High Level Design (HLD)

Prediction of LC50 value using quantitative structure activity relationship models (QSAR models)

Revision Number: 1.0

Last date of revision: 20 / 03/ 2024

Document Version Control

Date Issued	Version	Description	Author
20 / 03 / 2024	1	Initial HLD – V1.0	Rizal Muhammed

High Level Design (HLD)

Contents

Abstract

Thousands of chemical substances for which no ecological toxicity data are available can benefit from QSAR modelling to help prioritize testing. One of the data set encompassing in vivo test data on fish for hundreds of chemical substances using the ECOTOX database of the US Environmental Protection Agency, you can check that dataset through this link: ECOTOX Database and additional data from ECHA. We can utilize this to develop QSAR models that could forecast two sorts of end points: acute LC50 (median lethal concentration) and points of departure akin to the NOEC (no observed effect concentration) for any period (the “LC50” and “NOEC” models, respectively). Study factors, such as species and exposure route, were incorporated as features in these models to allow for the simultaneous use of many data types. To maximize generalizability to other species, a novel way of substituting taxonomic categories for species dummy variables were introduced.

The goal here is to build an end-to-end automated Machine Learning model that predicts the LC50 value, the concentration of a compound that causes 50% lethality of fish in a test batch over a duration of 96 hours (about 4 days), using 6 given molecular descriptors.

Prediction LC50 value using
QSAR models

1 Introduction

1.1 Why High-Level Design Document?

The purpose of this High-Level Design Document (HLD) is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

- Present all the design aspects and define them in detail
- Describe the user interface being implemented
- Describe the hardware and software interfaces
- Describe the performance requirements
- Include design features and the architecture of the project
- List and describe the non-functional attributes like
    - Reliability
    - Maintainability
    - Portability
    - Reusability
    - Application compatibility
    - Resource utilization
    - Serviceability

1.2 Scope

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system.

1.3 Definitions

LC 50	Lethal concentration 50 (LC50) is the amount of a substance suspended in the air required to kills 50% of a test animal during a predetermined observation period
Database	Collection of all the information monitored by this system
IDE	Integrated Development Environment

2 General Description

2.1 Product Perspective

Prediction of LC50 value using Quantitative Structure Activity Relationship Models (QSAR Models) is a Machine learning based regression model that will help us to predict the LC50 value, which will be helpful in Environmental Research.

2.2 Problem Statement

To create a Machine learning solution that predicts LC50 value, the concentration of a compound that causes 50% lethality of fish in a test batch over a duration of 96 hours (about 4 days), using 6 given molecular descriptors

2.3 Proposed Solution

The solution proposed here predict the LC 50 value based on 6 molecular descriptors: MLOGP (molecular properties), CIC0 (information indices), GATS1i (2D autocorrelations), NdssC (atom-type counts), NdsCH ((atom-type counts), SM1_Dz(Z) (2D matrix-based descriptors). The proposed solution can enhance environmental research.

2.5 Technical Requirements

2.6 Data Requirements

Data requirement is completely dependent on our problem statement.

Prediction LC50 value using
QSAR models

- We need more than 500 data instances for prediction of LC50 value with increased accuracy. The more data the better
- Missing values should be less than 10%
- Each data instance contains 6 features MLOGP (molecular properties), CIC0 (information indices), GATS1i (2D autocorrelations), NdssC (atom-type counts), NdsCH (atom-type counts), SM1_Dz(Z) (2D matrix-based descriptors)
- MLOGP (molecular properties) values typically fall within a range of –5 to 5. Negative values indicate that a molecule tends to be more soluble in polar solvents like water, suggesting higher hydrophilicity, On the other hand, positive values indicate that a molecule is more soluble in nonpolar solvents like octanol, suggesting higher lipophilicity.
- CIC0 refers to a specific molecular descriptor known as "Constitutional Information Content index 0." It's a numerical value used to describe the complexity or information content of a molecule based on its constitutional structure. Typically, CIC0 values are positive, and higher values indicate greater complexity or information content in the molecule's structure.
- Typically, GATS1i (2D autocorrelations) values are numerical and may be positive or negative, indicating different spatial correlations within the molecule. The GATS1i descriptor captures information about the spatial arrangement of atoms within a molecule, specifically considering the ionization potential of atoms as a weighting factor. It can provide insights into the electronic structure and chemical reactivity of molecules.

2.7 Tools Used

Python programming language and frameworks such as Flask, Numpy, Pandas, Matplotlib, Scikit-learn, etc. are used to build the whole model

- VS Code is used as IDE
- For visualization of the plots, Matplotlib and Seaborn are used
- MySQL is used to store and retrieve data
- Front end development is done using HTML/CSS
- Python, Flask are used for blackened development
- Git is used as version control system

2.7.1 Scikit-learn

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms, and is designed to interoperate with the Python numerical and scientific libraries Numpy and Scipy.

2.8 Constraints

The front end should be user-friendly. The prediction should have high accuracy and should be explainable, as users should not be required to know any of the workings of underlying machine learning models.

2.9 Assumptions

The main objective of the project is to solve the above mentioned problem statement for new data instance entered by the user. An end-to-end novel Machine learning regression model is used for predicting LC50 value based on user inputs for the following 6 molecular descriptors: MLOGP (molecular properties), CIC0 (information indices), GATS1i (2D autocorrelations), NdssC (atom-type counts), NdsCH ((atom-type counts), SM1_Dz(Z) (2D matrix-based descriptors).

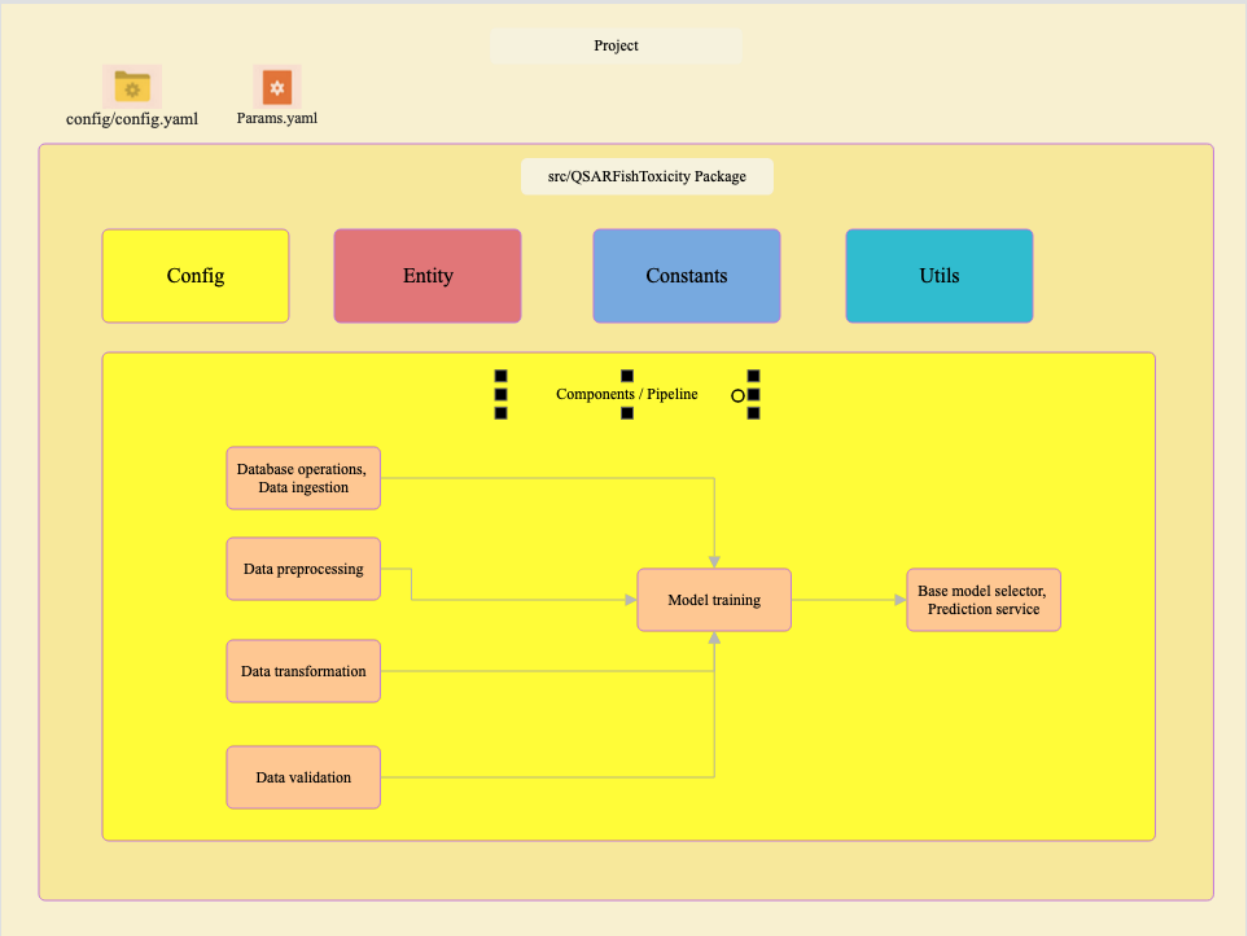
It is also assumed that all aspects of this project could work together in the way the designer is expecting.

3 Design Details

3.1 Workflow

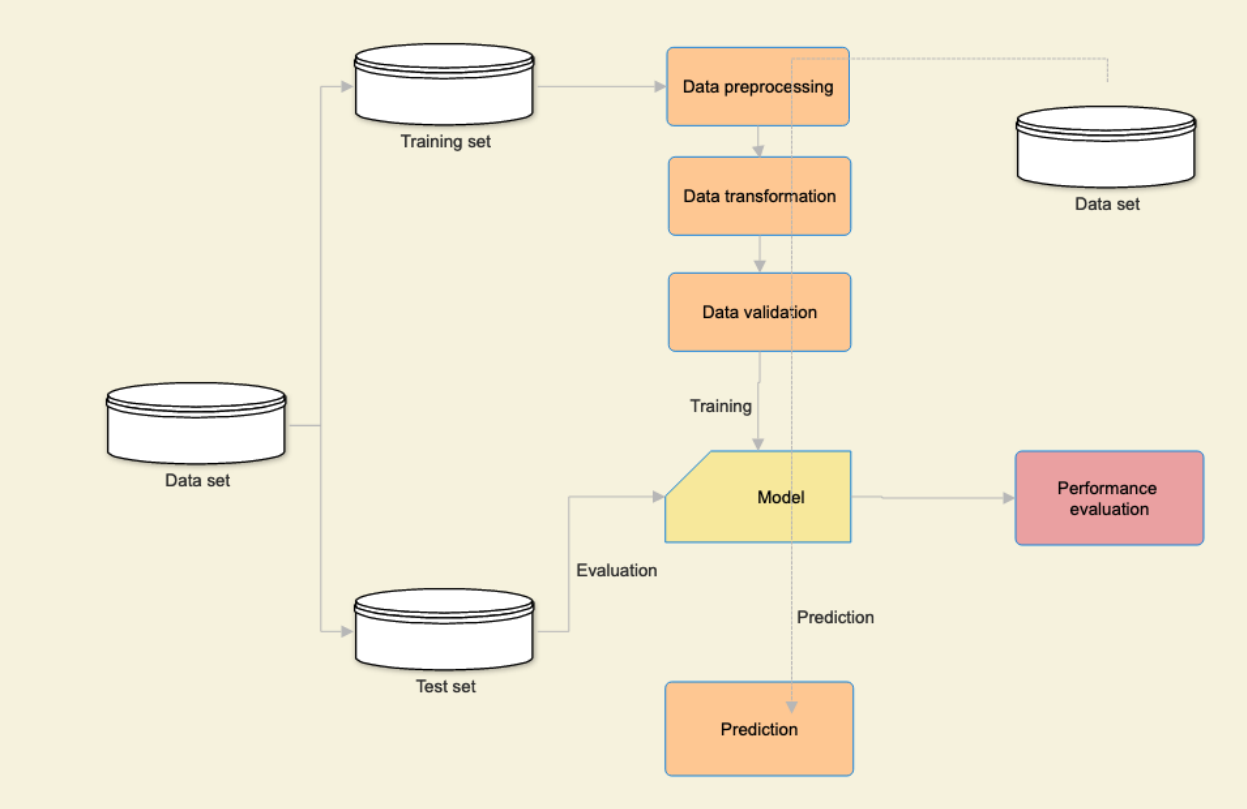
Process flow diagram for the Machine learning based model is as follows

High Level Design (HLD)



Prediction LC50 value using
QSAR models

3.1.1 Model Training and Evaluation



3.3 Event log

The system should log every even so that the user will know what process is running internally

Initial step by step description

1. The system identifies at what stage logging required
2. The system should be able to log each system flow
3. Developers can choose logging methods. Database logging or File logging
4. Logging is mandatory because, we can easily debug issues

3.3 Exception Handling

In encounter of exceptions, an explanation will be displayed as to what went wrong? An exception is defined as anything that falls outside the normal and intended usage.

4 Performance

The proposed solution prediction LC50 value, the concentration of a compound that causes 50% lethality of fish in a test batch over a duration of 96 hours (about 4 days), using 6 given molecular descriptors which will be referenced for environmental research. Therefore, the prediction should be as accurate as possible, to ensure that the researchers are not leading to erroneous conclusions.

4.1 Reusability

The code should be reusable

4.2 Scalability

The proposed solution should be scalable to incorporate future requirements

4.3 Application Compatibility

The different components of this project will be using Python as an interface between them. Each component will have its own task to perform, and it is the job of the Python to ensure proper transfer of information.

4.4 Resource Utilization

When training of the proposed solution is performed, it will likely use all the processing power available until the training is completed

5 Dashboards

Dashboards will be used to display and indicate certain KPIs and relevant indicators for the unveiled problems that if not addressed in time could cause catastrophes of unimaginable impact.

As and when the system starts to capture the historical / periodic data for a user, the dashboards will be included to display charts over time with progress on various indicators or factors.

5.1 KPIs (Key Performance Indicators)

1. Key indicators displaying a summary of the LC50 value prediction
2. Taking adequate evidence and action on environmental monitoring and management: LC50 values can serve as benchmarks for monitoring environmental quality and assessing the effectiveness of pollution control measures. Monitoring changes in LC50 values over time can help detect trends in environmental contamination and evaluate the success of remediation efforts in reducing chemical toxicity.
3. **Summary of safer chemicals developed utilizing prediction model**: Predictive models for LC50 values can guide the design and development of safer chemicals with reduced environmental impact. By understanding the structure-activity relationships associated with chemical toxicity, researchers can identify chemical properties that contribute to toxicity and use this information to design less harmful alternatives.

6 Conclusion

The Machine learning regression model will predict LC50 value, the concentration of a compound that causes 50% lethality of fish in a test batch over a duration of 96 hours (about 4 days), using 6 given molecular descriptors. This can be helpful in environmental research for assessment of chemical toxicity, environmental risk assessment, development of safer chemical alternatives, environmental monitoring and management, etc.