Predicting and Cross Validating kNN, LR, and SVM Models the Biodegradability of Chemicals from QSAR Data

Mafazul Huda Muhammad Syed, 230118234,

***Abstract*—This study assesses chemical biodegradability using QSAR data, employing machine learning models like k-Nearest Neighbors (kNN), Logistic Regression (LR), Support Vector Machines (SVM), and Partial Least Squares Discriminant Analysis (PLSDA). Key steps include data preprocessing, feature scaling, and 5-fold cross-validation for model robustness. Performance evaluation is based on accuracy, precision, recall, and AUC. Logistic Regression is identified as the optimal model due to its high accuracy (86.25%), and balanced precision and recall. This approach offers a valuable tool for environmental impact assessment of chemicals.**

***Index Terms*—Article submission, IEEE, IEEEtran, journal, LaTeX, paper, template, typesetting.**

I. Introduction

A. Background, Introduction, and Ethics

IN the expansive domain of environmental science and pharmacology, the evaluation of chemical biodegradability is essential. This assessment is not only crucial for evaluating the environmental footprint of various chemicals but also plays a pivotal role in preventing their potential accumulation and dispersion, which can lead to long-lasting ecological and health impacts. The ability to predict whether a chemical is readily biodegradable or not is therefore essential for environmental protection and sustainable chemical design.

QSAR is a method employed in chemoinformatics, a field that merges chemistry with computational modeling, to create models that relate the structural attributes of molecules to their biological or chemical activities. By leveraging QSAR models, it is possible to infer a chemical's biodegradability based on its molecular structure, providing a critical tool in environmental chemistry and pharmacology for assessing the ecological impact of chemicals before they are even synthesized.

The QSAR biodegradation dataset [1] serves as the foundation for this study. This dataset is a collection of 1,055 chemicals, each characterized by 41 distinct molecular descriptors. These descriptors include various molecular properties such as the leading eigenvalue from the Laplace matrix, the Balaban-like index, the number of heavy atoms, and frequencies of specific molecular bonds, among others [4]. The dataset is structured in a way that each chemical is categorized into one of two classes: ready biodegradable or not ready biodegradable, offering a binary classification problem that is suitable for the application of machine learning techniques.

B. Objective

This report aims to utilize machine learning to build, cross validate, and verify a predictive model that can accurately classify chemicals into these two categories based on their QSAR profiles. The method involves an exploration of various machine learning algorithms, including k-Nearest Neighbors, Logistic Regression, Support Vector Machines, and Partial Least Squares Discriminant Analysis, each offering unique perspectives and capabilities in handling the dataset's complexities.

II. Data Processing

The foundation of any robust machine learning model lies in meticulous data processing. In this study, the QSAR dataset, including 1,055 chemicals each described by 41 molecular descriptors, underwent several preprocessing steps to ensure data quality and suitability for modelling.

A. Initial Data Handling and Preprossessing

Initially, the dataset was checked for missing values. Any instances with missing data were identified and removed to maintain data integrity, ensuring that the models trained on this dataset are based on complete information.

Duplicate entries in the dataset can skew the model's learning process, leading to overfitting. Therefore, all duplicate rows were identified and removed, ensuring that each chemical in the dataset is unique. This step was crucial in maintaining the diversity and representation of the dataset.

Given the varying scales of the molecular descriptors, feature scaling was applied to standardize the data. This process involved normalizing the features (subtracting the mean and dividing by the standard deviation), which is essential in models where distance metrics are crucial, such as k-Nearest Neighbors (k-NN).

B. Dimensionality Reduction and Data Transformation

Principal Component Analysis (PCA): To address the high dimensionality of the dataset and to extract the most informative features, PCA was applied. This technique transforms

This paper was produced by Mafazul Huda Muhammad Syed, in Sheffield, UK.

Manuscript received April 19, 2021; revised August 16, 2021.

the original features into a set of linearly uncorrelated components, capturing the maximum variance in the data with fewer dimensions. The number of components retained, which cumulatively explained at least 95% of the variance, was determined based on the explained variance criterion.

The dataset was divided into training and test sets for model validation, which is essential for assessing the model's performance on unseen data and ensuring the generalizability of the predictive models.

The preprocessing of the QSAR dataset was conducted with the objective of enhancing the quality and reliability of the subsequent models. By accurately addressing missing and duplicate data, appropriately scaling features, and reducing dimensionality through PCA, the dataset was optimally prepared for applying various machine learning techniques. This thorough data processing lays a solid foundation for developing robust and accurate models for predicting the biodegradability of chemicals.

III. Methodology

The methodology of this study involves the implementation and analysis of four distinct machine learning models: k-Nearest Neighbors (k-NN), Logistic Regression, Support Vector Machines (SVM), and Partial Least Squares Discriminant Analysis (PLSDA). Each model was rigorously tested using 5-fold cross-validation to ensure robustness and to mitigate overfitting.

A. k-Nearest Neighbor Approach

The k-Nearest Neighbor (k-NN) model is a versatile and intuitive machine learning algorithm used in this study to predict the biodegradability of chemicals based on QSAR data.

The k-NN algorithm functions on the principle of feature similarity, predicting the classification of a new sample based on the majority vote of its 'k' nearest neighbors in the feature space. Mathematically, the similarity between data points is often measured using distance metrics, with Euclidean distance being the most common. For a given test sample, the algorithm identifies 'k' training samples closest to it and predicts the class based on the predominant class among these neighbors. [6]

A range of 'k' values [1, 3, 5, 7, 9] was considered for the k-NN model. The optimal 'k' value was determined based on the model's performance, balancing between bias and variance.

To prevent overfitting and assess the model's generalizability, 5-fold cross-validation was implemented. The dataset was divided into five subsets, with each subset serving as a test set while the remaining subsets formed the training set in each fold. This approach ensures that every data point is used for both training and testing, providing a comprehensive evaluation of the model's performance.

Choosing the appropriate 'k' value based on cross-validated performance ensured a balance between underfitting and overfitting.

B. Logistic Regression

The Logistic Regression model is a fundamental machine learning technique used in this study for the binary classification of chemical biodegradability based on QSAR data.

Logistic Regression is a statistical method for modeling the relationship between a dependent binary variable and one or more independent variables. The model estimates the probability of a binary response based on one or more predictor variables using the logistic function. The logistic function, often referred to as the sigmoid function, is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

where z is the linear combination of features and weights (coefficients); the logistic function is used to model the probability of a chemical being biodegradable. [7]

To account for the intercept in the logistic regression model, a column of ones was added to the feature matrix. The logistic regression model parameters (weights) were estimated using the Newton-Raphson method, an iterative optimization algorithm. The algorithm updates the model parameters by approximating the Hessian matrix and gradient of the loss function. Ridge regularization (L2 penalty) was incorporated into the model to prevent overfitting by penalizing large coefficients. This regularization technique is crucial when dealing with multicollinearity or when the number of predictors is high compared to the number of observations. The iterative process continued until the change in the weight vector was less than a predefined threshold (epsilon), ensuring convergence to an optimal solution.

C. Support Vector Machine (SVM) Model

The Support Vector Machine (SVM) model is a powerful and widely-used machine learning algorithm, employed in this study for the classification of chemical biodegradability based on QSAR data.

SVM is a supervised learning model that finds a hyperplane in an N-dimensional space (N - number of features) that distinctly classifies the data points into different categories. In the context of binary classification:

The goal is to find the optimal hyperplane that separates the data points with maximum margin. The hyperplane is defined mathematically as the set of points x satisfying

$$w^T x + b = 0 \tag{2}$$

where w is the weight vector and b is the bias. [8]

Support Vectors: These are the data points nearest to the hyperplane, and they help in determining the position and orientation of the hyperplane. The model aims to maximize the margin between the hyperplane and the nearest points from each class, known as support vectors. For non-linearly separable data, SVM uses a kernel function to transform the input space into a higher-dimensional space where a linear separation is possible. In this study, the Radial Basis Function (RBF) kernel was utilized.

The SVM implementation in this study inherently includes regularization, where the trade-off between maximizing the margin and minimizing the classification error is balanced. This regularization helps prevent the model from fitting too closely to the training data.

D. Partial Linear Squares Discriminant Analysis (PLSDA)

The Partial Least Squares Discriminant Analysis (PLSDA) model is a statistical technique used in this study for predicting the biodegradability of chemicals based on QSAR data.

PLSDA is an extension of Partial Least Squares (PLS) regression, tailored for classification problems. It combines features of both principal component analysis (PCA) and multiple regression, and is particularly useful when predictors are many and highly collinear.

PLSDA starts with PLS regression, which models the relationship between independent variables (X) and dependent variables (y) by extracting a set of orthogonal latent variables (LVs) that maximize the covariance between X and y.

In PLSDA, the continuous output from PLS regression is used for classification. By setting a threshold (typically 0.5 for binary classification), the continuous predictions are converted into class memberships.

PLSDA was applied using an optimized number of latent variables (numLVs) to balance model complexity and predictive accuracy, ensuring a crucial balance between overfitting and capturing essential variance. This approach was aligned with methods from previous studies [3] and used for verification against the original data source [1].

E. Cross Validation and Performance Evaluation Method

1) K-Fold Cross Validation: In this study, a 5-fold cross-validation approach was employed for each model to assess their performance and prevent overfitting. The dataset was divided into five subsets, with each subset serving as a test set while the remaining subsets form the training set. Every data point is used for both training and testing, providing a comprehensive evaluation of the model's performance. A single parameter, 'k', refers to the number of groups that a given data sample is to be split into; In this case, 'k' is set to 5, as it was the optimal balance between computational time and accuracy [2]. K-Fold cross-validation evaluates a model's generalization ability, reducing overfitting risk and offering a more accurate performance estimate.

2) Performance Metrics: The accuracy, precision, recall (sensitivity), AUC, and confusion matrix were calculated for each fold and then averaged across all folds to provide a comprehensive assessment. This uniform approach in evaluating the models provides a fair basis for comparison across different algorithms and ensures that the strengths and weaknesses of each model are accurately captured and understood.

IV. Model Analysis

The data visualisation via confusion matrices for all four models are shown in Figures 1 to 4. The accuracy, recall, and precision of each model is calculated from the data in the confusion matrices.

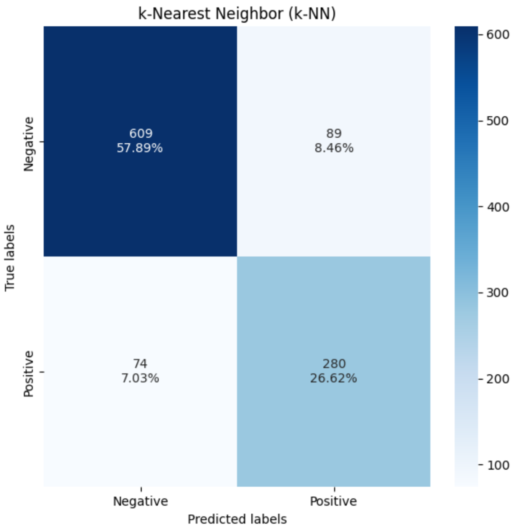

Fig. 1. Total Confusion Matrix for kNN Model - Averaged Across All Folds

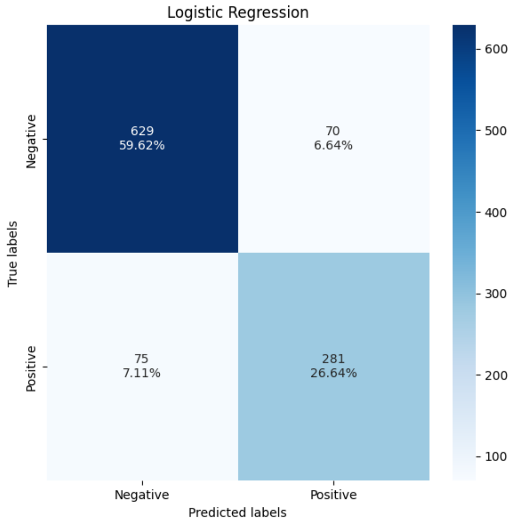

Fig. 2. Total Confusion Matrix for Logistic Regression Model - Averaged Across All Folds

A. Confusion Matrices for Each Model

B. Performance Metrics

Table I presents a comparative summary of four machine learning models based on performance metrics such as accuracy, AUC (Area Under the Curve), precision, and recall. Precision measures the accuracy of positive predictions, recall assesses how well a model identifies all positive cases, and average AUC (Area Under the Curve) reflects the model's ability to distinguish between classes, with higher values indicating better performance [5].

The kNN model demonstrated a solid performance with an accuracy of 85.93%, an AUC of 0.90, precision at 0.89, and recall at 0.87. These results suggest a balanced classification capability for both positive and negative cases. Logistic Regression slightly surpassed kNN, achieving an accuracy of 86.25%, a marginally higher AUC of 0.91, and maintaining similar precision but with an improved recall of 0.90. This

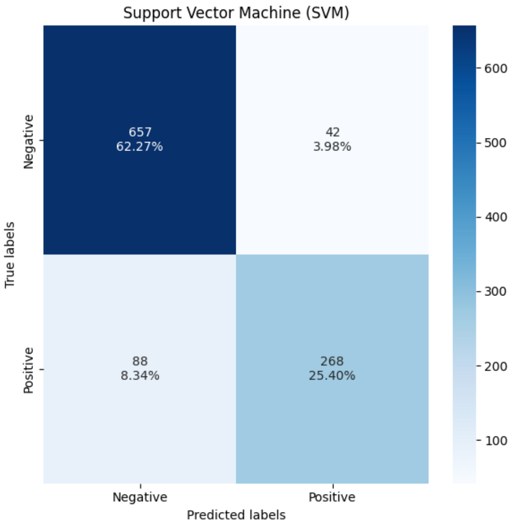

Fig. 3. Total Confusion Matrix for SVM Model - Averaged Across All Folds

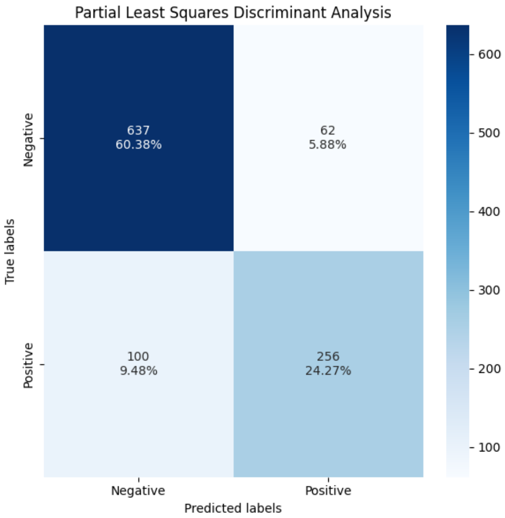

Fig. 4. Total Confusion Matrix for PLSDA Model - Averaged Across All Folds

indicates a slightly better performance in identifying positive cases while maintaining balanced precision.

The SVM model outperformed the others in terms of accuracy and AUC, registering 87.68% and 0.93, respectively. Although its precision was marginally lower at 0.88, its recall was notably higher at 0.94, indicating a particular strength in correctly identifying positive instances. Lastly, the PLSDA model, while having the lowest accuracy of 84.64%, showed a competitive AUC of 0.91. Its precision was slightly lower at 0.87, but it achieved a high recall of 0.91, suggesting a strong ability to identify positive cases, albeit with a slight trade-off in overall accuracy.

Model	Accuracy	AUC	Precision	Recall
kNN	85.93%	0.90	0.89	0.87
Logistic Regression	86.25%	0.91	0.89	0.90
SVM	87.68%	0.93	0.88	0.94
PLSDA	84.64%	0.91	0.87	0.91

TABLE I
PERFORMANCE METRICS OF MACHINE LEARNING MODELS

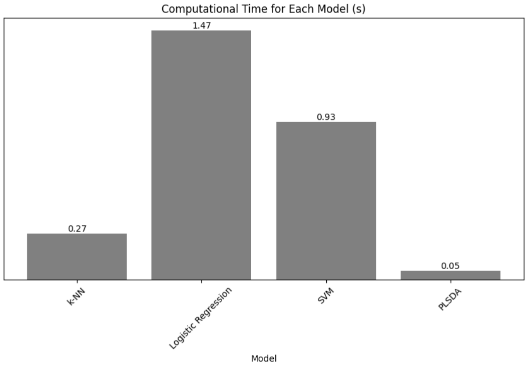

Fig. 5. Computational Time for Each Model (in seconds)

In terms of computational efficiency, shown in Figure 5, logistic regression consumes the most time, largely due to its Newton-Raphson iteration method. PLSDA is the fastest, leveraging built-in functions and serving as a verification mode. kNN follows PLSDA in efficiency, with SVM, despite its high accuracy, being less efficient than kNN. The accuracies of all four models are within 3% of those reported in [1], confirming their reliability and validating their effectiveness for this study.

V. CONCLUSION AND RECOMMENDATION

In this study, various machine learning models were compared for predicting the biodegradability of chemicals from QSAR data [1]. Among k-Nearest Neighbors, Logistic Regression, Support Vector Machines, and Partial Least Squares Discriminant Analysis, Logistic Regression is recommended due to its higher accuracy (86.25%), balanced precision and recall, and superior Area Under the Curve (AUC) of 0.91. With being relatively computationally efficient, LR's robustness, adaptability, and clear interpretability of results make it a suitable model for this application.

REFERENCES

[1] K. Mansouri, et. al. “Quantitative structure–activity relationship models for ready biodegradability of chemicals,” *Journal of Chemical Information and Modeling*, vol. 53, no. 4, pp. 867–878, 2013.

[2] [2]I. Kofi Nti, et. al. “Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation,” International Journal of Information Technology and Computer Science

[3] [3]C. Cassel, et. al. “Robustness of partial least-squares method for estimating latent variable quality structures,” Journal of Applied Statistics, vol. 26, no. 4, pp. 435–446, May 1999, doi: https://doi.org/10.1080/02664769922322.

[4] OpenML. (2023). [Online]. Available: https://www.openml.org/search?type=data&sort=runs&id=1494&status=active [Accessed Dec. 11, 2023].

[5] [5]D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,”, Oct. 2020, Available: https://arxiv.org/abs/2010.16061

[6] [6]Subhash Ajmani, et. al. “Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation,” Journal of Chemical Information and Modeling, doi: https://doi.org/10.1021/ci0501286.

[7] [7]Z. Y. Algamal and M. H. Lee, “A novel molecular descriptor selection method in QSAR classification model based on weighted penalized logistic regression,”

[8] [8]J.-P. Doucet, et. al. “Nonlinear SVM Approaches to QSPR/QSAR Studies and Drug Design,” Current Computer Aided-Drug Design,