Predicting the Biodegradability of Chemicals from QSAR Data

Ruimian Dai

***Abstract*—Non-biodegradable chemicals are significant contributors to ecological degradation. This study examined the utilization of machine learning models to categorize and forecast the biodegradability of compounds using Quantitative Structure-Activity Relationship (QSAR) data. We employed three commonly utilized techniques, namely logistic regression, k-nearest neighbor, and naïve Bayes, to categorize and forecast the biodegradability of a total of 1055 compounds. Prior to applying each technique, we conducted data preprocessing. Subsequently, we evaluated each model using accuracy, sensitivity, and ROC. The results demonstrate that the k-nn model has superior accuracy, specificity, and sensitivity. The work we conducted has developed essential methods for promptly evaluating the biodegradability of newly synthesized organic compounds in environmental hazard assessments.**

***Index Terms*—Machine learning, QSAR, logistic regression, k-nearest neighbor, naive bayes.**

I. Introduction

BIODEGRADATION is the process by which microorganisms break down organic matter, thereby reducing the accumulation of pollutants. In the face of the increasing pollution of the environment by non-biodegradable chemicals, the importance of this biological mechanism has become particularly obvious. The non-biodegradable substances pose complex threaten to environmental sustainability and public health.

Within this context, the need for accurate assessment methods to predict the biodegradability of chemicals is not only practical but also a regulatory requirement. Instruments like the European Union's REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals) framework reflect a legislative response to this need by integrating biodegradability as a critical parameter in the chemical evaluation process. Due to the limited availability of biodegradability data, the use of accurate predictive models is essential.

Utilizing QSAR models can aid in resolving this issue. These models utilize the molecular structure of compounds to predict their biodegradability, employing a technique to evaluate chemicals that have not been studied. The recent progress in computational techniques, namely in machine learning, has significantly enhanced the potential of QSAR models. Logistic Regression, K-Nearest Neighbors (KNN), and Naive Bayes are more accurate methods for categorizing compounds. This study investigates the pragmatic utilization of these machine learning models on QSAR data, with the preprocessing of the data prior to implementation. We facilitate the succeeding study by thoroughly analyzing each stage, encompassing data cleansing, normalization, and employing statistical techniques to handle outliers. We want to enhance the chemical biodegradability predictive modeling by conducting a thorough evaluation of the model's performance using quantifiable measures. We evaluate current proficiency and potential advancements in this particular field.

II. Data Processing

A. Biodegradability Dataset

Biodegradability was categorized for 1055 chemical substances to achieve this study's goals. These chemicals were selected from a larger pool from the National Institute of Technology and Evaluation of Japan (NITE) database (mansouri2013). This ensured that the QSAR model was relevant and trustworthy, ensuring its implementation. This dataset is used to construct reliable prediction models to help identify chemical biodegradability, hence its integrity is crucial.

B. Data Cleaning

We removed rows of missing data as well as duplicates, and used the Z-score method to deal with outliers data that could affect the model's fitting results.

C. Exploratory Data Analysis

We applied the exploratory data analysis (EDA) phase and generated a histogram for each characteristic to analyze the data distribution, then using violin plots to show the distribution in further depth and identify any correlations among the variables.

D. Feature Engineering and Selection

We applied the StandardScaler function to scale the features, ensuring that all variables have an equal impact on the model's performance by adapting them to a common scale. Mean centering and variance scaling are essential for models that are susceptible to fluctuating size, such as KNN.

E. PCA for dimensionality reduction

Principal Component Analysis (PCA) was utilized to reduce the dimensionality of the dataset, while preserving 95% of its variance. This approach decreased computational complexity and alleviated the curse of dimensionality, which can affect model performance, particularly in datasets with a high number of dimensions, such as QSAR datasets.

III. Methodology

Three classification modeling methods were applied in order to find the appropriate relationship between molecular structures, encoded in molecular descriptors, and the biodegradability of chemicals: Logistic Regression, k-nearest neighbors (kNN) and Naïve Bayes. The application of methods based on different mathematical strategies aimed to better explore the chemical space and balance potential biases related to each single modeling algorithm.

A. Logistic Regression

Logistic Regression (LR) has long been recognized as a critical tool for QSAR modeling, offering a robust framework for the classification and prediction of chemical properties. [2] Logistic Regression maps any real-valued number into a value between 0 and 1, interpretable as the probability of a particular class or event occurring. The logistic function is expressed mathematically as:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{1}$$

where $p(x)$ is the probability of the occurrence of the positive class, β_0 is the intercept term, β_1 represents the vector of coefficients, and x is the vector of input features.
To enhance the Logistic Regression model and guard against overfitting—a scenario where the model performs well on training data but poorly on unseen data—we implement regularization. Regularization techniques adjust the complexity of the model, penalizing the magnitude of the coefficients, thereby encouraging simpler models that perform better on new data. Our model incorporates a regularization term in the loss function, leading to the following adjusted loss function:

$$\mathcal{L} = L_0 + \frac{\lambda}{2}\|\beta\|^2 \tag{2}$$

Here, L_0 is the original loss function based on maximum likelihood estimation, and λ represents the regularization strength, which has been empirically set to 0.01 in our work. This regularization term adds a constraint to the size of the coefficients, effectively shrinking them towards zero and, as a result, reduces model complexity and potential overfitting. The weight update rule is modified to include this regularization term:

$$\Delta W = \Delta W + \lambda \cdot \text{weights} \tag{3}$$

Thus, each update to the weights during the training phase takes into account both the error reduction and the regularization term, which promotes a model that generalizes well.

B. k-Nearest Neighbors (k-NN)

The k-nearest neighbor algorithm (k-NN) is a method to make classification and prediction based on the features of the nearest points [3]. The proximity between data points is calculated using the Euclidean distance, expressed mathematically as:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two points in the dataset.
Additionally, to avoid overfitting or underfitting, we need to choose a optimal k. Because, if k is too small, the model will capture the noise of the data, if k is too large, the model will overlook some feature making the prediction too smooth.
So, we used a cross-validation approach which involves partitioning the training set into a number of subsets, or folds. The model is trained on all but one-fold and validated on the remaining fold. This procedure is iterated, with each fold serving as the validation set once. The average accuracy will be calculated to determine the best k value. Thus, it can prevent overfitting and underfitting to a certain extent.

C. Naive Bayes

The Naive Bayes classifier is a probabilistic model in machine learning that applies Bayes' theorem assuming the independence of features [4]. The Naive Bayes classifier generates the posterior probabilities which were given out directly based on the core function, as defined by:

$$P(y|x_1, \ldots, x_n) = P(y)\prod_{i=1}^{n} P(x_i|y) \tag{5}$$

where y is the class variable, and x_i are the individual features. We calculate the prior probability $P(y)$ and the likelihood $P(x_i|y)$ for each class. Each feature in the class is modeled using a Gaussian distribution, where the likelihood is determined by the mean and variance derived from the training data. The probability density function for the Gaussian distribution is given by:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{6}$$

When making predictions, the model calculates the log of these values to avoid underflow issues common with small probabilities. The log posterior for each class is the sum of the log prior and the sum of the log likelihoods for all features. The predicted class is then the one with the highest posterior probability.

The incorporation of this probabilistic framework, along with the assumption of feature independence, naturally prevents overfitting by reducing the intricacy of the model. Naive Bayes can provide reliable predictions even when dealing with the complex nature of QSAR datasets by utilizing feature probabilities that are averaged throughout the entire dataset. This allows for good generalization from the training data to unseen data, despite the high dimensionality of the datasets.

IV. Model Analysis

A. Performance Metrics Introduction

A complete set of indicators is used to thoroughly evaluate the performance of our QSAR models during critical assessment. To compare the results of different model, the classification models were evaluated based on the counts of true positives (TP), true negatives(TN), false positives (FP), and false negatives (FN). The sensitivity is precisely defined as $Sn = \frac{TP}{TP+FN}$, measuring the model's ability to correctly identify

positive instances. The specificity is $Sp = \frac{TN}{TN+FP}$, reflecting the model's proficiency in recognizing negative instances [5]. The overall predictive accuracy is $Q = \frac{TP+TN}{TP+TN+FP+FN}$, and the balanced accuracy, $BA = \frac{Sn+Sp}{2}$, ensures even consideration of both classes irrespective of their distribution [5].

The publication by Cheng [6] in 2012 featured the receiver operating characteristic (ROC) curve. The receiver operating characteristic curve, sometimes referred to as the ROC curve, simplifies the assessment of the trade-offs between the rates of correctly identified positive cases and incorrectly identified positive cases. This facilitates the establishment of the optimal model and threshold for decision-making, aligning with the organization's best interests.

The metrics not only offer a thorough and comprehensive evaluation of the model's performance, but they also provide extensive and profound insights into the model's ability to predict chemicals that are easily broken down (RB) and chemicals that are not easily broken down (NRB). Sensitivity and specificity are intricately linked in a mutually advantageous feedback loop.

B. Data Split

We used the `train_test_split` function from the scikit-learn library to split the processed dataset into train set and test set by a ratio of 7:3.

C. Model Evaluation and Results

Based on the data presented in Table I, it can be observed that the KNN model demonstrates the highest overall accuracy (Q) of 0.88 and balanced accuracy (BA) of 0.86. This indicates that it is the most consistent model across both classes (RB and NRB of molecules) among the three models that were evaluated. Naive Bayes and Logistic Regression both have overall accuracy that are comparable to one another, coming up at 0.85 and 0.83, respectively. However, Logistic Regression has a higher balanced accuracy, which suggests that it may be able to handle class imbalances more effectively than Naive Bayes.

TABLE I
COMPARISON OF MODEL PERFORMANCE METRICS

Model	**Sp**	**Sn**	**Ba**	**Q**
Logistic regression	0.88	0.80	0.84	0.85
KNN	0.92	0.80	0.86	0.88
Naïve bayes	0.88	0.74	0.81	0.83

With a value of 0.92, the KNN model has the highest specificity (Sp), which is a measurement of the true negative rate. This indicates that it is the most effective model for accurately recognizing NRB molecules at this time.The true positive rate, also known as sensitivity (Sn), is the same for Logistic Regression and KNN, both of which have a value of 0.80. This value is greater than the value of 0.74 for Naive Bayes. Given this information, it can be deduced that Logistic Regression and KNN are superior to Naive Bayes when it comes to recognizing RB molecules.

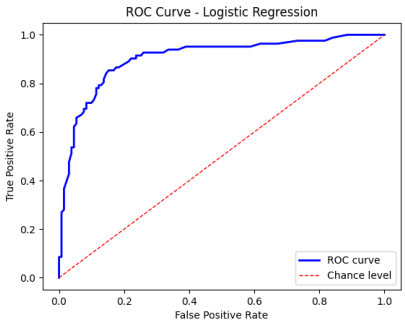

Fig. 1. Logistic Regression

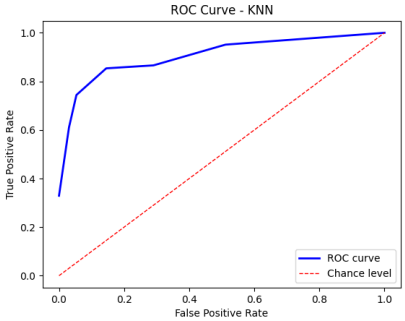

Fig. 2. KNN

Because the ROC curves for all three models are located above the chance level (diagonal line), it can be deduced that all of the models have a satisfactory capacity to differentiate between the two levels of classification.

According to Figures 1, 2, and 3, the ROC curve of the KNN model is closer to the top-left corner. This indicates that the KNN model has a better trade-off between the true positive rate and the false positive rate, which represents superior performance in classification tasks.Logistic Regression and Naive Bayes both provide good ROC performance; however, the ROC curve for Logistic Regression indicates that it has a marginally higher capacity for discriminating than Naive Bayes.

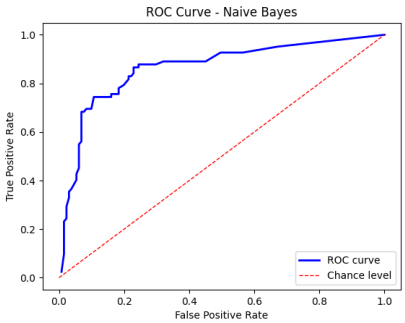

Fig. 3. Naive Bayes

V. Conclusion and Recommendation

We performed a thorough investigation and assessment of three distinct machine learning techniques to construct a dependable classification QSAR model for predicting ready biodegradability. The K-Nearest Neighbors (KNN) algorithm demonstrated superior performance, reaching the highest overall accuracy and balanced accuracy. Hence, it is the optimal selection for categorizing molecules as either readily biodegradable (RB) or non-readily biodegradable (NRB).

However, each model possesses advantages and disadvantages. Although KNN exhibits great accuracy, it is computationally demanding, posing a problem for handling expanding datasets. Logistic regression, in contrast, yields useful probabilistic outcomes and exhibits a minor advantage in distinguishing between categories, as evidenced by the ROC curves. In addition, although Naive Bayes may have poorer accuracy in this specific scenario, it is widely respected for its computational efficiency and remains a suitable choice for huge datasets with high dimensions when speed is crucial.

Considering these criteria, we suggest use the KNN model for estimating the biodegradability of compounds. The model exhibits exceptional accuracy and a favorable balance between sensitivity and specificity, making it one of the most efficient of the models examined.

References

[1] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni, "Quantitative Structure–Activity Relationship Models for Ready Biodegradability of Chemicals," *Journal of Chemical Information and Modeling*, vol. 53, no. 4, pp. 867–878, 2013, doi: 10.1021/ci4000213.

[2] W. Dongbin, Z. Aiqian, W. Zhongbo, H. Shuokui, and W. Liansheng, "A Case Study of Logistic QSAR Modeling Methods and Robustness Tests," *Ecotoxicology and Environmental Safety*, vol. 52, no. 2, pp. 143–149, 2002, doi: 10.1006/eesa.2002.2168.

[3] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[4] P. Watson, "Naïve Bayes Classification Using 2D Pharmacophore Feature Triplet Vectors," *Journal of Chemical Information and Modeling*, vol. 48, no. 1, pp. 166–178, 2008, doi: 10.1021/ci7003253.

[5] M. Lee and K. Min, "A Comparative Study of the Performance for Predicting Biodegradability Classification: The Quantitative Structure–Activity Relationship Model vs the Graph Convolutional Network," *ACS Omega*, vol. 7, no. 4, pp. 3649–3655, 2022, doi: 10.1021/acsomega.1c06274.

[6] F. Cheng et al., "In Silico Assessment of Chemical Biodegradability," *Journal of Chemical Information and Modeling*, vol. 52, no. 3, pp. 655–669, 2012, doi: 10.1021/ci200622d.