Prediction of LC50

using (QSAR) models

-Aditya Rokade

Objective:-

Development of web app that can be used in chemical labs to calculate the concentration of LC50. The scients are calculating a concentration by experimenting on fish. In this near about 50% fish are dead due to high concentration of LC50. We find the best solution on that QSAR(Quantitative structure–activity relationship) model.

Benefits:-

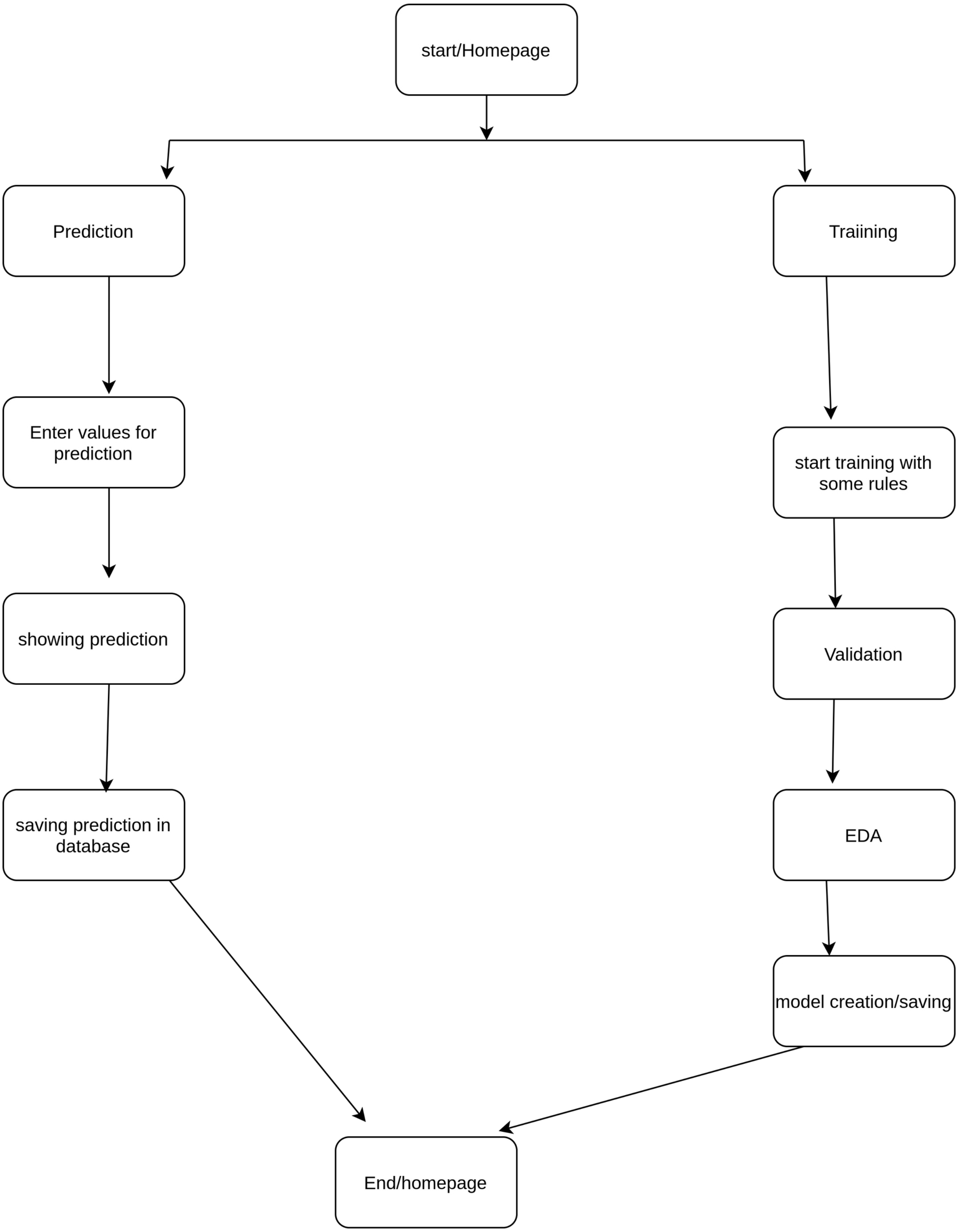
>User Friendly

>No code required

>Gives better responce

>Without doing experiment on any living things we can calculate the contration of LC50.

>Saves a lots of time and work


start/Homepage
Prediction
Traiining
Enter values for prediction
start training with some rules
showing prediction
Validation
saving prediction in database
EDA
model creation/saving
End/homepage

Start/Home:-

when we start or run the project then first we can see the home page .from home page we can go for prediction by clicking on calculate button.and from homepage we can go for retraining part also by clicking a training in navbar at the top.

Prediction:-

when we enter in prediction we can see the one input form on webpage. Then we need to enter the values of 6 molecular dispeters in each input box repetevely.then we need to press the calculate button.

Values for prediction:-

For prediction we need values in 2D array but we can accept the values in single mode ,then we program is converted the values in 2D array. Then forwarted to prediction.

Show prediction:-

when prediction is done by saving module we need to show to user.by using the table the prediction data is displayed to user and the concentration of LC50 is highlited.

Prediction saving:-

for future reference we need to store the prediction values somewhere.
To store the deteils we use the sqlite3 database which is provided by django inbuilt.
In database store the all details of user and values -username,labname,user email,input,output of values

Training:-

To train model first we need to provide a data file in which input and output also avaliable.

To train a model we need to store a data in “../training_data/” folder,number of files,number of columns ,etc.

Validation:-

1)file level validation:-

In validation first we do a file level validation like-number of files,number of columns,file name,etc. The validated files are moved to Good data folder and the files which are not validated are moved to bad data folder.
And after file level validation the data is stored in database(cassandra cloud) for future reference.

2)Data validation:-

After file validation there are another validation is data validation,inside data validation there are column validation ,giving column names,missing values in column and etc.
If there are any column having total NaN values then this files are moved from Good data folder to Bad data folder.
After data validation the data is go for EDA.

EDA:-

Before training EDA is important,inside EDA pandas report are created inside that almost everythings are covred-data format,missing value ,zero values.etc
After EDA the data is go for splittng of data for train and test of model

Model creation/Saving of model:-

inside model creation program is train a rainforest algorithm for creating a model ,
inside jupyter notebook we tryed a different a models ,our local system is not that much stronger to test a multiple model at a time.our model giving a near 70% accurecy ,after thar model is stored in one file for futute reference.

After training the goes to back to homepage...

Q & A

Q1)What’s the source of data?
Ans-The data is submitted by client in file format in perticular folder.

Q 2) What was the type of data?

Ans- The data is in numerical format-int,flout

Q 3) What’s the complete flow you followed in this Project?

Ans- for this please refer the page number 2,3 for better understanding

Q 4) How logs are managed?

Ans- Logger is implemented in each and every module and each and evry function with perticular file seperate.

Q 5) What techniques were you using for data pre-processing?

Ans- Following techniques are used :

Removing unwanted attributes

Visualizing relation of independent variables with each other and output variables

Checking and changing Distribution of continuous values

Removing outliers

Cleaning data and imputing if null values are present.

Splitting data

Q 6) How training was done or what models were used?

Training is done by using splitting data.in training Rainforest algorithm is used .

The score of this model is near 70&,and after model creation score checking is done.

Q 8) What are the different stages of deployment?
Ans8) The web app will get deployed in:

AWS
Azure
GCP
Heroku