High Level Design (HLD)
Prediction of LC50 value

Revision no.: 0
Date : 2/10/2021

Document version control

Date Issued	Version	Description	Author
2/10/2021	1.0	LC50 prediction	Ramesh RVS

Scope:

This HLD document presents structure of the system , such as database architecture, application flow, deployment process. The HLD uses non technical to mildly technical terms , which should be understandable to the administrators of the system.

Definitions:

LC50	Lethal content 50%
Database	Collection of values arrived during testing
IDE	Integrated development environment
AWS	Amazon web servers

Data Set 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).

Attribute 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)]

Problem Statement:

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.
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.

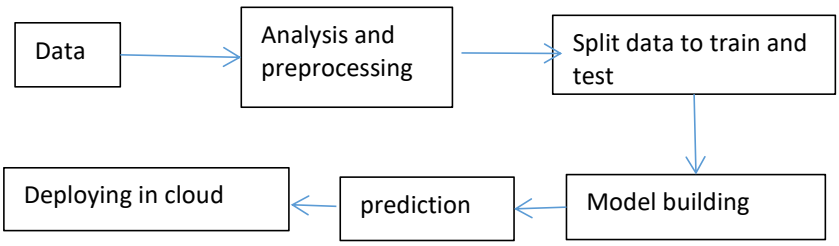
Proposed solution:

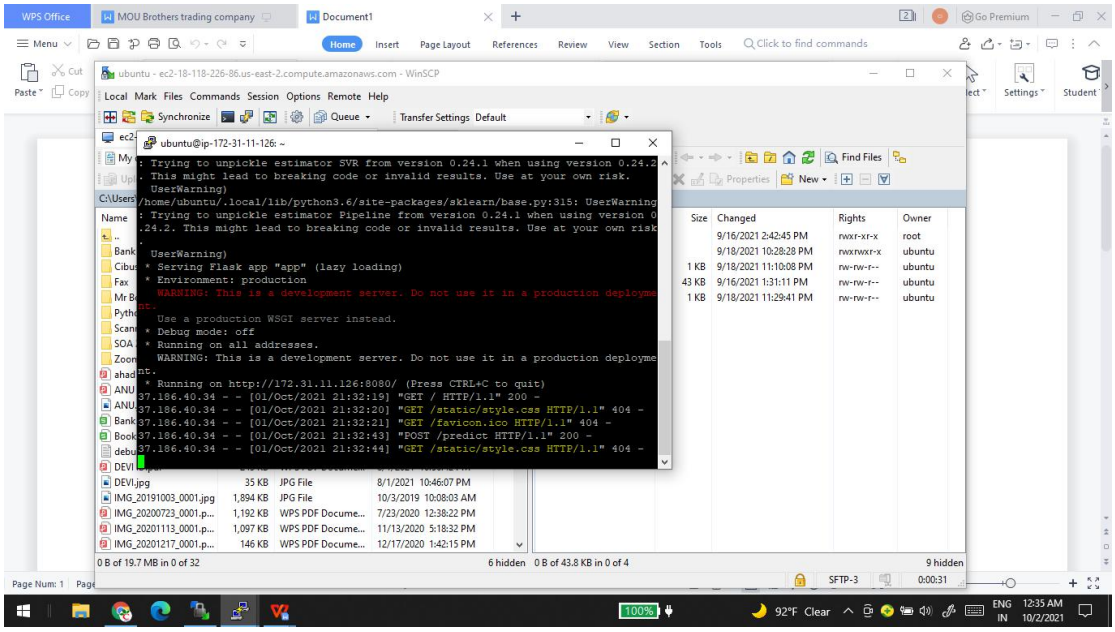
- Analyse the data
- Preprocessing of data
- Fit the data as training and testing to models.
- Compare models with accuracy
- Choose the best model
- Tuning the model
- Deploy the model in cloud

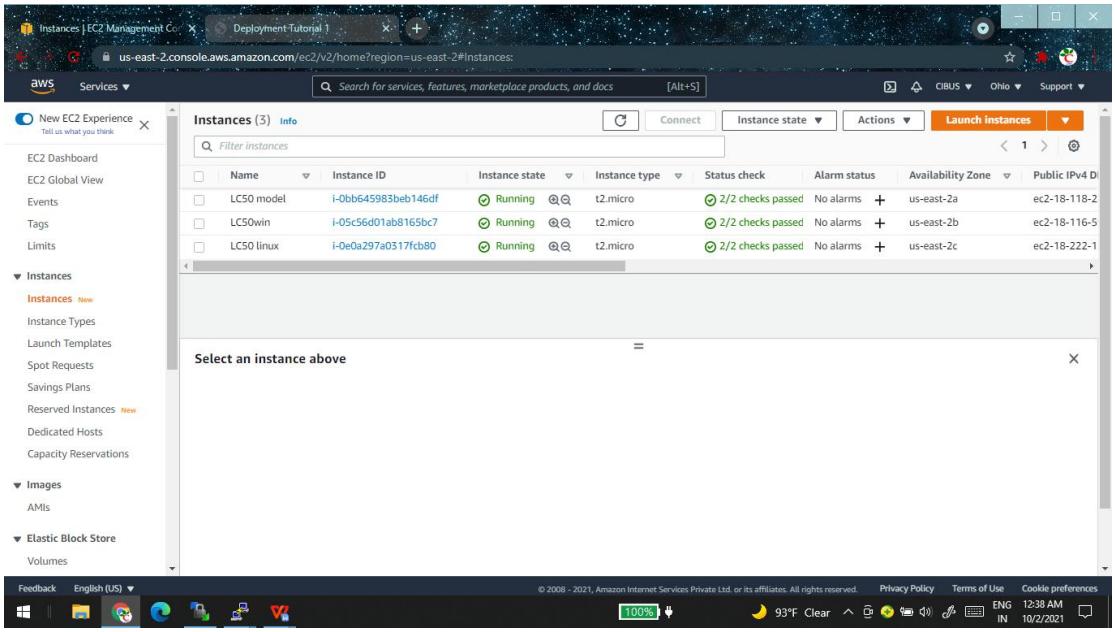
Tools used

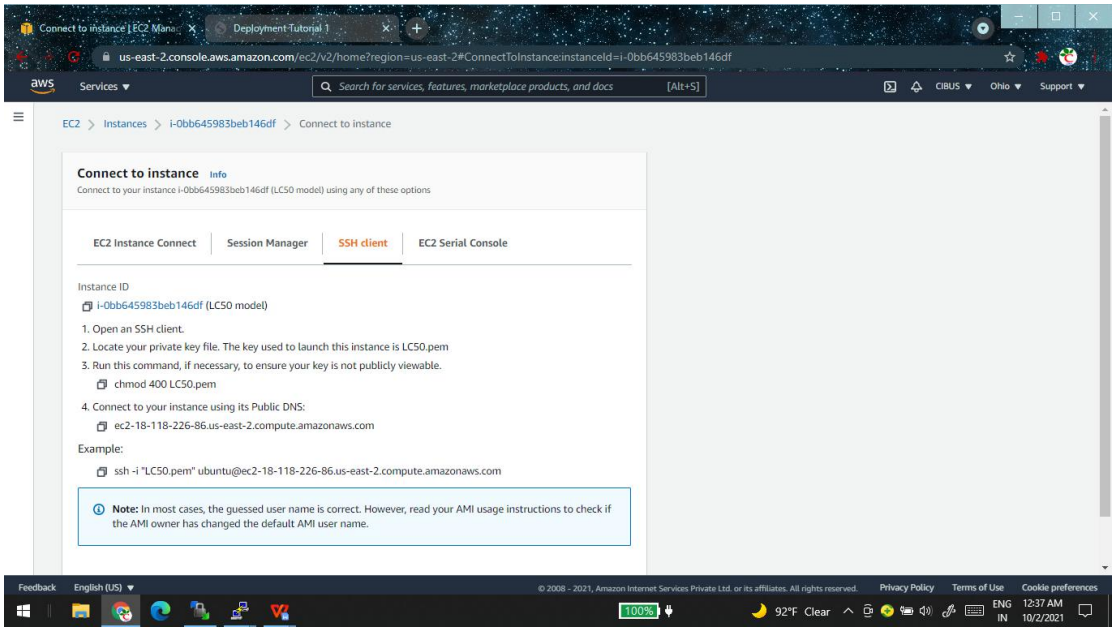
Numpy
Pandas
Sklearn
Flask
Spyder
AwS

Model flow




WPS Office
MOU Brothers trading company
Document1
ubuntu - ec2-18-118-226-86.us-east-2.compute.amazonaws.com - WinSCP
ubuntu@ip-172-31-11-126: ~
* Serving Flask app "app" (lazy loading)
* Environment: production
WARNING: This is a development server. Do not use it in a production deployment.
Use a production WSGI server instead.
* Debug mode: off
* Running on all addresses.
WARNING: This is a development server. Do not use it in a production deployment.
* Running on http://172.31.11.126:8080/ (Press CTRL+C to quit)
37.186.40.34 - - [01/Oct/2021 21:32:19] "GET / HTTP/1.1" 200 -
37.186.40.34 - - [01/Oct/2021 21:32:20] "GET /static/style.css HTTP/1.1" 404 -
37.186.40.34 - - [01/Oct/2021 21:32:21] "GET /favicon.ico HTTP/1.1" 404 -
37.186.40.34 - - [01/Oct/2021 21:32:43] "POST /predict HTTP/1.1" 200 -
37.186.40.34 - - [01/Oct/2021 21:32:44] "GET /static/style.css HTTP/1.1" 404 -


Instances | EC2 Management Co
Deployment Tutorial 1
us-east-2.console.aws.amazon.com/ec2/v2/home?region=us-east-2#Instances:
Instances (3) Info
Connect
Instance state
Actions
Launch instances
Select an instance above



LC50 value

cico | chem2 | chem3 | chem4 | chem5 | chem6 | LC50 value

LC50 value 5.46