QSAR Bio-concentration classification for drug detection using Artificial Neural Network using Synthetic Minority Oversampling Technique(SMOTE)

Project Author - Raj (201951123), Project Supervisor - Dr. Jignesh Bhatt

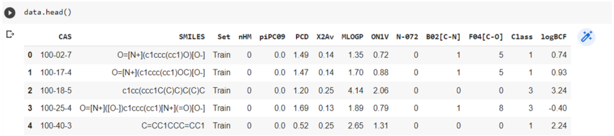
Introduction

A data set of manually-curated BCF (Bio concentration factor) for 779 chemicals was used to determine the mechanisms of bioconcentration, i.e. to predict whether a chemical:

1. is mainly stored within lipid tissues

2. has additional storage sites (e.g. proteins)

3. is metabolized/eliminated.



```
data.head()
```

	CAS	SMILES	Set	nHM	piPC09	PCD	X2Av	MLOGP	ON1V	N-072	B02[C-N]	F04[C-O]	Class	logBCF
0	100-02-7	O=[N+](c1ccc(cc1)O)[O-]	Train	0	0.0	1.49	0.14	1.35	0.72	0	1	5	1	0.74
1	100-17-4	O=[N+](c1ccc(cc1)OC)[O-]	Train	0	0.0	1.47	0.14	1.70	0.88	0	1	5	1	0.93
2	100-18-5	c1cc(ccc1C(C)C)C(C)C	Train	0	0.0	1.20	0.25	4.14	2.06	0	0	0	3	3.24
3	100-25-4	O=[N+]([O-])c1ccc(cc1)[N+](=O)[O-]	Train	0	0.0	1.69	0.13	1.89	0.79	0	1	8	3	-0.40
4	100-40-3	C=CC1CCC=CC1	Train	0	0.0	0.52	0.25	2.65	1.31	0	0	0	1	2.24

Attribute Information

Three Compound identifiers:
1. CAS number
2. Molecular SMILES
3. Train/test splitting

Nine molecular descriptors (independent variables)
4. nHM
5. piPC09
6. PCD
7. X2Av
8. MLOGP
9. ON1V
10. N-072
11. B02[C-N]
12. F04[C-O]

One experimental responses:
1. Bio accumulation class (three classes)

Previous works on the data

1. Data were randomly split into a training set of 584 compounds (75 percent) and a test set of 195 compounds (25 percent), preserving the proportion between the classes.

2. Two QSAR classification trees were developed using CART (Classification and Regression Trees) machine learning technique coupled with Genetic Algorithms.

3. The file contains the selected Dragon descriptors (9) along with CAS, SMILES, experimental BCF, experimental/predicted KOW and mechanistic class (1, 2, 3).

4. Further details on model development and performance along with descriptor definitions and interpretation are provided in the original manuscript (Grisoni et al., 2016).

Problem with the data set

1. Classes of target variable is imbalanced and need to be handled as our model will be creating bias for other classes larger in number.

2. Features that can be used to predict are less in number and Compound identifiers don't make sense and are not usable.

3. We are given a very small data set (not large enough to train on pre trained Convolutional Neural networks like ALEXNeT or VGGNet or ResNet with skip connection.

4. The training and validation accuracy difference keeps on increasing due to overfitting issue (exploding gradient problem).

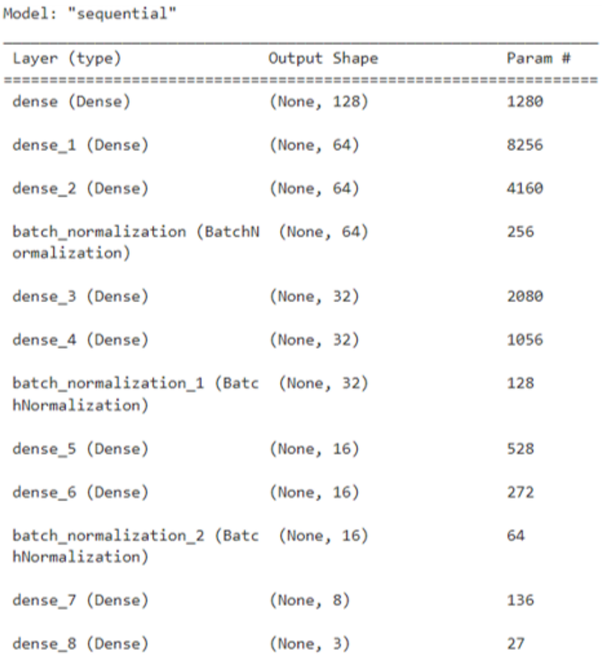
5. We have very less features to be able to use ANN's.

Method-1 Applying Artificial Neural Network without handling imbalanced data

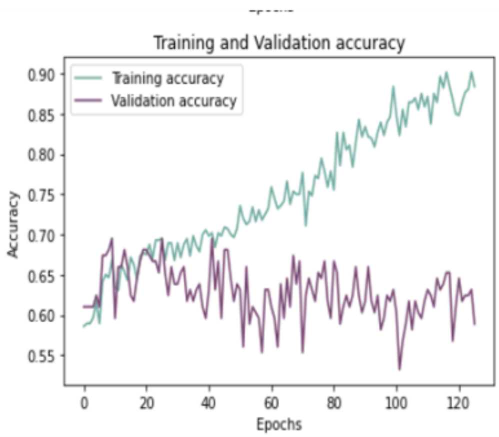
1. Data were randomly split into a training set of 584 compounds (75 percent) and a test set of 195 compounds (25 percent), preserving the proportion between the classes.

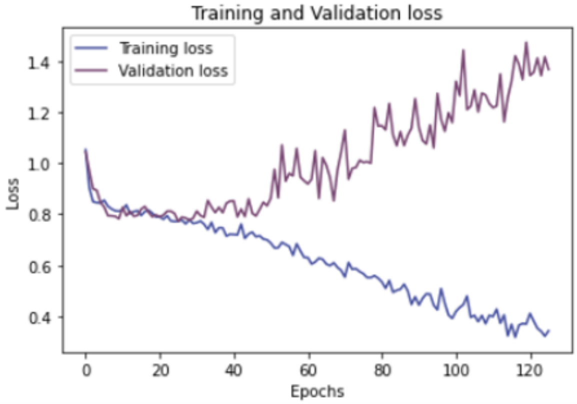
2. Two QSAR classification trees were developed using CART (Classification and Regression Trees) machine learning technique coupled with Genetic Algorithms.

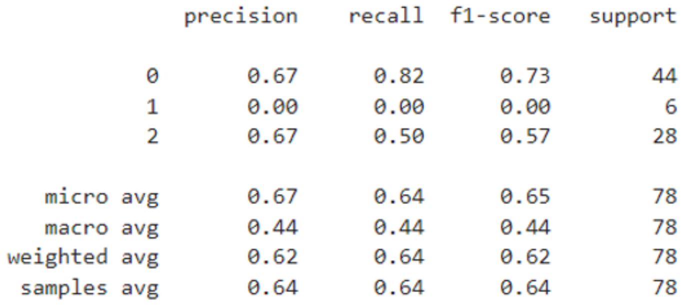
3. The file contains the selected Dragon descriptors (9) along with CAS, SMILES, experimental BCF, experimental/predicted KOW and mechanistic class (1, 2, 3).



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 128)               1280

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 64)                4160

 batch_normalization (BatchN  (None, 64)               256
 ormalization)

 dense_3 (Dense)             (None, 32)                2080

 dense_4 (Dense)             (None, 32)                1056

 batch_normalization_1 (Batc  (None, 32)               128
 hNormalization)

 dense_5 (Dense)             (None, 16)                528

 dense_6 (Dense)             (None, 16)                272

 batch_normalization_2 (Batc  (None, 16)               64
 hNormalization)

 dense_7 (Dense)             (None, 8)                 136

 dense_8 (Dense)             (None, 3)                 27
```





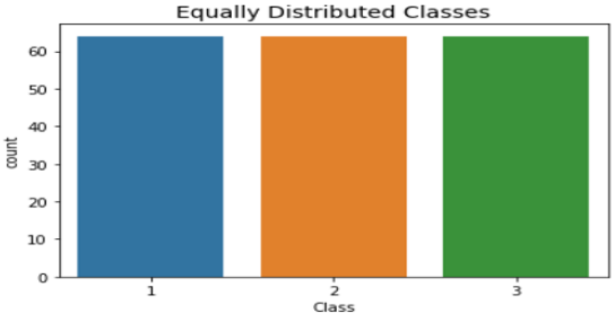


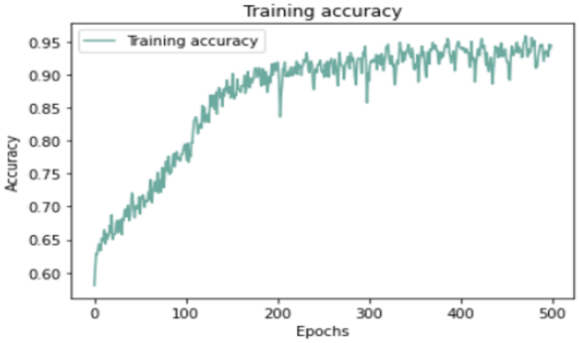
```
              precision    recall  f1-score   support

           0       0.67      0.82      0.73        44
           1       0.00      0.00      0.00         6
           2       0.67      0.50      0.57        28

   micro avg       0.67      0.64      0.65        78
   macro avg       0.44      0.44      0.44        78
weighted avg       0.62      0.64      0.62        78
 samples avg       0.64      0.64      0.64        78
```

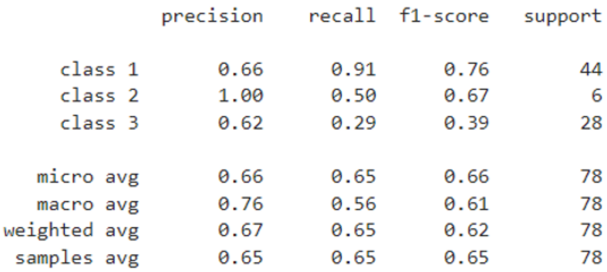
Method-2 Applying Artificial Neural Network with down sampling the data.

1. We are making other classes equal to the the class with least value.

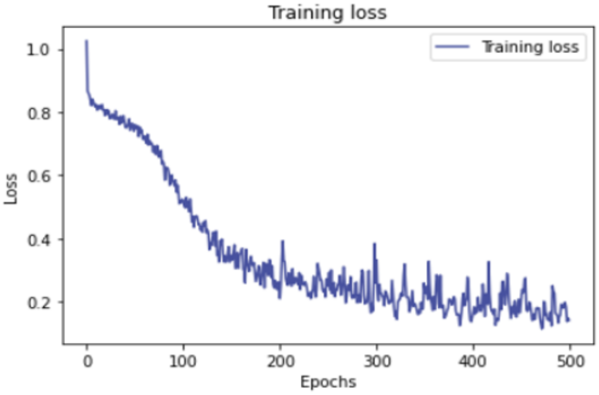
2. The data set is already less and after down sampling the data set decreases furthur and will result in more loss and precision will decrease even more.







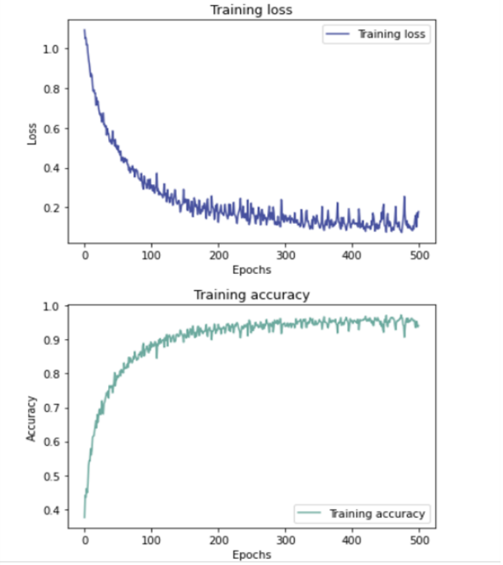
```
              precision    recall  f1-score   support

     class 1       0.66      0.91      0.76        44
     class 2       1.00      0.50      0.67         6
     class 3       0.62      0.29      0.39        28

   micro avg       0.66      0.65      0.66        78
   macro avg       0.76      0.56      0.61        78
weighted avg       0.67      0.65      0.62        78
 samples avg       0.65      0.65      0.65        78
```

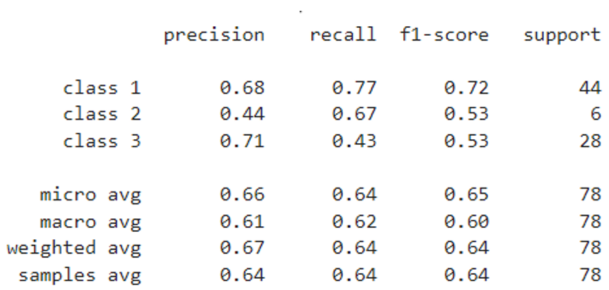


Method-3 Applying Artificial Neural Network with SMOTE (Synthetic Minority Oversampling Technique).

1. We are creating artificial data points using KNN algorithms for the class with least value.

2.This technique is better than down sampling but we are adding synthetic data over medical data set so it risky to use and can't always be trusted.





```
               precision    recall  f1-score   support

     class 1       0.68      0.77      0.72        44
     class 2       0.44      0.67      0.53         6
     class 3       0.71      0.43      0.53        28

   micro avg       0.66      0.64      0.65        78
   macro avg       0.61      0.62      0.60        78
weighted avg       0.67      0.64      0.64        78
 samples avg       0.64      0.64      0.64        78
```

Conclusions

We are getting the maximum precision using SMOTE technique. We are still not getting good metric values like precision, accuracy, f1-score as compared to the model used by paper using CART (Classification and Regression Tree) algorithm with a mixture of genetic algorithm.

Since, the data is less in quantity we can use Random Forest Classification Technique as this technique performs better under less features and data.

References

[1] https://archive.ics.uci.edu/ml/datasets/QSAR+Bioconcentration+classes+dataset

[2] Francesca Grisoni (francesca.grisoni '@' unimib.it), Viviana Consonni (viviana.consonni '@' unimib.it), Marco Vighi, Sara Villa, Roberto Todeschini

[3] F. Grisoni, V.Consonni, M.Vighi, S.Villa, R.Todeschini (2016). Investigating the mechanisms of bioconcentration through QSAR classification trees, Environment International, 88, 198-205

[4] F. Grisoni, V.Consonni, M.Vighi, S.Villa, R.Todeschini (2016). Investigating the mechanisms of bioconcentration through QSAR classification trees, Environment International, 88, 198-205. F. Grisoni, V. Consonni, S. Villa, M. Vighi, R. Todeschini (2015). QSAR models for bioconcentration: Is the increase in the complexity justified by more accurate predictions?. Chemosphere, 127, 171-179.