Matlab Assignment: Predicting the Biodegradability of Chemicals from QSAR Data

Nilsu Atlan - 230221176

***Abstract*—This report delves into the application of data modeling and machine learning algorithms for predicting chemical biodegradability using the Quantitative Structure-Activity Relationship (QSAR) dataset. Emphasizing robust data processing techniques, the report encompasses outlier and duplicate removal, and feature standardization to optimize data quality. It methodically explores implementation of Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Logistic Regression with Gradient Descent as predictive tools. The performance of the models is evaluated through a series of performance metrics.**

I. Introduction

Predicting biodegradability of chemicals, as it is a significant environmental concern, can aid in assessing environmental impact [1]. This report focuses on constructing data processing and training models to predict chemical biodegradability of chemicals using the provided Quantitative Structure-Activity Relationship (QSAR) dataset. QSAR modeling creates predictive models by correlating a compound's structure and molecular traits with its biological activities or physicochemical properties, using machine learning and statistical analysis [2]. This approach is particularly beneficial in environmental science as it aids in the proactive identification of potentially harmful chemicals, thereby informing regulatory decisions and protective measures [3]. The objective of this report is to develop a fully-validated model using advanced data modeling and machine learning techniques. Tools and functionalities available in Matlab are leveraged to facilitate the application of the following algorithms; Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Logistic Regression with Gradient Descent. These methods were chosen due to their proven efficacy in classification tasks, common appearance in QSAR modelling examples, and their ability to handle the complexity and diversity of QSAR data. The SVM algorihm is renowned for its effectiveness in high-dimensional spaces and memory efficiency, making it suitable for datasets with a large number of features like QSAR [4]. The KNN algorithm is a user-friendly and noise-tolerant method that remains effective and interpretable, even with large datasets [4]. Logistic Regression, with its probabilistic approach and gradient descent optimization, effectively predicts class probabilities, allowing threshold adjustments to balance sensitivity and specificity, as shown in ROC curve analysis [5].

While delving into the predictive modeling of chemical biodegradability utilizing QSAR, it's imperative to consider the ethical implications, both in the specific application and in machine learning at large. The use of QSAR modeling is dependent on integrity and transparency of the data. As predictions made relate to human health, environmental policies and regulations, it is crucial to ensure accuracy and reliability of the models and labeling. Additionally, potential bias affecting labeling and data collection should be acknowledged and mitigated.

Throughout the report implementation and comparison of the models are discussed as predictions from the models are observed. The process is crucial for practical implications in environmental science, particularly in predicting biodegradability of chemicals using the presented list of features for that chemical.

Atlan is with the Department of Automatic Control and Systems Engineering, The University of Sheffield. E-mail: natlan1@sheffield.ac.uk

II. Data Processing

The QSAR dataset comprises 1055 molecules, each characterized by 42 distinct attributes, including 41 features and a label indicating biodegradability. The features of the QSAR dataset represent various chemical properties. Regarding biodegradability the data can be visualized with Fig. 1. In terms of, data preprocessing for the data to be ready to trained, duplicate entries and outliers were identified and removed while maintaining the integrity of the dataset. "`Number of rows decreased from 1055 to 1052 with 3 repeating rows removed`" (retrieved from the command window). After the removal of the duplicates, histograms were used to visualize the data as exampled in Fig. 2.

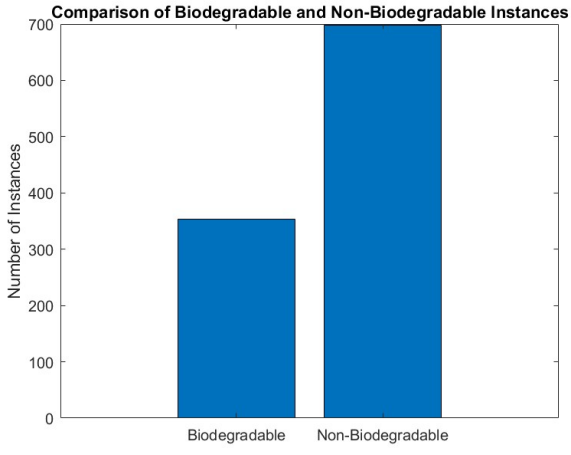

Fig. 1. Comparison of Biodegradable and Non-Biodegradable Instances of the Data

zscore method was used to eliminate the outliers where the code "`Removed 342 rows containing outliers`".

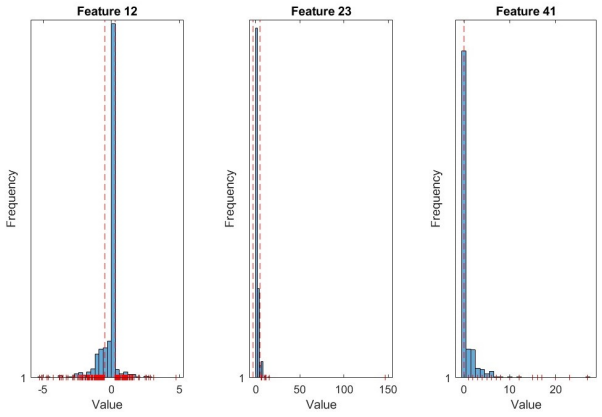

Fig. 2. Examples: Outliers' Effect on Data Distribution

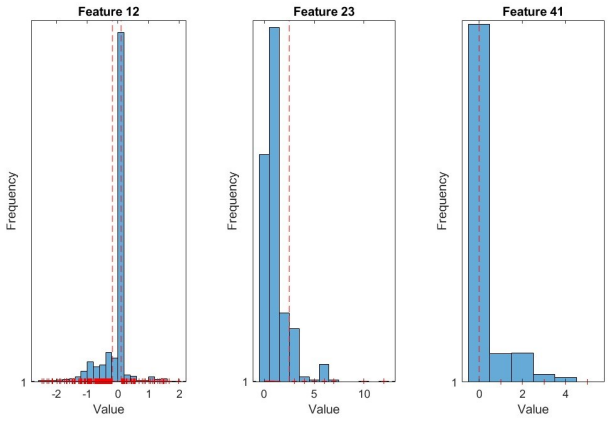

Fig. 3. Examples: Removal of Outliers' Effect on Data Distribution

The effect of removed outliers can be observed in Fig. 3. Standardization of features was conducted to normalize the data range, enhancing the comparability and effectiveness of the machine learning models. The dataset was split into training and testing sets to ensure the robustness of the models. This split was crucial for evaluating the models' performance on unseen data and avoiding overfitting. This preprocessing not only prepared the data for modeling but also provided insights into its underlying structure and distributions.

III. Methodology

This section details over-fitting prevention strategies and the mathematical foundations of utilized training mechanisms. By splitting the data into train (80%) and test (20%) versions prevention of over-fitting was ensured. The methodologies of SVM, KNN, and Logistic Regression were employed. SVM was chosen for its efficiency in higher-dimensional spaces, KNN for its simplicity, and Logistic Regression for its interpretability and suitability for linearly separable data.

A. Support Vector Machines (SVM)

For the SVM method a built-in function is implemented in Matlab. This method is utilized to compare model evaluating metrics. SVMs provide a powerful approach for nonlinear classification by augmenting the feature space with nonlinear basis functions, denoted as $\Phi(\mathbf{x})$. This transformation increases the dimension of decision variables from $\mathbb{R}^n$ to $\mathbb{R}^m$ where $m > n$, posing computational challenges. SVMs overcome this by fitting a model of the form:

$$[f(\mathbf{x};\theta) = \theta_0 + \sum_{i \in S} \theta_i K(\mathbf{x}, \mathbf{x}_i)] \tag{1}$$

where S denotes the indices of support vectors, and $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a kernel function, measuring distance. This model primarily depends on the support vectors, significantly reducing the number of parameters involved, thus decreasing computational complexities [5].

B. K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) classifier operates by first selecting a positive integer k and a test observation $\mathbf{x}_0$. Different values of 'k' ranging from 1 to 20 were tested to identify the most effective number of neighbors.The distance between $\mathbf{x}_0$ and each point in the training set is computed, typically using the Euclidean distance given by:

$$[\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}] \tag{2}$$

where x_i and y_i are the coordinates of the two points in n-dimensional space.

The algorithm identifies the k points in the training data that are closest to $\mathbf{x}_0$, denoted as N_0. It then estimates the conditional probability for each class j as the fraction of points in N_0 belonging to class j:

$$[\Pr(Y = j | X = \mathbf{x}_0) = \frac{1}{k} \sum_{i \in N_0} I(y_i = j)] \tag{3}$$

where $I(y_i = j)$ is an indicator function equal to 1 if $y_i = j$ and 0 otherwise. Finally, KNN classifies the test observation $\mathbf{x}_0$ into the class with the highest estimated probability [6].

C. Logistic Regression with Gradient Descent

The logistic regression model with gradient descent is implemented as follows:

Initial hyper-parameters are established, including the initialization of the model's parameters θ as a zero vector. The maximum number of iterations is set, and ranges for the learning rate (α) and convergence threshold (ϵ) are defined.

$$\text{alphas} = [0.001, 0.01, 0.1, 1] \tag{4}$$

$$\text{epsilons} = [1e-4, 1e-6, 1e-8] \tag{5}$$

For each combination of α and ϵ, the logistic regression algorithm undergoes the following iterative process:

- Compute the weighted sum of inputs (z) using the current parameters (θ):

$$z = X_{\text{train}}\theta \tag{6}$$

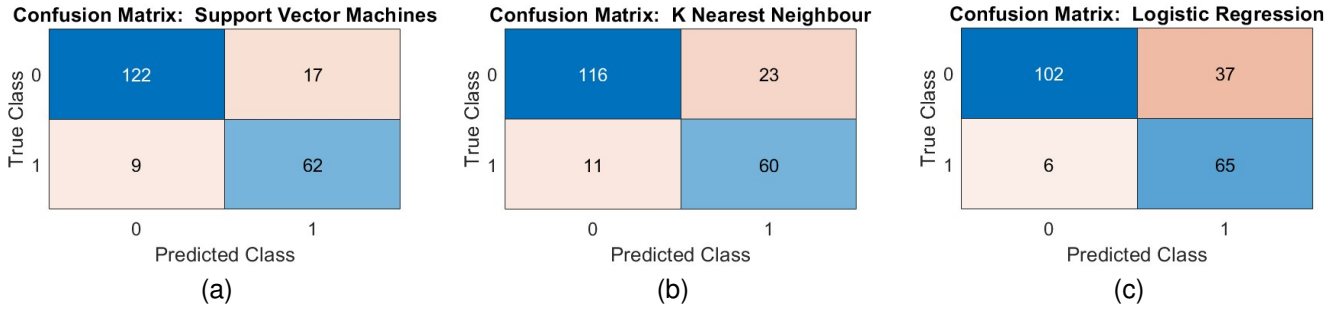

Fig. 4. Confusion Matrices of Training Models (a) Support Vector Machines (SVM) (b) K-Nearest Neighbors (KNN) (c)Logistic Regression with Gradient Descent

- Apply the logistic (sigma) function to z, obtaining the predicted probabilities (σ):

$$\sigma = \frac{1}{1 + e^{-z}} \tag{7}$$

- Calculate the gradient of the cost function:

$$\text{gradient} = \frac{X_{\text{train}}^T(\sigma - y_{\text{train}})}{m} \tag{8}$$

- Update the model parameters (θ) by taking a step proportional to the gradient:

$$\theta_{i+1} = \theta_i - \alpha \times \text{gradient} \tag{9}$$

The iterative process continues until the parameters converge (change in θ becomes negligible) or the maximum number of iterations is reached. After each iteration, predictions are made, and accuracy is calculated to determine the best combination of hyper-parameters. This process ensures the fine-tuning of the model to achieve optimal performance [7].

Each model includes mechanisms to mitigate over-fitting, ensuring robustness and generalization of the results.

IV. Model Analysis

Each model's performance was evaluated using accuracy, sensitivity (True Positive Rate), specificity (True Negative Rate), precision, recall, and the F1 score. Data splits for training and testing were outlined. The analysis includes detailed discussions on confusion matrices and a Receiver Operating Characteristic (ROC) curves, highlighting the models' strengths and weaknesses in predicting chemical biodegradability.

A. Model Evaluation Metrics

- **Accuracy**: Measures the overall correctness of the model, defined as the proportion of true results (both true positives and true negatives) among the total number of cases examined [8].
- **Precision**: Known as the positive predictive value, precision is the proportion of positive identifications that were actually correct. It reflects the model's ability to avoid false positives [8].
- **Recall (Sensitivity)**: Indicates the proportion of actual positives that were correctly identified by the model, showcasing its ability to detect positive instances [8].
- **F1 Score**: The harmonic mean of precision and recall, F1 Score is a single metric that balances both precision and recall, especially useful in uneven class distributions [8].
- **True Positive Rate (Sensitivity)**: The ratio of correctly predicted positive observations to all actual positives. It measures the proportion of actual positives correctly identified as such [8].
- **False Positive Rate**: The ratio of incorrectly predicted positive observations to the total actual negatives. It measures the proportion of actual negatives that were incorrectly classified as positives [8].
- **True Negative Rate (Specificity)**: The ratio of correctly predicted negative observations to all actual negatives. It measures the proportion of actual negatives correctly identified as such [8].
- **False Negative Rate**: The ratio of incorrectly predicted negative observations to all actual positives. It measures the proportion of actual positives that were incorrectly classified as negatives [8].
- **Confusion Matrix**: A table often used to describe the performance of a classification model on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm [8].
- **ROC Curve**: A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings [8].

B. Evaluation

1) Support Vector Machines (SVM):

- Accuracy: 87.62%, indicating a strong overall performance.
- Sensitivity: 0.87 and Specificity: 0.88, suggesting effective identification of both biodegradable and non-biodegradable chemicals.

2) K-Nearest Neighbors (KNN):

- Optimal performance with k = 6, accuracy: 85.71%.
- Higher False Positive Rate (0.17) compared to SVM, indicating lower specificity (0.83).

3) Logistic Regression with Gradient Descent:

- Optimal performance with α = 1 and ϵ = 1e-8, accuracy: 79.52%, the lowest among the three models.
- High Sensitivity: 0.92, effective in identifying biodegradable chemicals but with a higher false positive rate (0.27).
- The ROC curve for the Logistic Regression with Gradient Descent model, presented in (Fig. 5) demonstrates a high True Positive Rate (TPR) across varying thresholds. With an Area Under Curve (AUC) of approximately 0.915, the model exhibits a strong discriminatory ability between the biodegradable and non-biodegradable classes. The curve's shape indicates a favorable balance between the TPR and the False Positive Rate (FPR), suggesting that the model is robust in distinguishing between the two outcomes with high confidence.

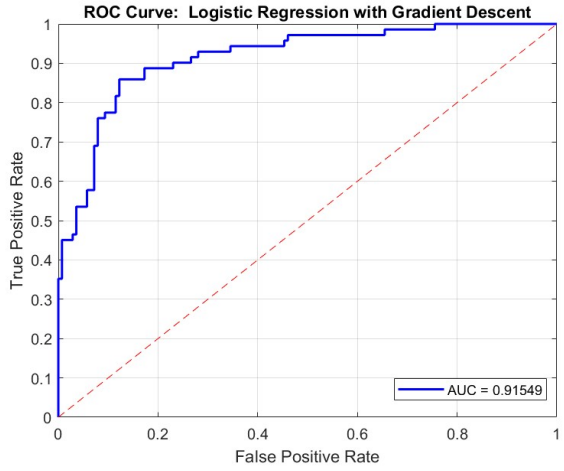

Fig. 5. The ROC Curve for the Logistic Regression with Gradient Descent Model

C. Comparison of Models

- SVM demonstrated superior balance in performance metrics, particularly excelling in precision (78%), which underscores its efficacy in minimizing false positive predictions. This characteristic makes SVM particularly suitable for applications where false alarms are costly. The confusion matrix (Fig. 4a) illustrates SVM's balanced approach towards both biodegradable and non-biodegradable classifications.
- Logistic Regression showed exceptional sensitivity (92%), indicating its strength in correctly identifying biodegradable chemicals. However, it trailed in precision (64%), suggesting a higher rate of false positives. In scenarios where missing a biodegradable chemical is unacceptable, Logistic Regression could be the preferred model, despite its higher false positive rate, as depicted in its confusion matrix (Fig. 4c).
- KNN, while slightly lagging behind SVM in overall accuracy, stands out for its computational simplicity and ease of implementation. It offers a viable alternative when computational resources are limited or for initial exploratory analysis. The confusion matrix for KNN (Fig. 4b) shows a relatively balanced performance but with slightly higher false negatives compared to SVM.

V. Conclusion and Recommendation

In conclusion, SVM emerged as the most robust model for this task however, it was a built-in function implementation. From the manually implemented functions, K-Nearest Neighbors performed better. It is important to emphasize that, the choice of the model should be tailored to the specific requirements of the application, considering factors like the importance of identifying biodegradable chemicals and computational constraints. The report concludes by comparing the advantages and disadvantages of the employed machine learning methods. Based on the analysis, a recommendation is made for the most suitable model for predicting chemical biodegradability. The recommendation is justified based on the models' performance metrics and their applicability to the QSAR dataset.

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.

[2] Chanin Nantasenamat, Chartchalerm Isarankura-Na-Ayudhya, Thanakorn Naenna, and Virapong Prachayasittikul, "A practical overview of quantitative structure-activity relationship," vol. 8, Jul. 2009, doi: https://doi.org/10.17877/de290r-690.

[3] G. Piir, I. Kahn, A. T. García-Sosa, S. Sild, P. Ahte, and U. Maran, "Best Practices for QSAR Model Reporting: Physical and Chemical Properties, Ecotoxicity, Environmental Fate, Human Health, and Toxicokinetics Endpoints," *Environmental Health Perspectives*, vol. 126, no. 12, Dec. 2018, doi: https://doi.org/10.1289/EHP3264.

[4] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," Decision Analytics Journal, vol. 3, p. 100071, Jun. 2022, doi: https://doi.org/10.1016/j.dajour.2022.100071.

[5] M. Jones. (2023). ACS6427 DATA MODELLING AND MACHINE INTELLIGENCE, *WEEK 6: FURTHER CLASSIFICATION* [PDF document].

[6] Gareth Michael James, D. Witten, T. J. Hastie, and R. Tibshirani, *An introduction to statistical learning : with applications in R*. New York: Springer, 2013.

[7] X. Zou, Y. Hu, Z. Tian and K. Shen, "Logistic Regression Model Optimization and Case Analysis," *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Dalian, China, 2019, pp. 135-139, doi: 10.1109/ICCSNT47585.2019.8962457.

[8] Ž. . Vujovic, "Classification Model Evaluation Metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, doi: https://doi.org/10.14569/ijacsa.2021.0120670.