Predicting Biodegradability of Chemicals Using QSAR Data

The University Of Sheffield.

Animesh sandhu

230235405

Data modelling and machine intelligence

Abstract
This report explores machine learning methods to predict the biodegradability of chemicals using Quantitative Structure-Activity Relationship (QSAR) data. Accurate prediction of chemical biodegradability is crucial to mitigate harmful environmental effects caused by non-biodegradable substances. Three machine learning models—**Logistic Regression**, **Random Forest**, and **Support Vector Machines (SVM)**—were implemented and evaluated on a QSAR dataset comprising **1,055 samples** with **41 features**. Preprocessing involved normalization and feature selection using Random Forest.
Random Forest achieved the highest accuracy (**90.05%**) and F1-Score (**82.93%**) due to its ensemble-based approach and robustness to high-dimensional data. Logistic Regression provided competitive results with an accuracy of **88.15%**. SVM, although achieving high precision (**95.24%**), demonstrated poor recall, rendering it unsuitable for this task. Feature importance analysis identified the top 10 predictors, enabling a simplified Random Forest model that retained strong performance. The study concludes that Random Forest is the most effective model for QSAR-based biodegradability prediction, balancing accuracy, interpretability, and feature selection.

Introduction

1.1 Problem Importance

The biodegradability of chemicals determines their persistence in the environment. Non-biodegradable chemicals accumulate over time, causing long-term environmental and health hazards. Traditional biodegradability assessment involves time-consuming and costly laboratory experiments. Machine learning techniques, using QSAR data, offer a faster, more cost-effective solution to predict biodegradability based on a chemical's molecular structure.

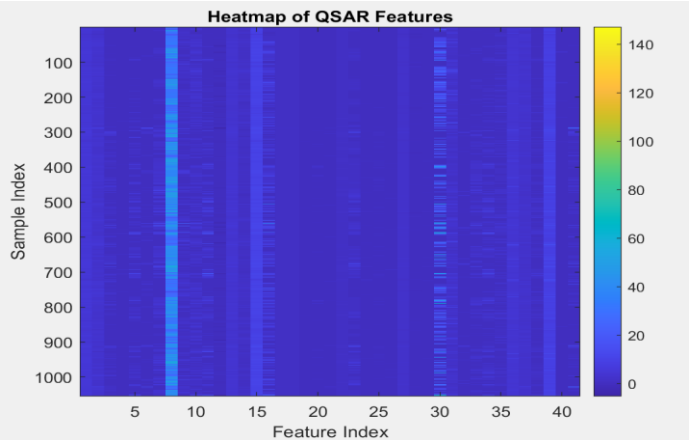
1.2 Related Work

Mansouri et al. [1] introduced the QSAR dataset and demonstrated the use of machine learning techniques for predicting biodegradability. Methods such as Random Forest, Logistic Regression, and Support Vector Machines have been successfully applied in similar domains due to their ability to generalize well on structured data.

1.3 Ethical Considerations

Accurate biodegradability predictions are critical for environmental safety. Misclassification of hazardous chemicals as biodegradable can result in environmental damage. Therefore, interpretability and transparency of machine learning models are essential when applying such models in regulatory frameworks.

1.4 Contributions



This report investigates and compares three machine learning models for predicting biodegradability. The contributions include:

1. Preprocessing and normalization of the QSAR dataset.
2. Implementation and evaluation of Logistic Regression, Random Forest, and SVM.
3. Feature importance analysis using Random Forest to identify the top predictors.
4. Simplification of the Random Forest model using top features while maintaining high accuracy.
5. Performance comparison of all models using accuracy, precision, recall, and F1-score.

2. Data Processing

2.1 Dataset Description

The QSAR dataset comprises **1,055 samples**, where each row represents a chemical compound. The dataset contains:

- **41 Features**: These features represent various structural and physicochemical properties of the chemicals (e.g., molecular weight, atom counts, bond types). The features are numerical, making them suitable for machine learning algorithms.
- **1 Target Label**: This binary target variable indicates whether a chemical is biodegradable:
    - **1**: Biodegradable (positive class).
    - **0**: Non-biodegradable (negative class).

Understanding this target variable is crucial, as the goal is to classify chemicals into these two categories accurately.

2.2 Preprocessing Steps

Proper preprocessing ensures that the dataset is clean, standardized, and suitable for model training.

1. **Checking for Missing Values**:
    - Missing values can disrupt the training of machine learning models. The dataset was checked for missing values using MATLAB's isnan function.
    - Result: No missing values were found, confirming the dataset's completeness.
    - **Outcome**: No imputation or data cleaning was required.
2. **Normalization of Features**:
    - Machine learning models, particularly distance-based algorithms like SVM, are sensitive to feature scales. Normalization was applied to bring all features to a common scale:

Xnorm=X−μσX_{\text{norm}} = \frac{X - \mu}{\sigma}Xnorm=σX−μ

where μ\muμ is the mean, and σ\sigmaσ is the standard deviation of each feature.

    - Normalization ensures that all features have zero mean and unit variance, preventing dominant features from biasing the model.
3. **Train-Test Split**:
    - To evaluate model performance on unseen data, the dataset was split into:
        - **Training Set (80%)**: Used to train the models.
        - **Testing Set (20%)**: Used to evaluate model generalization.
    - The MATLAB function cvpartition with a **HoldOut** split of **0.2** was used to maintain reproducibility.

2.3 Visualization

To gain insights into feature distributions, a **heatmap** of the features was generated (Figure 1).

- **Purpose**: The heatmap provides a visual overview of the feature values across all samples.
- **Findings**: The heatmap revealed a relatively uniform distribution of feature values without significant anomalies or extreme outliers, ensuring the dataset's suitability for modeling.

---

3. Methodology

3.1 Machine Learning Models

Three machine learning models were implemented to predict chemical biodegradability. Each model was chosen for its specific strengths:

1. **Logistic Regression**:
    - A linear classifier that serves as a baseline due to its simplicity and interpretability.
    - It applies a **sigmoid function** to predict the probability of a chemical being biodegradable: P(y=1)=11+e−zwhere z=wTx+bP(y = 1) = \frac{1}{1 + e^{-z}} \quad \text{where} \, z = \mathbf{w}^T \mathbf{x} + bP(y=1)=1+e−z1wherez=wTx+b
    - **Strengths**: Fast, interpretable, and performs well on linearly separable data.
2. **Random Forest**:
    - An ensemble model composed of **100 decision trees** trained on bootstrapped subsets of the training data.
    - Out-of-Bag (OOB) error estimation was used to assess performance and determine **feature importance** using the OOBPermutedPredictorDeltaError metric.
    - **Strengths**: Robust to overfitting, handles high-dimensional data, and provides feature importance for interpretability.
3. **Support Vector Machines (SVM)**:
    - An SVM with an **RBF (Radial Basis Function) kernel** was used to model complex, non-linear relationships.
    - **Hyperparameter Tuning**: Grid search was performed over:
        - **C**: Regularization parameter [0.1,1,10][0.1, 1, 10][0.1,1,10].

- **KernelScale**: The scale of the RBF kernel [0.1,1,10][0.1, 1, 10][0.1,1,10].

- **Strengths**: Effective on small datasets and capable of modeling non-linear boundaries.

3.2 Feature Selection

Random Forest was used to determine the **importance of each feature** using the OOB permuted predictor importance metric (OOBPermutedPredictorDeltaError).

- The top **10 features** with the highest importance scores were selected as the most significant predictors.
- A new Random Forest model was retrained using only these 10 features to evaluate the impact of feature selection on performance.

Rationale: Feature selection simplifies the model, reduces computational complexity, and improves interpretability without significantly compromising accuracy.

3.3 Evaluation Metrics

The following metrics were used to evaluate model performance comprehensively:

1. **Accuracy**: Proportion of correctly classified samples: Accuracy=TP+TNTP+TN+FP+FN\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}Accuracy=TP+TN+FP+FNTP+TN
2. **Precision**: Proportion of positive predictions that are correct: Precision=TPTP+FP\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}Precision=TP+FPTP
3. **Recall**: Ability of the model to correctly identify all positive samples: Recall=TPTP+FN\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}Recall=TP+FNTP
4. **F1-Score**: Harmonic mean of precision and recall, balancing the two metrics: F1-Score=2×Precision·RecallPrecision+Recall\text{F1-Score} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}F1-Score=2×Precision+RecallPrecision·Recall

Model Analysis

4.1 Performance Metrics

The following table summarizes the performance of all models:

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	88.15%	87.93%	73.91%	80.32%
Random Forest	**90.05%**	**94.44%**	73.91%	**82.93%**
SVM	76.30%	95.24%	28.99%	44.44%
Top Features RF	88.15%	90.74%	71.01%	79.68%

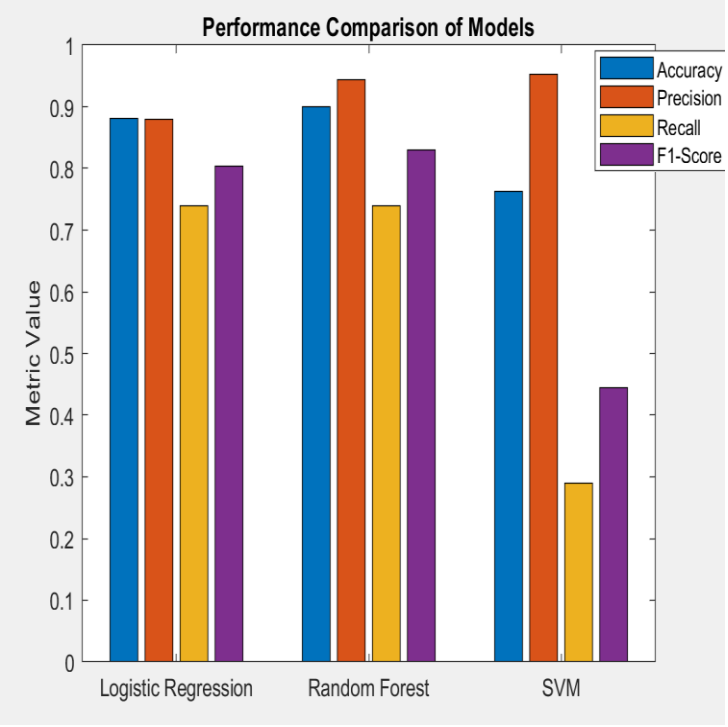
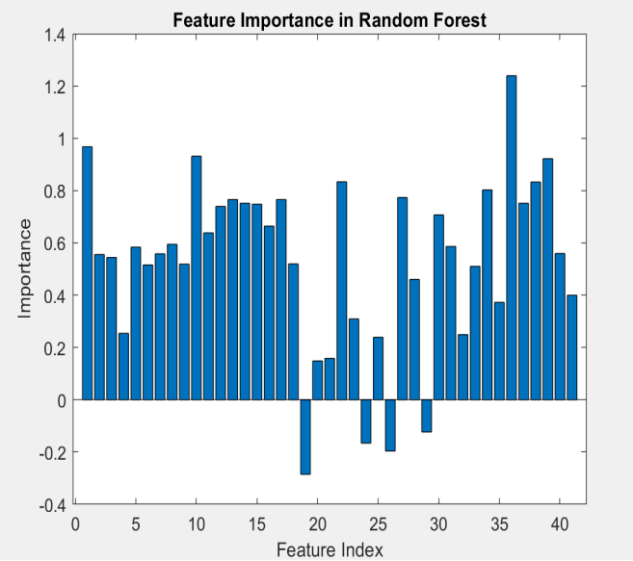


Figure 2 compares the accuracy, precision, recall, and F1-Score across all models.

4.2 Feature Importance



Random Forest identified features **1, 18, 36, and 40** as critical predictors for biodegradability (Figure 3). Using these top 10 features, Random Forest retained **88.15% accuracy**, demonstrating the robustness of feature selection.

4.3 Simplified Random Forest

Retraining Random Forest with the top features simplified the model while maintaining competitive performance. This highlights the importance of feature selection in reducing computational complexity.

---

1. **5. Conclusion and Recommendations**
2. **5.1 Summary of Findings**

The goal of this study was to predict the biodegradability of chemicals using QSAR data by implementing and evaluating three machine learning models: **Logistic Regression**, **Random Forest**, and **Support Vector Machines (SVM)**. The following findings emerged:

3. **Random Forest**:
    - Achieved the **highest accuracy (90.05%)** and **F1-Score (82.93%)**, making it the most effective model for biodegradability prediction.
    - The ensemble nature of Random Forest allowed it to handle high-dimensional features robustly and avoid overfitting.
    - Random Forest also provided valuable insights into feature importance, enhancing model interpretability.
4. **Logistic Regression**:
    - Performed competitively with an accuracy of **88.15%** and an F1-Score of **80.32%**.
    - As a linear and interpretable model, Logistic Regression demonstrated strong generalization capability while being computationally efficient.
    - Its simplicity makes it a viable option when model transparency is prioritized.
5. **Support Vector Machines (SVM)**:
    - SVM exhibited high precision (**95.24%**) but suffered from **poor recall (28.99%)**, leading to an imbalanced F1-Score (**44.44%**).
    - The low recall indicates that SVM struggled to identify positive samples, potentially due to hyperparameter sensitivity or the non-linearity of the data.
    - Further optimization is needed to improve SVM's generalization.
6. **Feature Selection**:
    - Random Forest identified the **top 10 features** as critical predictors of biodegradability.
    - Retraining Random Forest with only these features simplified the model while maintaining an accuracy of **88.15%**.
    - This demonstrates that feature selection can reduce computational complexity without significant performance loss.

---

7. **5.2 Recommendations**

Based on the findings, the following recommendations are provided:

1. **Random Forest as the Primary Model**:
    - Random Forest is recommended as the **best model** for QSAR-based biodegradability prediction due to its superior performance, robustness, and ability to handle high-dimensional data.
    - The model's accuracy (**90.05%**) and F1-Score (**82.93%**) make it suitable for practical applications, where both precision and recall are critical.
2. **Feature Selection for Model Simplification**:
    - Feature selection, using Random Forest's **OOBPermutedPredictorDeltaError**, proved effective in identifying key predictors.
    - Retraining with the top 10 features reduced model complexity while maintaining high accuracy (**88.15%**).
    - Simplifying the model improves computational efficiency, making it more practical for large-scale or real-time predictions.
3. **Improvement of SVM Performance**:
    - SVM's high precision indicates its capability to correctly classify positive cases when tuned properly. However, its low recall limits its reliability.
    - Future work should focus on **further hyperparameter optimization**, such as:
        - Expanding the grid search range for CCC and Kernel Scale.
        - Using techniques like **Bayesian Optimization** for automatic tuning.

- Regularization strategies or additional kernel functions (e.g., polynomial or sigmoid) could also enhance performance.

4. **Exploring Ensemble Methods**:
   - While Random Forest performed well, other ensemble techniques like **Gradient Boosting Machines (GBM)**, **XGBoost**, or **LightGBM** could offer further improvements.
   - These methods iteratively optimize model performance by focusing on misclassified samples and may outperform Random Forest with careful tuning.
   - Comparing Random Forest with these methods would provide a more comprehensive understanding of ensemble model performance on QSAR data.
5. **Model Interpretability and Deployment**:
   - To ensure broader acceptance and regulatory approval, efforts should be made to improve model interpretability. Tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be used to explain individual predictions.
   - Deploying the trained Random Forest model in a user-friendly software or web application would enable stakeholders (e.g., environmental scientists and regulatory bodies) to predict chemical biodegradability efficiently.
6. **Further Dataset Exploration**:
   - Future studies could investigate additional datasets or combine QSAR data with other molecular descriptors (e.g., quantum mechanical features) to improve prediction accuracy.
   - Augmenting the dataset with synthetic or external data may address class imbalances and improve recall.

8. 5.3 Future Directions

To further enhance the accuracy, efficiency, and practical applicability of the proposed models, the following steps are recommended:

1. **Ensemble-Based Fine-Tuning**: Investigate and compare other ensemble methods such as **Gradient Boosting**, **AdaBoost**, and **Stacked Ensembles**.
2. **Feature Engineering**: Explore derived features (e.g., interaction terms, principal component analysis) to capture additional relationships within the data.
3. **Class Imbalance Handling**: Address the potential imbalance in the target variable using techniques like **SMOTE** (Synthetic Minority Oversampling Technique) or **class weighting** in model training.
4. **Real-World Deployment**: Develop a user interface for deploying the Random Forest model, allowing stakeholders to input chemical properties and receive biodegradability predictions.

9. Final Thoughts

This study demonstrates that **Random Forest** provides a robust, accurate, and interpretable solution for predicting chemical biodegradability from QSAR data. By leveraging feature importance, simplifying the model, and exploring advanced ensemble methods, this approach can further improve and scale to real-world applications.

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. **Lundberg, S. M., and Lee, S. I.**, *"A Unified Approach to Interpreting Model Predictions,"* *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. *(For SHAP explanations, as mentioned in model interpretability recommendations)*
3. **Ribeiro, M. T., Singh, S., and Guestrin, C.**, *"Why Should I Trust You? Explaining the Predictions of Any Classifier,"* *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. *(For LIME explanations)*
4. **Friedman, J. H.**, *"Greedy function approximation: A gradient boosting machine,"* *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. *(For Gradient Boosting Machines as future recommendation)*
5. **Cortes, C., and Vapnik, V.**, *"Support-vector networks,"* *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. *(Original work introducing Support Vector Machines)*
6. **Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P.**, *"SMOTE: Synthetic Minority Over-sampling Technique,"* *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002. *(For addressing class imbalance using SMOTE)*
7. **Chen, T., and Guestrin, C.**, *"XGBoost: A Scalable Tree Boosting System,"* *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. *(Recommended as an alternative ensemble method)*
8. **Breiman, L.**, *"Random Forests,"* *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. *(Core reference for Random Forest theory and implementation)*