Architecture Design Document

Prediction of LC50 value Using QSAR Models

Document Version Control

Date Issued	Version	Description	Author
21/03/2024	1.0	First Draft	Rizal Muhammed

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 sets 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.

1 Introduction

1.1 Why this Architecture Design Document?

The purpose of this document is to present a detailed description of Prediction of LC50 value using QSAR Models system. It will explain the purpose and features of the system, the interface of the system, what the system will do, the constraints under which it will operate, etc. This document is intended for both stake holders and the developers of the system and will be proposed to the higher management for its approval.

The objective of the project is to predict LC50 value, which is the concentration that causes death in 50% of test fish over a test duration of 96 hours (about 4 days), using Quantitative Structure Activity Relationship Models (QSAR Models) based on 6 attributes (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 project shall be delivered in two phases

Phase 1: Backend functionalities like database operations, model building, etc.

Phase 2: Integration UI to all the functionalities

1.2 Scope

This software system will be a web application. This system will be designed to detect LC50 value using QSAR models based on 6 attributes (molecular descriptors) as mentioned above. QSAR models offer a cost-effective alternative to traditional experimental methods for toxicity testing, QSAR models using computational models to predict toxicity contribute to the reduction of animal testing by minimizing the number of live animals required for toxicity testing. QSAR models provide insight into the chemical properties and structural features of compounds that contribute to their toxicity. Understanding these relationships can guide the design of safer chemicals and drugs by allowing researchers to modify molecular structures to reduce toxicity while maintaining desired biological activity. Also, prediction of LC50 value can enhance environmental research.

1.3 Constraints

We will be only considering 6 molecular descriptors for LC50 value prediction model.

1.4 Risks

QSAR models are typically developed for specific chemical classes or endpoints, and their performance may degrade when applied to compounds outside the domain of applicability. Using QSAR models to predict LC50 values for chemicals that differ significantly from those in the training set can lead to erroneous predictions.

While QSAR models can be valuable tools for toxicity prediction, regulatory acceptance of QSAR predictions for safety assessment purposes may vary depending on the jurisdiction and the specific regulatory agency.

1.5 Out of Scope

Mechanistic insights: QSAR models are primarily concerned with correlating chemical structures with biological activity or toxicity, but they typically do not provide mechanistic insights into how a compound interacts with biological systems at a molecular level. Understanding the underlying mechanisms of toxicity often requires additional experimental and computational studies beyond the scope of QSAR modeling.

Non-chemical factors: QSAR models focus on chemical structure-activity relationships and may not explicitly account for non-chemical factors that can influence toxicity, such as environmental conditions, organism-specific factors, and metabolic pathways.

2 Technical Specifications

2.1 Dataset

Source: The dataset is available at UCI Machine Learning Repository. Please refer here.

Additional Variable Information

6 molecular descriptors and 1 quantitative experimental response:

1. CIC0
2. SM1_Dz(Z)
3. GATS1i
4. NdsCH
5. NdssC
6. MLOGP
7. quantitative response, LC50 [-LOG(mol/L)]

Data set containing values for 6 attributes (molecular descriptors) of 908 chemicals used to predict quantitative acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow).

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Physics and Chemistry	Regression

Feature Type	# Instances	# Features
Real	908	6

Feature Name	Data type	Required?
C1C0	Float	Yes
SM1_Dz	Float	Yes
GATS1i	Float	Yes
NdsCH	Int	Yes
NdssC	Int	Yes
MLOGP	Float	Yes

2.2 Predicting LC50 value

- The system displays fields for inputting 6 molecular descriptors mentioned above.

- The system should be able to predict the corresponding LC50 value based on the user inputs

2.3 Logging

We should log every activity.

- The system identifies at what step logging is required
- The system should be able to log every system flow.
- Developers can choose logging methods. Database logging / File logging
- Logging is mandatory since it helps to resolve issues

2.4 Database

The system stores the data received after data ingestion stage into a database. MySQL database was chosen for this purpose.

2.5 Deployment

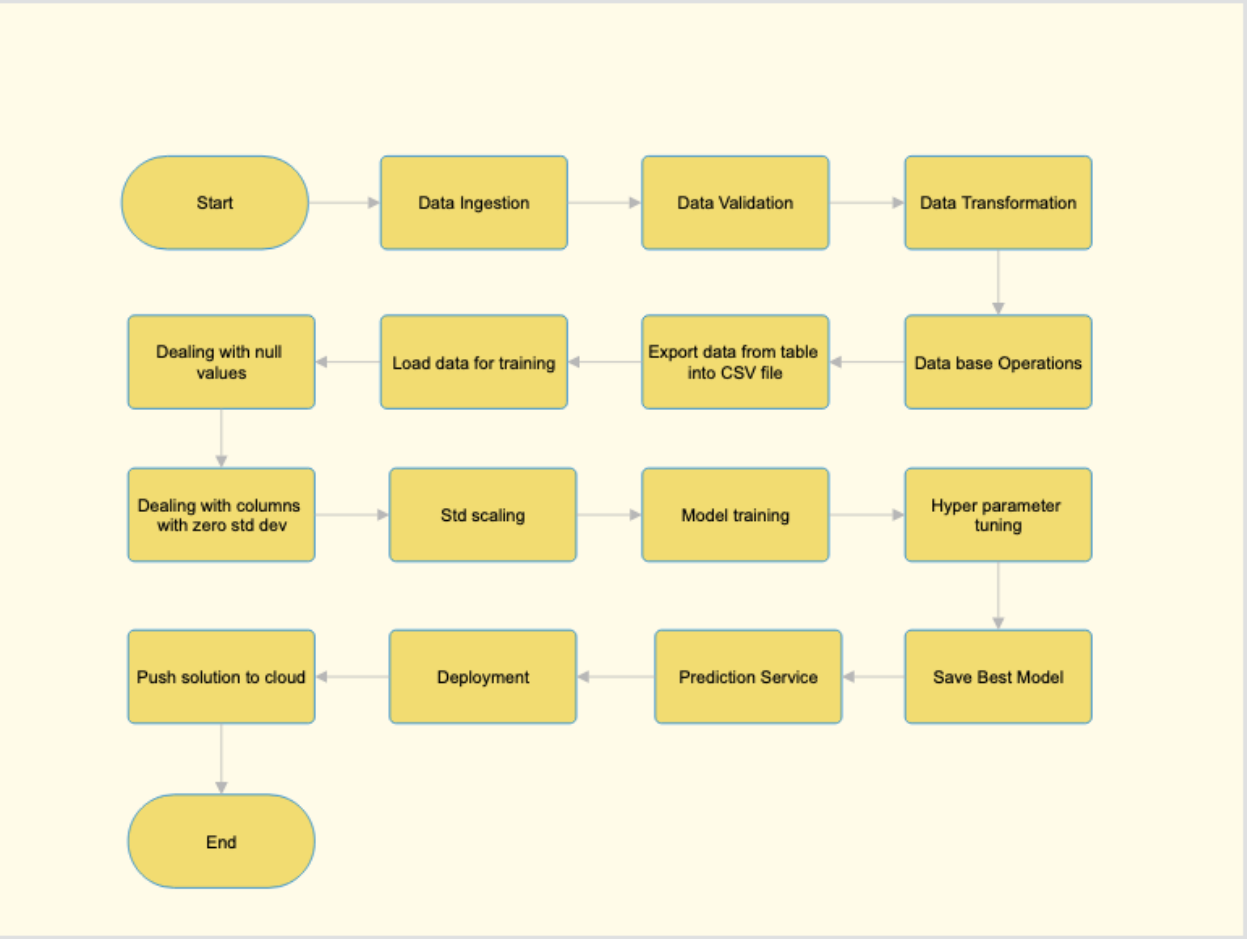
3 Technology Stack

Front end	HTML, CSS
Backend	Python, Flask, Sci-kit learn
Database	MySQL
Deployment	Heroku

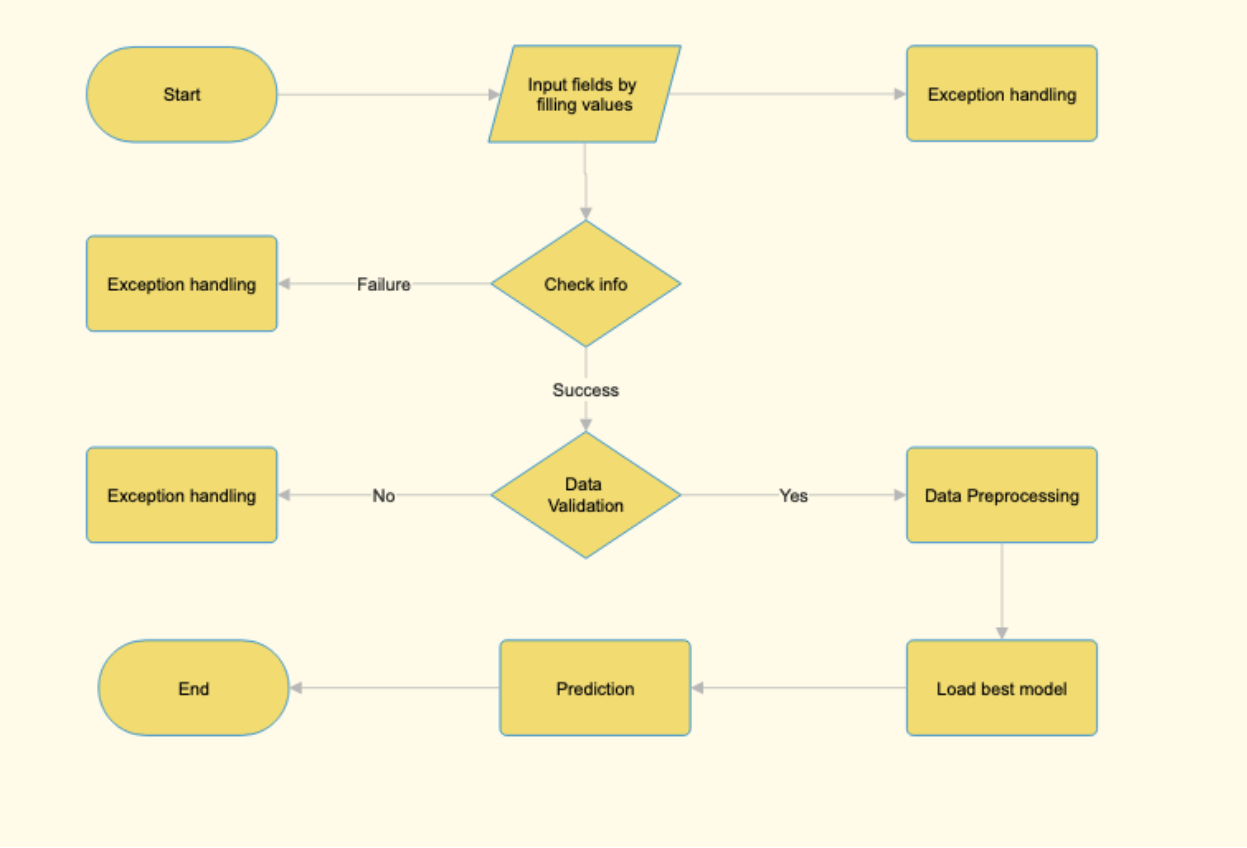
4 Proposed Solution

The prediction of LC50 value using 6 molecular descriptors is a regression problem that can be solved with machine learning algorithms. We'll try different machine learning algorithms like linear regression, support vector machine, random forest. Based on evaluation, the best model along with corresponding metrics will be stored for prediction and further reference respectively.

5 Model Training / Validation Workflow



6 User I/O Workflow



7 Test Cases

Test Case Description	Pre-Requisite	Expected Result
Verify whether the application URL is accessible to the user	1. Application URL should be defined	Application URL should be accessible to the user
Verify whether the application loads completely for the user when the URL is accessed	1.Application URL is accessible 2. Application is deployed	Application should load completely for the user when the URL is accessed
Verify whether user can see input fields on accessing the URL	1. Application is accessible	User should be able to successfully see input fields
Verify whether user can edit all input fields	1. Application is accessible 2. All the input fields are visible	User should be able to edit all input fields
Verify whether user gets Predict button to submit the inputs	1. Application is accessible 2. All the input fields are visible 3. User can edit input fields	User should be able access Predict button
Verify whether the corresponding prediction is displayed	1. Application is accessible 2. All the input fields are visible 3. User can edit input fields 4. User can submit input	User should be able to get the prediction displayed

8 Key Performance Indicators (KPI)s

- Key indicators displaying a summary of the LC50 value prediction
- 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.
- **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