Low Level Design (LLD) Document

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

Document Version Control

Date Issued Version Description Author

25/08/2022	1.0	Introduction, General Description	Darshan
25/08/2022	1.1	Dataset Information, Architecture	Darshan
13/09/2022	1.2	Final Revision	Darshan

Contents

Introduction

What is Low Level Design Document (LLD)?

The Low-level design specifics the detailed description of all modules, which implies that the LLD involves all the system component's actual logic. The goal of the Low-level design document (LLDD) is to give the internal logic design of the actual program code .

What is the purpose?

A good low-level design document makes the program easy to develop when proper analysis is utilized to create a low-level design document. The code can then be developed directly from the low-level design document with minimal debugging and testing. Other advantages include lower cost and easier maintenance. The goal of LLD or a low-level design document (LLDD) is to give the internal logical design of the actual program code. Low-level design is created based on the high-level design.

Project Introduction

Thousands of chemical substances for which no ecological toxicity data are available can benefit from QSAR modelling to help prioritise 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 utilise 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 maximise generalizability to other species, a novel way of substituting taxonomic categories for species dummy variables was introduced.

In our industrialized society, a huge amount of chemical substances are used and produced every day. This increasing number of chemicals around us raises the problem of characterization, prediction and evaluation of their consequences to the human health and to the environment.

Toxicology provides the knowledge of mechanisms, rules, and data characterized by quality levels and defines the limits of safety of chemical agents. Chemistry offers knowledge on chemical descriptors and physico-chemical properties. Biology studies the mechanisms of the action of the chemicals on the animals and the other organisms used for tests.

Problem Statement

Prediction of LC50 value using Quantitative structure–activity relationship models (QSAR models). 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, using 6 given molecular descriptors.

Dataset Information

This dataset was used to develop quantitative regression QSAR models to predict acute aquatic toxicity towards the fish Pimephales promelas (fathead minnow) on a set of 908 chemicals. LC50 data, which is the concentration that causes death in 50% of test fish over a test duration of 96 hours, was used as 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).

The model comprised 6 molecular descriptors and 1 quantitative experimental response :

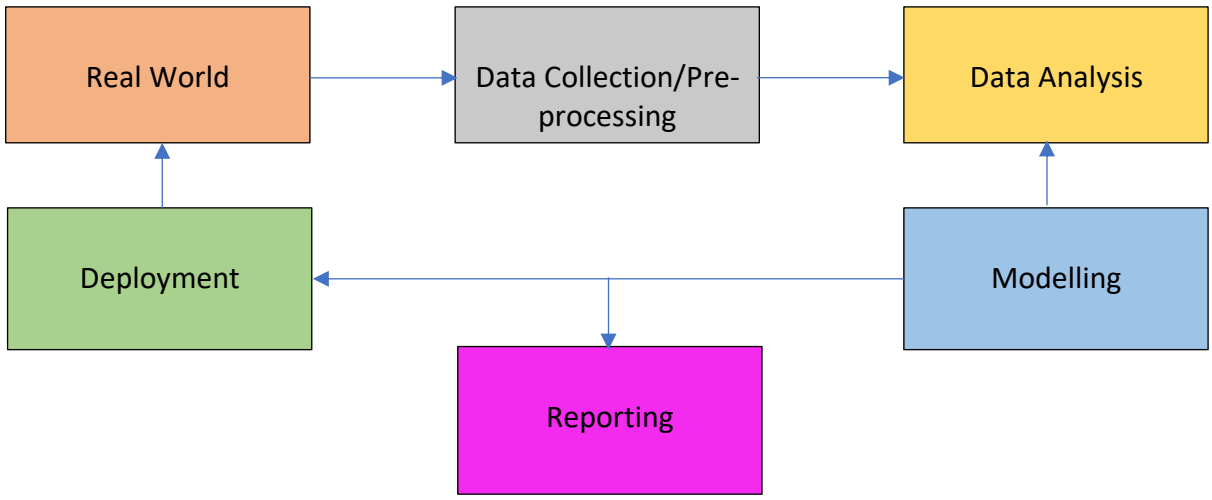
- MLOGP (molecular properties),
- CIC0 (information indices),
- GATS1i (2D autocorrelations),
- NdssC (atom-type counts),
- NdsCH ((atom-type counts),
- SM1_Dz(Z) (2D matrix-based descriptors).
- quantitative response, LC50 [-LOG(mol/L)]

Details can be found in the quoted reference: M. Cassotti, D. Ballabio, R. Todeschini, V. Consonni. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (Pimephales promelas), SAR and QSAR in Environmental Research (2015), 26, 217-243; doi: 10.1080/1062936X.2015.10189380

Data Set Characteristics:	Multivariate	**Number of Instances:**	908	**Area:**	Physical
Attribute Characteristics:	Real	**Number of Attributes:**	7	**Date Donated**	2019-09-23
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	46489

Source : UCI Machine Learning Repository

Architecture



Architecture Description

Raw Data Collection

The Dataset was taken from iNeuron's Provided Project Description Document.

Link

Data Pre-Processing

Before building any model, it is crucial to perform data pre-processing to feed the correct data to the model to learn and predict. Model performance depends on the quality of data feeded to the model to train.
This Process includes-

1. Handling Null/Missing Values
2. Handling Skewed Data
3. Outliers Detection and Removal
4. Remove duplicate or irrelevant observations
5. Filter unwanted outliers
6. Renaming required attributes

Exploratory Data Analysis (EDA)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Reporting

Reporting is a most important and underrated skill of a data analytics field. Because being a Data Analyst you should be good in easy and self- explanatory report because your model will be used by many stakeholders who are not from technical background.

1. a) High Level Design Document (HLD)
2. b) Low Level Design Document (LLD)
3. c) Architecture
4. d) Wireframe
5. e) Detailed Project Report
6. f) Power Point Presentation

Modelling

Data Modelling is the process of analysing the data objects and their relationship to the other objects. It is used to analyse the data requirements that are required for the business processes. The data models are created for the data to be stored in a database. The Data Model's main focus is on what data is needed and how we have to organize data rather than what operations we have to perform.