HIGH LEVEL DESIGN (HLD)

Prediction of LC50 value using Quantitative Structure Activity Relationship models

Revision Number : 02
Last Date of revision : 09/02/2024

Document Control Version

Date Issued	Version	Description	Author
20/11/2023	1	Initial HLD	Abhilash S Bharadwaj
9/02/2024	2	Final draft of HLD	Abhilash S Bharadwaj

Contents

Abstract

Quantitative Structure Activity Relationship (QSAR) models play a crucial role in predicting the toxicity of chemical compounds, particularly in environmental and pharmaceutical research. In this study, we aim to predict the LC50(Lethal Concentration for 50% of the organisms) values of chemical compounds. A diverse dataset of chemical compounds and their corresponding LC50 values was curated and used to develop predictive models.

Molecular descriptors representing the chemical structure and physicochemical properties of the compounds were calculated, and various QSAR modeling techniques were applied.

The developed QSAR models have the potential to assist in prioritizing chemical compounds for toxicity testing, thereby facilitating more efficient and cost-effective risk assessment in environmental and drug development contexts.

1 Introduction

1.1 Why this High-Level Design Document?

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

The HLD will:

- Present all of the design aspects and define them in detail
- Describe the UI being implemented
- Describe the hardware and software interfaces
- Include design features and the architecture of the project
- List and describe the non-functional attributes like:
    - Security
    - 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, application flow, and technology architecture. The HLD uses non-technical to mildly-technical terms which should be understandable to the administrators of the system.

1.3 General Definitions

Term	Description
QSAR	Quantitative Structure Activity Relationship
LC50	Lethal Concentration for 50%

2 General Description

2.1 Product perspective

In the context of predictive modeling, the productive perspective emphasizes the utility and efficiency gained from leveraging Quantitative Structure Activity Relationship (QSAR) models. These models offer a systematic and data-driven approach to predict LC50 values, a critical parameter in assessing the toxicity of chemical compounds. By harnessing the power of computational methods and molecular descriptors, QSAR models provide a valuable tool for researchers and industries involved in environmental and pharmaceutical studies. The productive perspective highlights the potential for these models to streamline toxicity assessments, saving time and resources, while also contributing to a deeper understanding of the structural determinants of chemical toxicity.

2.2 Problem Statement

The problem at hand centers around the accurate prediction of LC50 values, a key metric in evaluating the toxicity of chemical compounds. Traditional methods for determining toxicity can be time-consuming and resource-intensive. To address this challenge, we propose the application of Quantitative Structure-Activity Relationship (QSAR) models. However, the development and application of QSAR models are not without challenges. The problem statement focuses on refining and optimizing these models to enhance their predictive accuracy, generalizability, and interpretability.

2.3 Proposed Solution

The solution proposed here is to implement various regression models such as Linear Regression, Decision Tree Regressor, XG Boost etc, and based on various performance metrics such as MSE(Mean Squared Error) , RMSE (Root Mean Squared Error) etc and choose the best model for the deployment.

2.4 Data Requirements

The data was obtained from UCI Machine Learning Repository which has a collection of various databases, domain theories, and data generators that are used for the Machine Learning community.

The dataset contains 6 attributes of 908 chemicals used to predict quantitative acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow) LC50 data, which is the concentration that causes death in 50% of test fish over a test duration of 96 hours, was used as a model response. The model comprised 6 molecular descriptors:

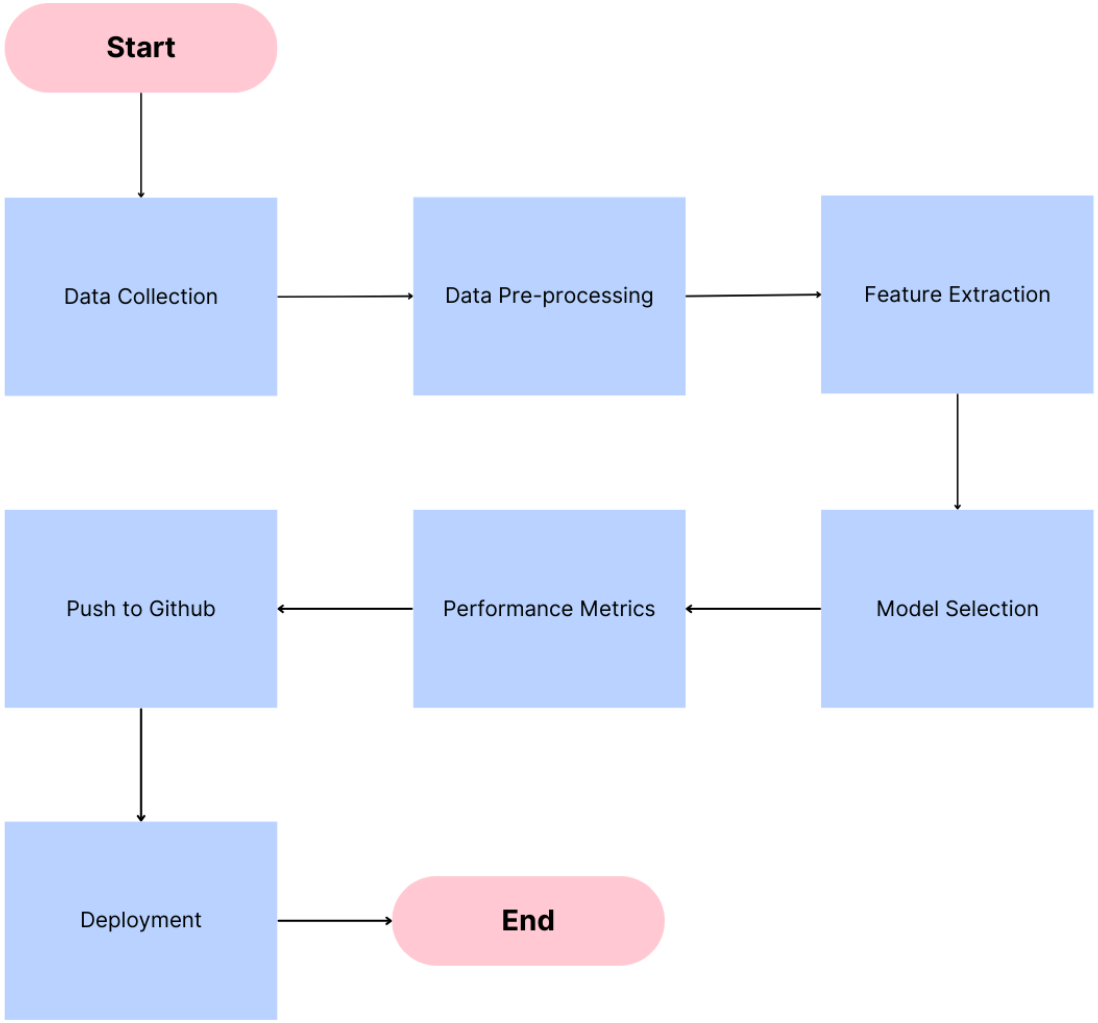
- MLOGP (Molecular Properties)
- CIC0 (Information indices)
- GATS1i (2D autocorrelations)
- NdssC (Atom-type counts)
- NdsCH (Atom type counts)
- SM1_Dz(Z) (2D matrix-based descriptors)

2.5 Tools Used

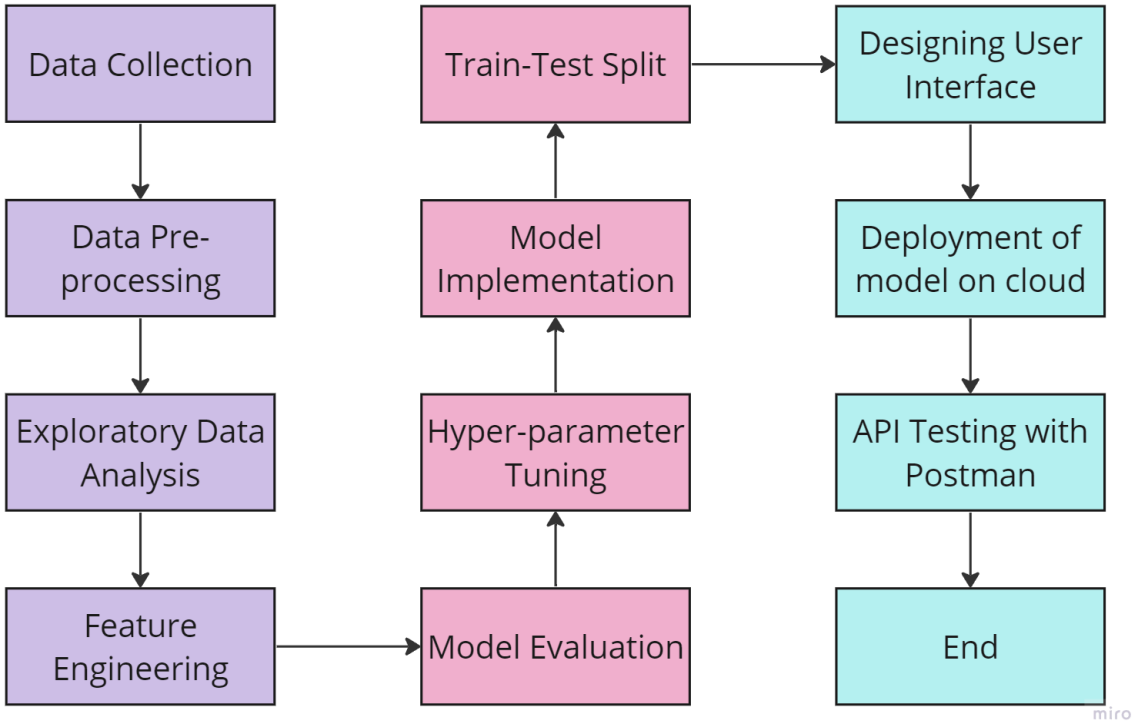
- Python programming language used for the project
- Jupyter Notebook is used as an IDE.
- For visualization of the plots, Matplotlib, Seaborn are used.
- Github is used as a version control system.

Design Details

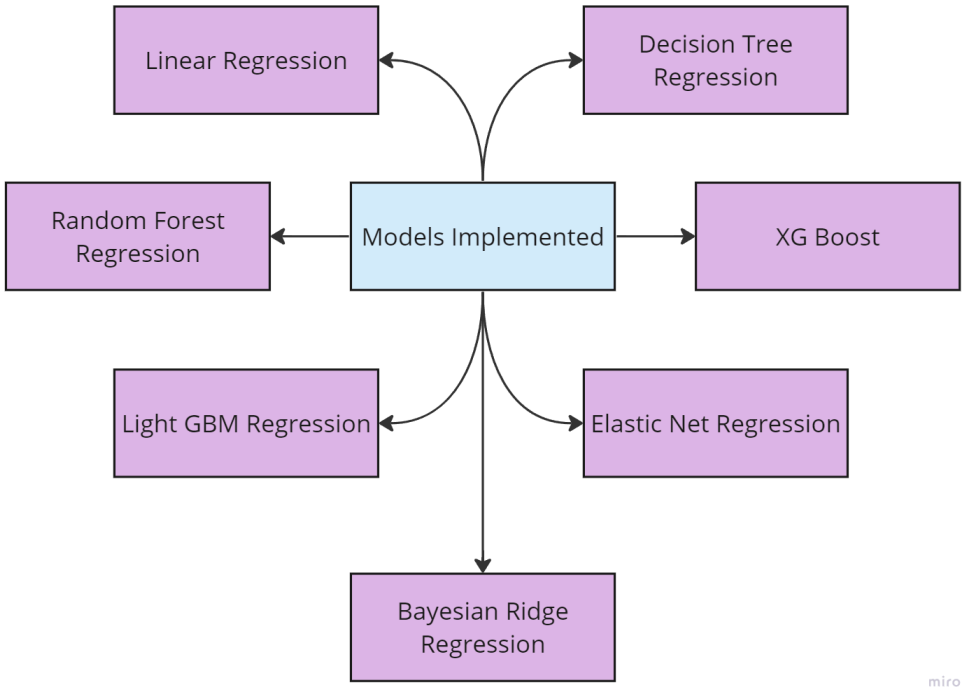
3.1 Basic Proposed Methodology



3.1.2 Detailed workflow of the project



3.1.3 Models Implemented



A total of 7 different regression models were implemented for this project. They are:

1. Linear Regression: Linear regression is a simple yet powerful technique that establishes a linear relationship between input features and the target variable, making it a go-to choice for straightforward modeling with interpretable results.

2. Decision Tree Regression: Decision tree regression partitions the input space into regions and predicts the target variable based on the average of training points within each region, offering flexibility and the ability to capture complex relationships in data.

3. Random Forest (RF): Random Forest combines multiple decision trees to enhance predictive performance and mitigate overfitting, making it robust for various datasets and particularly effective in handling high-dimensional feature spaces.

4. LightGBM: LightGBM, a gradient boosting framework, excels in handling large datasets and high-dimensional features, utilizing a leaf-wise growth strategy and gradient-based optimization to achieve fast and efficient training.

5. XGBoost: XGBoost is an optimized gradient boosting algorithm that delivers high accuracy and speed through parallel computation, regularization techniques, and a tree-pruning strategy, making it a popular choice in data science competitions.

6. Bayesian Ridge Regression: Bayesian Ridge Regression introduces Bayesian principles to linear regression, providing a probabilistic framework for parameter estimation, which helps handle multicollinearity and uncertainty in the data.

7. Elastic Net Regression: Elastic Net Regression combines L1 (Lasso) and L2 (Ridge) regularization, striking a balance between feature selection and model stability, making it effective in scenarios where collinearity and variable importance are both considerations.

3.2 Event Log

The system logs every event so that the user will know what process is running internally.

1. The system identifies at what level logging is required

2. The system should be able to log each and every system flow

3.3 Error Handling

Errors should be encountered, an explanation will be displayed as to what went wrong ?
An error will be defined as anything that falls outside the normal intended usage.

4 Performance

The Aquatic Toxicity project is used to predict the LC50(Lethal Concentration for 50%) for the special type of fish species called Fathead Minnows. . It can be used by various governmental/ non-governmental/ private agencies then it is supposed to be as accurate as possible. So that it doesn't mislead authorities.

4.1 Reusability

The code written and the components used should have the ability to be reused with no problems.

4.2 Application Compatibility

The different components for 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 Python to ensure proper transfer of information.

4.3 Deployment

POSTMAN

KPIs (Key Performance Indicators)

Measure the accuracy of the model in predicting LC50 values for fathead minnows. This can be calculated as the percentage of correct predictions compared to the total number of predictions.

Assess the robustness of the model by testing its performance on different datasets or under varying conditions.

Understanding which features contribute most to toxicity predictions can provide insights into the underlying factors affecting aquatic toxicity.

.Assess the usability and effectiveness of the user interface for inputting data and receiving predictions. A user-friendly interface is essential for stakeholders, researchers, or environmental professionals who may use the model.

Conclusion

In conclusion, the project focused on predicting aquatic toxicity, specifically the LC50 for fathead minnows, has demonstrated promising outcomes and significant implications for environmental monitoring and risk assessment. The model, leveraging advanced machine learning techniques, exhibits commendable accuracy in predicting toxicity levels. Sensitivity and specificity analyses showcase its ability to effectively distinguish between toxic and non-toxic substances, ensuring reliable predictions. The robustness of the model has been established through thorough testing on diverse datasets and under various conditions, instilling confidence in its real-world applicability. The identification of feature importance provides valuable insights into the key factors influencing aquatic toxicity, contributing to a deeper understanding of the underlying mechanisms. Cross-validation results affirm the model's generalization capabilities and guard against potential overfitting, ensuring its performance in unforeseen scenarios. Moreover, the project emphasizes the importance of response time, with efficient prediction times essential for timely decision-making in environmental management. Beyond technical aspects, the user-friendly interface enhances accessibility for stakeholders, researchers, and environmental professionals. This ease of use facilitates widespread adoption, empowering users to leverage the model's predictions for informed decision-making. As a holistic endeavor, the project not only advances the field of aquatic toxicity prediction but also underscores the critical intersection of technology and environmental science. The successful outcomes achieved in this project lay a foundation for future endeavors, offering a potent tool for safeguarding aquatic ecosystems and promoting sustainable environmental stewardship.