QSPR MODELING PROJECT

WINTER SEMESTER 2024-25

Shayan Hatami

January 22, 2025

ABSTRACT

This project focuses on designing an efficient workflow in KNIME to predict the pLC50 values of chemical compounds using molecular descriptors. The workflow integrates data preprocessing, outlier removal, feature selection, and machine learning model development. XGBoost was identified as the best-performing model, achieving an R^2 value of 0.92 on the test data, surpassing Random Forest. A separate testing workflow was developed to ensure scalability and applicability to new datasets. This report offers a detailed overview of the workflow and the methodology used.

1 Introduction

Accurately predicting pLC50 values, which represent the toxicity levels of chemical compounds, is crucial for environmental risk assessment and drug development. This project utilized cheminformatics tools within KNIME to develop a predictive model capable of generalizing to unseen datasets.

1.1 Project Objectives

- Preprocess and engineer features from molecular descriptor data.
- Handle outliers and normalize data to enhance model performance.
- Develop and validate a robust machine learning model.
- Create a modular testing workflow for new datasets.

2 Workflow Design

2.1 Training Workflow

2.1.1 Data Preparation and Exploration

Data Source: The dataset, provided in SDF format, was processed using RDKit nodes to extract molecular descriptors.
Target Analysis: The target column, pLC50, contained an extreme outlier, which was removed due to its significant impact on model performance.

2.1.2 Data Splitting

The dataset was split into 70% training and 30% testing. The test data was excluded from the training workflow to ensure unbiased evaluation.

2.1.3 Data Preprocessing

- **Variance Filtering:** Low-variance features were removed.
- **Outlier Handling:** The Closest Permitted Values Method replaced extreme values.
- **Feature Transformation:**
  - Box-Cox Transformation: Applied using a Python script to address skewness in feature distributions.
  - Two-Step Normalization:
    * Z-Score normalization for outlier management.
    * Min-Max scaling to standardize feature ranges.

2.1.4 Feature Selection

- **Correlation Filtering:** Features with correlations ¿0.9 were removed.
- **Backward Selection:** Using a loop with R^2 as the scoring metric and Random Forest, the most predictive features were identified. The final set consisted of 25 features.

2.1.5 Model Training

- **Cross-Validation:** 5-fold cross-validation was conducted using the X Partitioner and X Aggregator nodes.
- **Model Selection:** Both Random Forest and XGBoost models were trained. XGBoost showed superior performance and was saved using the Model Writer node.

2.2 Test data

After training the model, 30% of the data which was put away , was used for testing. After doing the same preprocessing on this test set, the result of R^2 value of 0.92, was achieved using the xgboost model.

2.3 Testing Workflow

To ensure the model's usability on new data, a separate testing workflow was developed inside the same file in a seperate node, for the undisclosed test data, with the following steps:

2.3.1 Data Preprocessing

Consistent preprocessing steps, including Box-Cox transformation, normalization, and feature selection, were applied to new datasets.

2.3.2 Model Application

The trained XGBoost model was loaded using the Model Reader node, and predictions were generated for the new data.

2.3.3 Output

The final output contained two columns: compound_id and pred_pLC50, which were saved as a CSV file.

3 Results and Observations

3.1 Model Performance

The XGBoost model achieved an R^2 value of 0.92 on the test data, indicating robust predictive performance.

3.2 Impact of Preprocessing

- Removing the single extreme outlier from the training dataset significantly improved model accuracy.
- The two-step normalization process and Box-Cox transformation enhanced data consistency.

3.3 Feature Selection

Using only 25 features optimized model performance while maintaining interpretability.

4 Discussion

- **Scalability:** The modular design ensures easy adaptation for new datasets.
- **Outlier Handling:** Outliers were addressed during training, but the testing workflow retains them to allow the model to generalize effectively.
- **Reproducibility:** Consistent preprocessing across workflows guarantees reliable results.

5 Conclusion

This project demonstrates the successful application of KNIME for predictive modeling of pLC50 values. Combining feature engineering, robust preprocessing, and model optimization resulted in a high-performing model. The creation of a separate testing workflow ensures the model's applicability to new datasets, making it a valuable tool for cheminformatics and toxicology studies.