Random Forest (RF) Regressor on QSAR models for predicting acute toxicity (LC_{50} 9 hours) on the *fathead minnow*

Srijan Acharya (224107213)

IIT Guwahati

1. Abstract

QSARs are one widely used in Quantitative Chemistry for studying Structure-Activity relationships. This present project uses a sanitised data set of 908 toxic chemicals to develop a supervised learning ensemble model for prediction of acute toxicity (label) on the *fathead* minnow from 6 molecular descriptors (features) and compares the effectiveness of decision trees over k-Nearest Neighbours (k=6) algorithm implemented by Cassotti et. al. (2015).

2. Introduction

Quantitative Structure-Activity relationship models (QSAR models) are regression or classification models used in chemical and biological sciences and engineering. [1] QSARs may be regression models (relating a set of predictor variables to the potency of a response variable to the "potency" of a response variable, like molecular structure to a measure of toxicity) or they may relate predictor variables to a categorical value of the response variable. One of the earliest applications of QSARs was in predicting the boiling points of liquids from molecular structure. QSARs exist on the bridge between the computer sciences and chemical sciences, and they have found extensive use in various sub-domains, notably toxicological analyses. Now that pollution of the atmosphere & hydrospheres - and eventually the biosphere

- being one the most potent problems plaguing the modern chemical industry, and the irreversible damage done to oceanic ecosystems threatening to backfire on the very existence of humanity, it is time we figure the chemicals which are safe for use moving forward. With 40,000 – 60,000 chemicals being produced annually, projected by double from 2017 to 2030 [2] and many new chemicals to come into commercial usage, it is smarter to figure out which chemical structures are supportive of sustainable development.

To this end, the Regulation on the registration, evaluation, authorisation and restriction of chemicals (REACH) is the main EU law for protection of human health and environment from the risks that can be posed by chemicals. [3] No such regulation exists in India, though she is developing her own REACH-like regulations [4]. REACH promotes alternative testing method for chemicals, like *in vitro* (in test tubes) and *in silico* (by means of simulation) methodologies

for chemical testing, and evaluation of short-term toxic effects of compounds imported/manufactured by more than 10 tonnes per year by EU member states [5]. A QSAR suitable model for this purpose must have a 'clearly defined' algorithm and end point, have an estimation of its 'domain of applicability' (all QSAR models have a specific Application Domain (AD)), the goodness-of-fit & predictivity of the model must be evaluated by appropriate strategies, and a mechanistic interpretation of model descriptors should be given, if possible. This project is based on the development of a QSAR model for prediction of acute toxicity, defined by LC_{50} 9 hours, towards the *fathead minnow*, of 908 organic chemicals, based on the Random Forest technique, and is based on [5], which used the kNN (k-Nearest Neighbours) methodology for solving the same problem. The data set provided for the problem was collected from the UCI Machine Learning Repository [6]. It was already sanitised at the time of obtaining it. **The predictive power of the algorithm was attempted to be put in relationship with aquatic toxicity and is estimated by means of thorough internal and external model validation procedures**.

3. Methodology

As mentioned above, the data was already cured and filtered at the time of obtaining it, with the help of KNIME (Konstanz Information Miner), and the technique, which can be found in [5] is being reproduced here. The raw data fed to KNIME was sourced from 3 databases, namely OASIS (OECD QSAR toolbox), ECOTOX (Ecotoxicology database, EPA) and EAT5 (ECETOC Aquatic Toxicity database) and data indicating ranges/thresholds were removed. Of the 4626 records thus obtained, both the CAS registry number (CAS-RN) and chemical name were available. To check that the CAS-RN and chemical name referred to the same structure, the ChemSpider database and the Chemical Identifier Resolver (CIR) of the CADD Group at NCI/NIH were used. Out of 4626 records (corresponding to 1139 unique CAS-RNs, plus 12 compounds lacking a CAS-RN), more than 50% presented mismatches (2422 records, corresponding to 518 different CAS-RNs and the 12 compounds lacking a CAS-RN). The records showing mismatches were exported and checked manually using PubChem & Sigma-Aldrich and also ChemSpider as additional sources. Records were deleted during this screening if they had a non-existent CAS-RN, a missing specification of which structural isomer(s) had been used, did not have an available molecular structure because the chemical only had a commercial name, or it proved to be impossible to be able to resolve a CAS-RN–chemical name mismatch. Finally, 2192 records corresponding to 441

different CAS-RNs were retained and merged with the 2204 records with matching structures, leaving 4396 records. Next it was checked whether each CAS-RN was associated with only one structure and vice versa. It was identified that 10 structures were associated with two different CAS-RNs. Since these CAS-RNs most probably obsolete were resolved only the CAS-RN indicated on the Sigma-Aldrich database was retained. Next 13 records with units coded as '%', '% v/v' and 'AI ng/L' were removed. The LC_{50} values of the remaining 1047 molecules were converted to $\log_{10}$ molarity units($-\log_{10}$ (mol/L)). Several molecules had multiple experimental values, which may occur because of: (a) different measurements; (b) the same measurement published in a paper that had been included in more source databases; and (c) the same measurement published in different papers. Since the median of the LC_{50} values would have been used for modelling, duplicates of the same measurement (cases b and c) had to be removed because they would distort the median. Duplicates of the same experiment from different databases (case b) were removed. Duplicates of the same experiment published in more papers (case c) often lacked references. Records with identical LC 50 value were considered as duplicates and only one record was retained. and the probability that two measurements would give the same LC_{50} value is, in principle, extremely low. To limit the model's AD to discrete organic molecules, only molecules with at least two carbon atoms and comprising only certain elements were allowed (H, Li, B, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, and I). Symbols specifying stereochemical configuration were removed from the Simplified Molecular Input Line Entry System (SMILES). Salts and mixtures were submitted to a dissociation algorithm in the OASIS Database Manager that first checked whether the species could be dissociated and then screened the potential dissociation products for non-toxic species. If more than one species was considered the source of toxicity, the record was removed. By doing so, it was possible to convert 50 mixtures and salts to a single organic component, assumed as the only source of the measured toxicity. Ions such as Na+, Mg 2+, Cl− were therefore not used for modelling. The dissociation products were neutralized, unless they were quaternary ammonium ions for which the charged form was retained. For three salts both the organic ion (acetate, benzoate and 2-hydroxybenzoate) and the inorganic counter-ion (K + or Na +) were considered nontoxic, and thus removed. For these three cases, the organic component was re-introduced and considered for modelling. In 15 cases, the dissociation product coincided with another molecule in the dataset, with toxicity values of these two species being very close for most instances, allowing for pooling the data. In four cases, mixtures of the type A+A+B were present and the dissociation algorithm returned only one molecule of A as source of toxicity.

The LC_{50} (molarity) values were accordingly doubled to correct for this approximation. At the end of this filtering stage, 929 molecules were retained. Final validation of the structures was made by comparing the SMILES in the dataset with those in the OpenTox database which lacked a structure for 58 compounds in the dataset. Large agreement in the structures of the remaining 871 molecules was observed. 9 mismatches were detected and solved by looking for the correct structure in the Sigma-Aldrich database. Only one case consisted of completely different compounds, whereas the other differences were mainly due to tautomers, valence and charge. For several compounds, multiple experimental values were available, showing differences of up to three logarithmic units. To reduce dependence on outlying toxicity data, the median, was calculated together with the corresponding standard deviation on the logarithmically transformed molarities ($-\log_{10}$(mol/L)) since it is more robust than the mean. The pooled standard deviation over the entire dataset was calculated (σ = 0.229 $\log_{10}$ (mol/L)) and used to derive an alert for inconsistent data (2σ = 0.458 $\log_{10}$ (mol/L)). Molecules with a standard deviation larger than 2σ were filtered out. Each experimental value was searched in the original scientific publication to detect errors in the compilation of the databases. If the scientific publication was not available/not found, the corresponding experimental value was deleted. 21 such chemicals with large standard deviations were removed because none of the original publication was accessible or found. The final dataset thus consisted of 908 organic molecules and is freely available.

The SMILES (Simplified Molecular-Input Line-Entry system) of these 908 species were used for calculating molecular descriptors using a DRAGON 6 software, details of which are available in original paper. In the end only 6 molecular descriptors were retained, and these were:

1) *MLOGP*: “log P”, the octanol-water partitioning coefficient, commonly used as measure of lipophilicity of organic compounds, and based on the Moriguchi model (hence the M)

2) *CIC0*: The 0^{th} order Complementary Information Content, belonging to a set of indices indicating neighbourhood symmetry, and is inversely related to the number of heteroatoms, where only the number of different elements is accounted for, and not the number of occurrences of each. It has historically proven to be significant at modelling acute toxicity.

3) *NdssC*: Number of unsaturated sp^2 carbon atoms of type =C< lying at the centre of multiple functional groups like ketones, carboxylic acids/carboxylates etc.

4) *NdsCH*: unsaturated sp^2 carbon atoms of type =C-H. encoding information about functional groups like aldehydes etc. Such electrophilic carbon atoms show substitution/addition reactions with nucleophiles which abound in biological systems. The hypothesis is that electrophilicity is highly correlated with toxicity, as in fact most common electrophiles are toxicants.

5) *SMI_Dz(Z)*: The spectral moment of order 1, the sum of eigenvalues of the Barysz matrix, accounting both for atomic number & bond order. It belongs set of descriptors calculated from 2D matrices, and low values imply absence of heteroatoms while high values imply several heteroatoms.

6) *GATSli*: The 2D Geary autocorrelation descriptor. Roughly, it is a measure of how much a molecule consists of bonds having differing ionization potentials.

This is essentially a regression machine learning problem. Random forests (RFs) are an ensemble learning method, based on the learning algorithm of Decision Trees, and are generally believed to perform better than decision trees. k-Nearest Neighbours (kNN) is also an ensemble learning method, and the authors of [5] have use k=6 to evaluate the dataset in question. The method used here is combines the kNN approach of 6-nearest neighbours, excluding the distance metric used in the paper.

The data was not scaled since all the values occur between the ranges of 1-10. The heatmap shows that MLOGP shows the highest correlation with the target metric, a fact agreed upon in toxicology, since lyophilicity is genuinely the major cause of toxicity, besides electrophilicity. A variety of train-test splits were tried, and in the end a 60%-40% train-test split was taken, since it showed the highest accuracy in the end. A kNN (base_mod) was constructed with n_neighbours = 6, and also an RF (RF) was constructed with 5000 estimators & max_features turned to 2. These values were chosen randomly, to test the performance against the kNN using (100 – %MAE) as a measure of accuracy besides tracking the mean square error (MSE). The aim was to perform better than the kNN in both error metrics (more accuracy and lower MSE). Following that, a new Random Decision Forest (rf) was constructed for hyperparameter tuning through Randomized Search Cross-Validation using the following code:

```
##Hyperparameter Tuning
from sklearn.model_selection import RandomizedSearchCV
```

```
n_e = [int(i) for i in np.linspace(1000, 5000, 100)]; #No. of estimators
m_f = ['auto', 'sqrt']; #, (int(i) for i in np.arange(10)); #Maximum
Features b_s = [True, False]; #bootstrap
r_grid = {'n_estimators': n_e, 'max_features': m_f, 'bootstrap': b_s};
rf_tuner = RandomizedSearchCV(estimator = rf, param_distributions = r_grid, n_iter =
100, cv = 3, verbose = 2, random_state=42, n_jobs=-1); #rf = Random Forest
rf_tuner.fit(qsar_train, y_train);

print(rf_tuner.best_params_);
```

As is evident, the no. of estimators (no. Of decision trees), maximum features (no. Of features selected at random and without replacement at split) and bootstrap (whether or not to extract data-points with sampling) hyper-parameters were tweaked to enhance model performance. Unfortunately, the number of parameters being too large, the Randomized Search did not converge for the 300 fits within 1 hours, proving to be intractable. Following that the number was reduced to 60 (number of iterations was changed to 20), but even so the algorithm proved intractable. Eventually, this approach was deemed futile and discarded, and a manual trial-and-error method of hyperparameter tuning was used. Finally, the 1st decision tree of the forest was plotted, following the procedure given in [8].

4. Results

1) Though the MAE was higher and the accuracy through MAE was lower for the random forest, the MSE was lower for it, giving conflicting results.

MAE tends to give equal importance to all errors, while MSE penalized large errors more than small errors. This explained by the fact that random forests use mse as the tuning criterion. However, the 2% difference caused implies the data is imbalanced.

2) The hyperparameter tuning revealed than no difference was caused by changing the number of estimators to from 5000 to 1. In fact, the MAE & MSE remained exactly the same for all values of estimators in the range.

This implies that a single decision tree, more than a random forest is more suited. This challenges the perception that a bootstrap aggregation is more suited for decision making than single decision trees.

3) A kNN calculates n = 908 quantities every time, and takes the decisions of the nearest neighbours while the decision tree makes ~ log2(909) which is about 9.8 ~ 10

comparisons for unsuccessful searches, which is vastly easier, for a lower MSE, which seems to be preferable over the MAE [9].

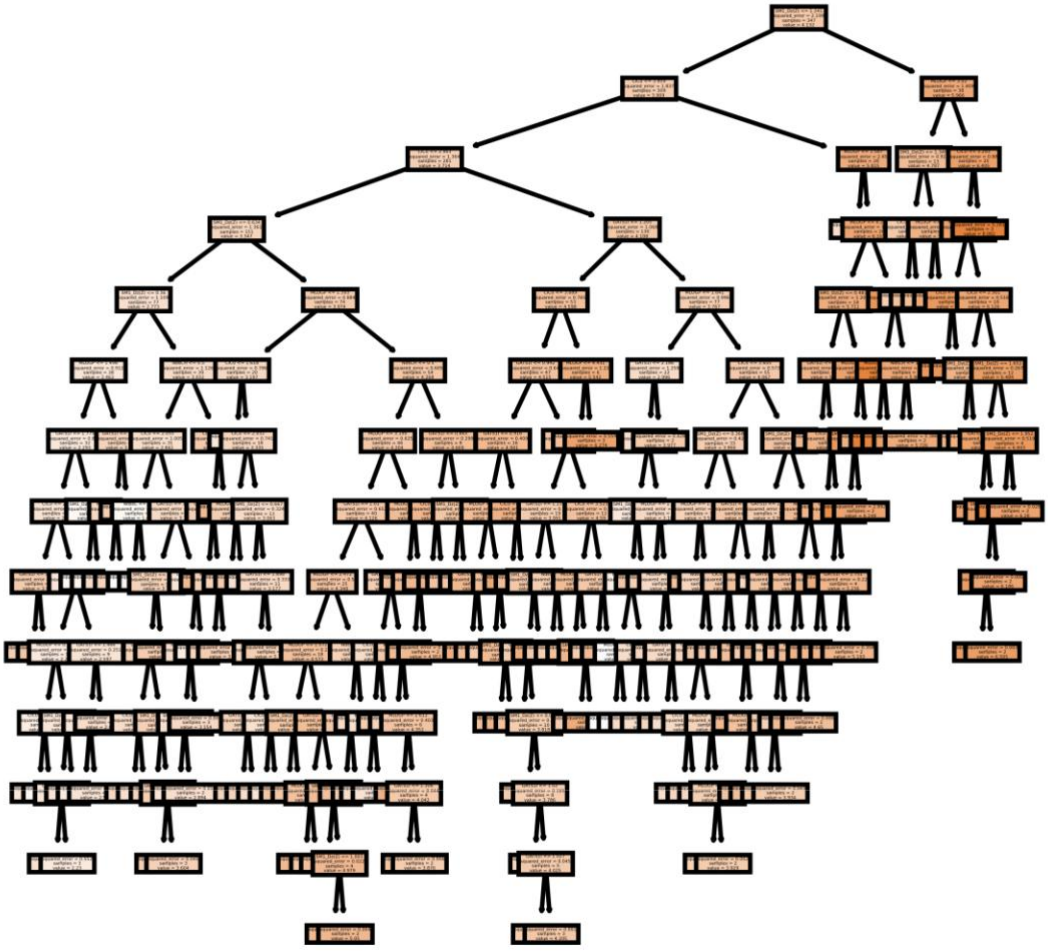

Fig. 1 The final decision tree (rf_0.png)

5. Conclusion

The project is inspired by a kNN model used on the same dataset & but produces a decision tree which works with a lower MSE but slightly higher MAE. Overall it reduces the time required for evaluation of the dataset.

The present work has the limitation that it applies only to the *fathead minnow* fish, and this needs to be further extended to other fish species to have a comprehensive and more general understanding of what is toxic to most fish, though much data seems to be lacking. Unlike REACH, Indian fish are nearly not studied as systematically, putting the life of a large sector of fish eaters' lives in possible jeopardy. Further research is most definitely needed in this area. Chemical Engineering, and more specifically Environmental Engineering and the related disciplines will benefit from such data, and such studies in water pollution may provide an impetus to conduct similar studies in Air Pollution and even Soil Pollution studies. Such studies may help in the study of aquatic ecosystems and contribute to their preservation and help in understanding which effluents from process industries are safe for disposal and precisely to what extent.

6. References

1. Mauri, A. Consonni, V. Todeschini, R. (2017). "Molecular Descriptors". Handbook of Computational Chemistry. Springer International Publishing. pp. 2065–2093. DOI:10.1007/978-3-319-27282-5_51. ISBN 978-3-319-27282-5.
2. Compendium of WHO and other UN guidance on health and environment, World Health Organisation (WHO).

(www.who.int/tools/compendium-on-health-and-environment/chemicals)

3. REACH Regulation, Energy, Climate Change, Environment, European Commission

(https://environment.ec.europa.eu/topics/chemicals/reach-regulation_en)

4. Overview of Chemical Regulations in India and Latest Developments, ChemSafetyPRO

(www.chemsafetypro.com/Topics/India/Overview_of_Chemical_Regulations_in_India)

5. Cassotti, M. Ballabio, D. Todeschini, R. Consonni, V. (2015). A similarity based QSAR model for predicting acute toxicity towards the fathead minnow (Pimophelas promelas) , SAR and QSAR in Environmental Research, 26:3, 217-243, DOI: 10.1080/1062936X.2015.1018938
6. Koerhsen, W. Random Forest in Python, Medium

(https://towardsdatascience.com/random-forest-in-python-24d0893d51c0)

7. Koerhsen, W. How to Visualise a Decision Tree from a Random Forest in Python using Scikit-Learn, Medium

(https://towardsdatascience.com/how-to-visualize-a-decision-tree-from-a-random-forest-in-python-using-scikit-learn-38ad2d75f21c)

8. Plot trees for a Random Forest in Python with Scikit-Learn, Stack Overflow (https://stackoverflow.com/questions/40155128/plot-trees-for-a-random-forest-in-python-with-scikit-learn)
9. Chai, T. Draxler, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 1247-1250, 2014, DOI: 10.5194/gmd-7-124-2014

7. Appendix

1) QSAR Data Set
2) Code
3) Decision Tree Diagram