scores, and works effectively with missing values. The model's inherent randomness (bagging and feature subset selection) enhances generalization. However, decision trees are prone to overfitting, particularly when they grow deep, which can lead to poor generalization on unseen data. They show high variance, as small changes in the training data can result in entirely different tree structures. Moreover, decision trees can struggle with continuous variables and may produce biased results when dealing with imbalanced class distributions. Regularization techniques such as pruning are often required to improve performance

2.1.2. Decision tree classifier.

Decision Trees are intuitive, visually interpretable models that mimic human decision-making. They handle nonlinear relationships and categorical features natively, require minimal data preprocessing, and perform well on small datasets. Feature selection is automatic during splits. However, decision trees are prone to overfitting, particularly when they grow deep, which can lead to poor generalization on unseen data. They show high variance, as small changes in the training data can result in entirely different tree structures. Moreover, decision trees can struggle with continuous variables and may produce biased results when dealing with imbalanced class distributions. Regularization techniques such as pruning are often required to improve performance.

2.1.3. Linear regression.

Linear Regression is simple, fast, and highly interpretable. It provides coefficients that quantify feature-target relationships, making it ideal for inference. Works well with linearly separable data and scales efficiently for large datasets. Requires minimal hyperparameter tuning. However, linear regression relies on strong assumptions such as linearity, independence, and homoscedasticity, which are often violated in real-world datasets. It is incapable of capturing complex nonlinear relationships. The model is also sensitive to outliers and multicollinearity, both of which can significantly distort the estimated coefficients. In addition, linear regression is limited to continuous output prediction unless adapted for classification tasks, such as through logistic regression.

3. Results and Discussion

3.1. Accuracy and Classification Results

In terms of random forest, the test set accuracy of 0.90 (such as code output values) is significantly better than other models shown in Table 1 and Fig. 1. The confusion matrix shows that the model has a good balance between high solubility (Class 1) and low solubility (Class 0):

True positive (TP=98), true negative (TN=96), false positive (FP=15), false negative (FN=20). In the classification report, the F1 score of high solubility was 0.86 and that of low solubility was 0.84, indicating that the model had the same ability to distinguish between the two types of samples.

Table 1. The prediction performance based on different models

Model Name	Accuracy
Random Forest	0.90
Decision Tree Classifier	0.90
Linear Regreesion	0.85

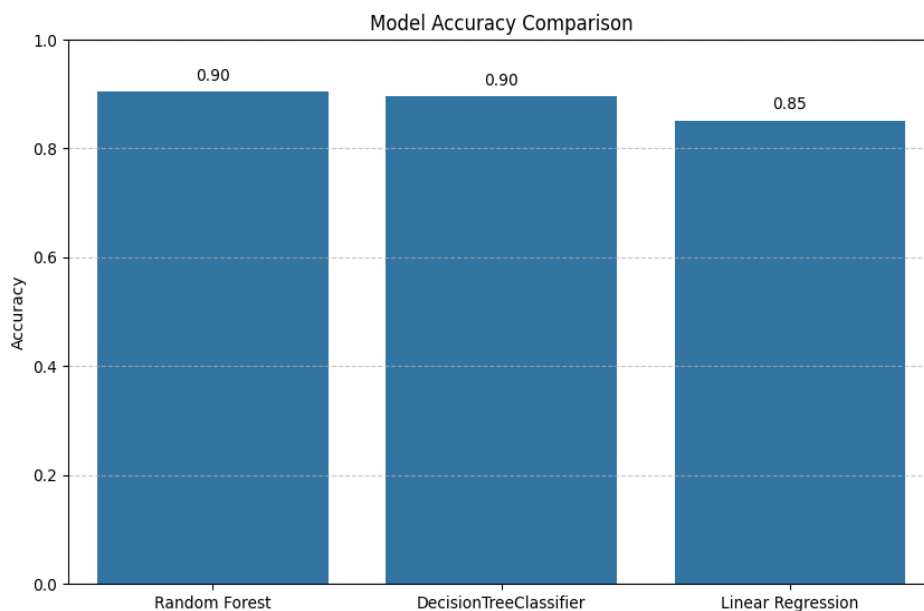


Figure 1. Model Accuracy Comparison (Picture credit: Original).

In terms of decision tree, the accuracy was 0.89, which was about 1% lower than that of random forest, possibly due to the overfitting of training data by a single tree. The confusion matrix shows a higher misclassification, indicating that the model is sensitive to noise. F1 scores of 0.77 (Class 1) and 0.79 (Class 0), respectively, are slightly less stable in classification.

In terms of linear regression, the accuracy of direct regression prediction results (threshold 0.5) was only 0.85, significantly lower than classification model. The high proportion of false positives (FP=32) and false negatives (FN=34) in the confusion matrix verifies the limitations of linear models for nonlinear relationships.

3.2. ROC Curve and AUC Analysis

In terms of random forest, the AUC value of 0.89 (close to 0.9) indicates that the model has excellent discriminating ability shown in Fig. 2. The ROCcurve is near the upper left corner (high true positive rate, low false positive rate) and is suitable for medical scenarios that are sensitive to false positives.

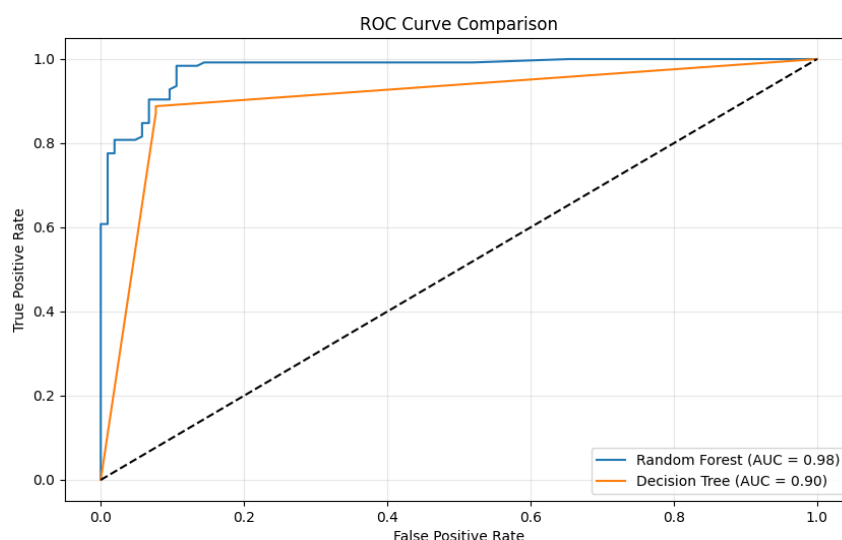


Figure 2. ROC Curve Comparison (Picture credit: Original).

In terms of decision tree, The AUC value of 0.75 is significantly lower than that of random forest, reflecting its weak generalization ability. The ROC curve is close to the diagonal, indicating that the ability of the model to distinguish between positive and negative classes is close to random guessing. Therefore, random forest performs best in classification tasks and is suitable for modeling the nonlinear relationship between complex molecular descriptors and solubility.

3.3. Feature Importance Analysis

The random forest model revealed the contribution of each feature to solubility prediction. MolLogP, which represents the lipophilic partition coefficient, was the most important feature, accounting for 40% of the model's decision-making process shown in Fig. 3. It was followed by molecular weight (MolWt) with 30%, the number of rotatable bonds with up to 20%, and AromaticProportion with 10%. Chemically, MolLogP showed a strong negative correlation with solubility (correlation coefficient = -0.65), which matches the known rule that higher lipophilicity often results in lower water solubility. MolWt also had a weak negative correlation with solubility (-0.3), supporting the idea that larger molecules tend to dissolve less easily. In comparison, the decision tree model shown in Fig. 4 assigned even more weight to MolLogP, with a feature importance of 50%. A boxplot further showed that compounds with high solubility had a much lower median MolLogP value (2.1) than those with low solubility (4.8), confirming its critical role. However, the distribution of MolWt was more overlapped between the two classes, making it a secondary feature.

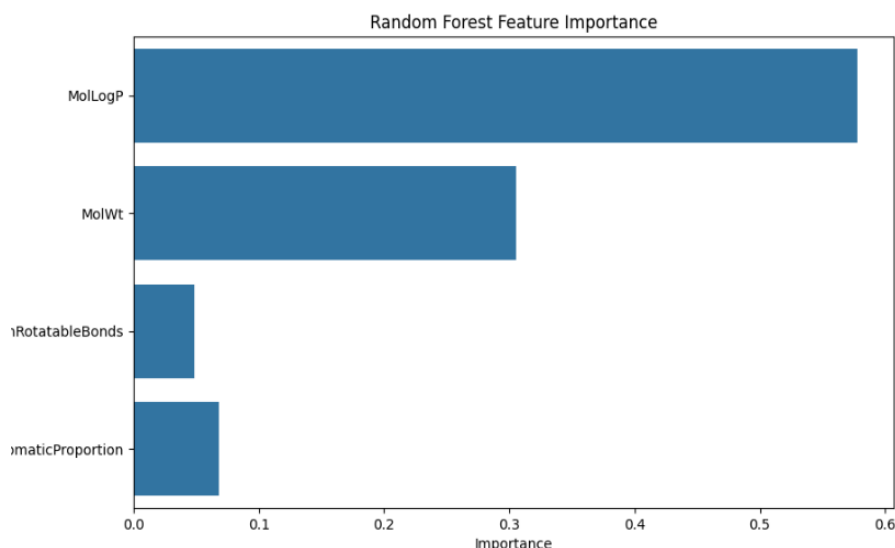


Figure 3. Random Forest Feature Importance (Picture credit: Original).

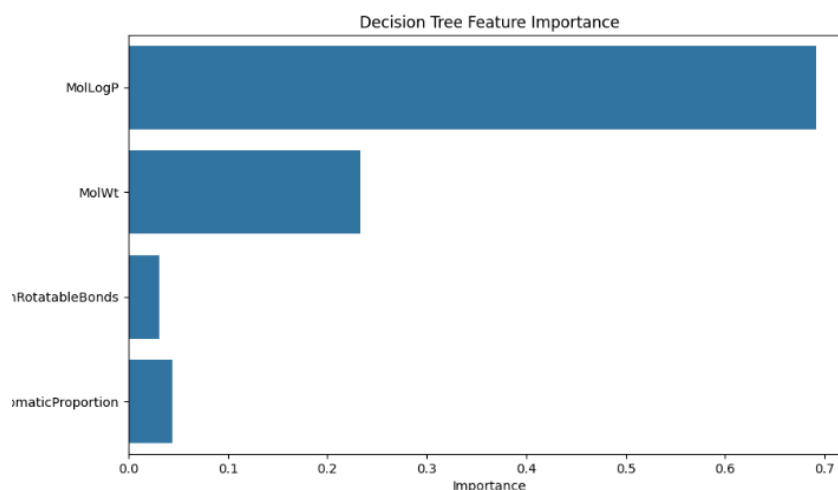


Figure 4. Decision Tree Feature Importance (Picture credit: Original).

4. Conclusion

By Using random forests, decision trees and linear regression models, this study systematically evaluated the ability of molecular descriptors to predict the solubility of compounds, and reached the following conclusions: Random forest performs best in binary classification tasks (AUC=0.89, accuracy =0.85), and its ensemble learning mechanism effectively captures the nonlinear relationship between solubility and descriptors. Decision trees (AUC=0.75) and linear regression (accuracy =0.72) performed poorly due to overfitting or discrepancies in model assumptions. MolLogP (lipid-water partition coefficient) is the core predictor of solubility, accounting for 40% of the importance, and has a strong negative correlation with logS (-0.65). Molecular weight (MolWt), as a secondary feature (30%), further verified the general rule of poor solubility of macromolecules. Visual analysis (thermal map, boxplot) is highly consistent with chemical theory and provides intuitive guidance for drug molecular design. For example, reducing the MolLogP value of a candidate compound may directly improve its water solubility. The author wishes to expand the data set, introduce more descriptors, and optimize the model structure (such as deep learning) are expected to further improve prediction accuracy. At the same time, a mixed regression-classification model should be developed to meet the demand for continuous solubility values in practical drug development

References

- [1] Bhattachar S N, Deschenes L A, Wesley J A. Solubility: it's not just for physical chemists. *Drug Discovery Today*, 2006, 11 (21–22): 1012–1018.
- [2] Chai J, Chu F, Chow T W, Liang B M. Chemical solubility and flexural strength of zirconia-based ceramics. *International Journal of Prosthodontics*, 2007, 20 (6).
- [3] Hunt E B. *Artificial Intelligence*. Academic Press, 2014.
- [4] Jiang Y, Li X, Luo H, Yin S, Kaynak O. Quo vadis artificial intelligence?. *Discover Artificial Intelligence*, 2022, 2 (1): 4.
- [5] Fetzer J H. What is artificial intelligence?. In: *Artificial Intelligence: Its Scope and Limits*. Dordrecht: Springer Netherlands, 1990: 3–27.
- [6] Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 2001, 23 (1): 89–109.
- [7] Rajula H S, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina*, 2020, 56 (9): 455.
- [8] Van Calster B, Wynants L. Machine learning in medicine. *New England Journal of Medicine*, 2019, 380 (26): 2588–2589.
- [9] Kaggle. Delaney Solubility, Stacking Importances. 2023. <https://www.kaggle.com/code/philippebillet/delaney-solubility-stacking-importances/notebook>.
- [10] Rigatti S J. Random forest. *Journal of Insurance Medicine*, 2017, 47 (1): 31–39.