Report on Machine Learning Mini Project

Prediction of Toxicity

Varshaneya V, I M.Tech C.S. 17559

Nikhil Rai, I M.Tech C.S. 17556

Aim:

To build a model for classifying toxicity of compounds in Tox21 dataset based on their reactivity towards 7 different assays.

Platform Used:

1. Anaconda python (version 2.7.14)
2. NumPy for array related manipulations.
3. Sklearn for different classifiers and performance metrics.
4. PyBioMed for extracting features like ECFP4 and ECFP6.
5. Keras for building and training neural network.
6. Imblearn for SMOTE.
7. Matplotlib for ploting of ROC curves.

Models used:

1. **Support Vector Machine**
   i. Linear kernel
   ii. Tolerance is 0.01
   iii. Decision function shape is “one versus rest”.
2. **Random Forest**
   i. 25 trees as estimators.
   ii. Gini index as the criterion of split.
3. **Neural Network**
   i. Input layer of dimension 1024 with tanh activation.
   ii. Hidden layer of dimension 512 with relu activation.
   iii. Output layer of single neuron which outputs either 0 or 1.
   iv. Optimizer used is Adam, with loss function as binary cross entropy.

Experiments Conducted:

The toxicity of compounds are based on their reactivity towards 7 assays namely NRAHR, NRAR, NRARLBD, NRARAROMATASE, NRER, NRERLBD and NRPPARGAMMA. The tox21 dataset used has the reacitivity of set of compunds with each of these assays. If a compund reacts with an assay then that is represented as 1 and 0 if there is no reaction. The reactivity of these compounds depend on their structures. So in order to capture their structures succinctly, features extraction methods like ECFP4 and ECFP6 are used. These methods give a 1024-bit representation of the compunds based on the arrangement of atoms. The 1024-bit representation forms the features for each of the compound.

The data for each of the assay is highly imbalanced with compunds reacting to a given assay being very very less than compared compounds that donot react with the assay. In other words the class 0 has got a large number of compounds compared to compounds belonging to class 1. So inoreder to account for class imbalance SMOTING technique has been employed as against the base-line of not using SMOTE. SMOTE stands for Synthetic Minority Over-sampling TEchnique.

The over-sampling with replacement doesn't significantly improve minority class recognition. SMOTE interprets the underlying effect in terms of decision regions in feature space. Essentially, as the minority class is over-sampled by increasing amounts, by identifying similar but more specific regions in the feature space as the decision region for the minority class.

In order to find an efficient model which consists of a feature extraction and a machine learning technique we have embarked upon considering 3 classifiers as mentioned in the "models used" section without using SMOTE and with using SMOTE. As far as feature extraction methods are concerned we have considered ECFP4 and ECFP6 features.

We have found AUC of ROC and accuracy using cross validation for the 3 classifiers both with and without SMOTING using both the features on 7 assays. The observations are tabulated in the observations section separately for each of the assay.

Dataset description:

Toxicity of the compounds were studied by studying their reaction with 7 different assays. So we have 7 datasets to work with as mentioned. Below is a statistics of the number of classes in each of the dataset.

Assay	Number of samples for class 0	Number of samples for class 1
NRAHR	7220	950
NRAR	8982	380
NRARLBD	8296	303
NRARAROMATASE	6866	360
NRER	6761	937
NRERLBD	8307	446
NRPPARGAMMA	7962	222

It can be noticed that there is high imbalance in the data pertaining to one of the classes. This is because of the fact that there are only a few componds that react with a particular assay. So to account for the class we have compared the classifiers with SMOTING as against the baseline which is without SMOTING. SMOTING was done as a part preprocessing step after feature extraction.

Observations:

NRAHR

ECFP6 features:

Model	AUC w/o SMOTE	Accuracy w/o SMOTE	AUC with SMOTE	Accuracy with SMOTE
Random Forest	0.7	92.59	0.97	96.43
SVM	0.74	89.67	0.93	92.46
Neural Network	0.69	92.85	0.93	94.71

Model	Class	Precision	Recall	f1-score	Support
Random forest without SMOTE	0	0.92	0.99	0.95	1793
	1	0.83	0.40	0.54	248
	avg/ total	0.91	0.92	0.90	2041
Random forest with SMOTE	0	0.95	0.98	0.97	1825
	1	0.98	0.95	0.96	1782
	avg / total	0.97	0.97	0.97	3607
SVM without SMOTE	0	0.94	0.94	0.94	1793
	1	0.56	0.53	0.54	248

	avg/ total	0.89	0.89	0.89	2041
SVM with SMOTE	0	0.96	0.89	0.92	1825
	1	0.89	0.96	0.93	1782
	avg / total	0.93	0.93	0.93	3607
Neural network without SMOTE	0	0.92	0.98	0.95	1793
	1	0.70	0.40	0.51	248
	avg / total	0.89	0.91	0.89	2041
Neural network with SMOTE	0	0.95	0.92	0.93	1825
	1	0.92	0.95	0.93	1782
	avg / total	0.93	0.93	0.93	3607

ECFP4 features:

Model	**AUC w/o SMOTE**	**Accuracy w/o SMOTE**	**AUC with SMOTE**	**Accuracy with SMOTE**
Random Forest	0.72	92.96	0.96	96.51
SVM	0.74	90.07	0.92	92.96
Neural Network	0.71	94.29	0.95	96.07

Model	**Class**	**Precision**	**Recall**	**f1-score**	**Support**
Random Forest without SMOTE	0	0.93	0.99	0.96	1793
	1	0.82	0.46	0.59	248
	avg / total	0.92	0.92	0.91	2041
Random Forest with SMOTE	0	0.95	0.98	0.96	1821
	1	0.98	0.95	0.96	1786
	avg / total	0.96	0.96	0.96	3607
SVM without SMOTE	0	0.94	0.95	0.94	1793
	1	0.58	0.53	0.55	248
	avg / total	0.89	0.90	0.89	2041
SVM with	0	0.96	0.88	0.92	1821

SMOTE	1	0.89	0.96	0.92	1786
	avg / total	0.92	0.92	0.92	3607
Neural Network with SMOTE	0	0.93	0.96	0.95	1793
	1	0.63	0.46	0.53	248
	avg / total	0.89	0.90	0.89	2041
Neural Network without SMOTE	0	0.97	0.93	0.95	1821
	1	0.93	0.97	0.95	1786
	avg / total	0.95	0.95	0.95	3607

NRAR:

ECFP6 features

Model	**AUC w/o SMOTE**	**Accuracy w/o SMOTE**	**AUC with SMOTE**	**Accuracy with SMOTE**
Random Forest	0.77	97.8	0.98	98.74
SVM	0.76	96.36	0.98	97.27
Neural Network	0.76	97.92	0.98	50.0

Model	**Class**	**Precision**	**Recall**	**f1-score**	**Support**
Random forest without SMOTE	0	0.98	0.99	0.99	2246
	1	0.73	0.54	0.62	94
	avg / total	0.97	0.97	0.97	2340
Random forest with SMOTE	0	0.98	0.99	0.98	2239
	1	0.99	0.98	0.98	2250
	avg / total	0.98	0.98	0.98	4489
SVM without SMOTE	0	0.98	0.98	0.98	2246
	1	0.50	0.55	0.52	94
	avg / total	0.96	0.96	0.96	2340
SVM with SMOTE	0	1.00	0.95	0.98	2239
	1	0.95	1.00	0.98	2250
	avg / total	0.98	0.98	0.98	4489

Neural network without SMOTE	0	0.98	1.00	0.99	2246
	1	0.86	0.52	0.65	94
	avg / total	0.98	0.98	0.97	2340
Neural network with SMOTE	0	1.00	0.96	0.98	2239
	1	0.97	1.00	0.98	2250
	avg / total	0.98	0.98	0.98	4489

ECFP4 features:

Model	**AUC w/o SMOTE**	**Accuracy w/o SMOTE**	**AUC with SMOTE**	**Accuracy with SMOTE**
Random Forest	0.77	97.71	0.98	95.75
SVM	0.76	96.83	0.97	96.99
Neural Network	0.76	97.81	0.97	97.95

Model	**Class**	**Precision**	**Recall**	**f1-score**	**Support**
Random Forest without SMOTE	0	0.98	0.99	0.99	2246
	1	0.71	0.54	0.61	94
	avg / total	0.97	0.97	0.97	2340
Random Forest with SMOTE	0	0.98	0.99	0.98	2239
	1	0.99	0.98	0.98	2250
	avg / total	0.98	0.98	0.98	4489
SVM without SMOTE	0	0.98	0.98	0.98	2246
	1	0.53	0.54	0.54	94
	avg / total	0.96	0.96	0.96	2340
SVM with SMOTE	0	1.00	0.94	0.97	2239
	1	0.94	1.00	0.97	2250
	avg / total	0.97	0.97	0.97	4489
Neural Network with SMOTE	0	0.98	0.99	0.99	2246
	1	0.79	0.52	0.63	94

	avg / total	0.97	0.98	0.97	2340
Neural Network without SMOTE	0	1.00	0.94	0.97	2239
	1	0.95	1.00	0.97	2250
	avg / total	0.97	0.97	0.97	4489

NRARLBD

ECFP6 features:

Model	AUC w/o SMOTE	Accuracy w/o SMOTE	AUC with SMOTE	Accuracy with SMOTE
Random Forest	0.78	98.23	0.99	99.26
SVM	0.78	97.63	0.99	98.8
Neural Network	0.5	98.0	0.98	99.0

Model	Class	Precision	Recall	f1-score	Support
Random forest without SMOTE	0	0.99	1.00	0.99	2078
	1	0.87	0.56	0.68	71
	avg / total	0.98	0.98	0.98	2149
Random forest with SMOTE	0	0.99	1.00	0.99	2064
	1	1.00	0.99	0.99	2082
	avg / total	0.99	0.99	0.99	4146
SVM without SMOTE	0	0.99	0.99	0.99	2078
	1	0.58	0.58	0.58	71
	avg / total	0.97	0.97	0.97	2149
SVM with SMOTE	0	1.00	0.97	0.99	2064
	1	0.97	1.00	0.99	2082
	avg / total	0.99	0.99	0.99	4146
Neural network without SMOTE	0	0.97	1.00	0.98	2078
	1	0.00	0.00	0.00	71

	avg / total	0.94	0.97	0.95	2149
Neural network with SMOTE	0	0.98	0.97	0.98	2064
	1	0.98	0.98	0.98	2082
	avg / total	0.98	0.98	0.98	4146

ECFP4 features:

Model	**AUC w/o SMOTE**	**Accuracy w/o SMOTE**	**AUC with SMOTE**	**Accuracy with SMOTE**
Random Forest	0.79	98.28	0.99	99.31
SVM	0.80	97.44	0.99	98.59
Neural Network	0.51	96.60	0.99	50.0

Model	**Class**	**Precision**	**Recall**	**f1-score**	**Support**
Random Forest without SMOTE	0	0.99	1.00	0.99	2078
	1	0.87	0.58	0.69	71
	avg / total	0.98	0.98	0.98	2149
Random Forest with SMOTE	0	0.99	0.99	0.99	2064
	1	0.99	0.99	0.99	2082
	avg / total	0.99	0.99	0.99	4146
SVM without SMOTE	0	0.99	0.99	0.99	2078
	1	0.73	0.61	0.66	71
	avg / total	0.98	0.98	0.98	2149
SVM with SMOTE	0	1.00	0.97	0.99	2064
	1	0.97	1.00	0.99	2082
	avg / total	0.99	0.99	0.99	4146
Neural Network with SMOTE	0	0.97	1.00	0.98	2078
	1	1.00	0.01	0.03	71
	avg / total	0.97	0.97	0.95	2149

Neural Network without SMOTE	0	1.00	0.98	0.99	2064
	1	0.98	1.00	0.99	2082
	avg / total	0.99	0.99	0.99	4146

NRARAROMATASE

ECFP6 features:

Model	AUC w/o SMOTE	Accuracy w/o SMOTE	AUC with SMOTE	Accuracy with SMOTE
Random Forest	0.68	96.59	0.99	98.58
SVM	0.72	94.46	0.97	96.87
Neural Network	0.5	95.8	0.97	97.4

Model	Class	Precision	Recall	f1-score	Support
Random forest without SMOTE	0	0.97	0.99	0.98	1723
	1	0.75	0.36	0.49	83
	avg / total	0.96	0.97	0.96	1806
Random forest with SMOTE	0	0.98	1.00	0.99	1741
	1	1.00	0.98	0.99	1690
	avg / total	0.99	0.99	0.99	3431
SVM without SMOTE	0	0.97	0.97	0.97	1723
	1	0.46	0.47	0.47	83
	avg / total	0.95	0.95	0.95	1806
SVM with SMOTE	0	1.00	0.94	0.97	1741
	1	0.94	1.00	0.97	1690
	avg / total	0.97	0.97	0.97	3431
Neural network without SMOTE	0	0.95	1.00	0.98	1723
	1	0.00	0.00	0.00	83
	avg / total	0.91	0.95	0.93	1806
Neural network with	0	0.99	0.95	0.97	1741

SMOTE	1	0.95	0.99	0.97	1690
	avg / total	0.97	0.97	0.97	3431

ECFP4 features:

Model	**AUC w/o SMOTE**	**Accuracy w/o SMOTE**	**AUC with SMOTE**	**Accuracy with SMOTE**
Random Forest	0.69	96.79	0.99	98.49
SVM	0.75	94.75	0.97	96.79
Neural Network	0.50	94.80	0.50	50.20

Model	**Class**	**Precision**	**Recall**	**f1-score**	**Support**
Random Forest without SMOTE	0	0.97	0.99	0.98	1723
	1	0.70	0.40	0.51	83
	avg / total	0.96	0.96	0.96	1806
Random Forest with SMOTE	0	0.98	0.99	0.99	1741
	1	0.99	0.98	0.99	1690
	avg / total	0.99	0.99	0.99	3431
SVM without SMOTE	0	0.98	0.97	0.97	1723
	1	0.46	0.53	0.49	83
	avg / total	0.95	0.95	0.95	1806
SVM with SMOTE	0	1.00	0.95	0.97	1741
	1	0.95	1.00	0.97	1690
	avg / total	0.97	0.97	0.97	3431
Neural Network with SMOTE	0	0.95	1.00	0.98	1723
	1	0.00	0.00	0.00	83
	avg / total	0.91	0.95	0.93	1806
Neural Network without SMOTE	0	0.51	1.00	0.67	1741
	1	0.00	0.00	0.00	1690
	avg / total	0.26	0.51	0.34	3431

NRER

ECFP6 features:

Model	AUC w/o SMOTE	Accuracy w/o SMOTE	AUC with SMOTE	Accuracy with SMOTE
Random Forest	0.66	89.96	0.95	94.74
SVM	0.63	87.07	0.84	83.9
Neural Network	0.5	89.8	0.83	89.8

Model	Class	Precision	Recall	f1-score	Support
Random forest without SMOTE	0	0.92	0.98	0.95	1698
	1	0.74	0.34	0.46	226
	avg / total	0.90	0.91	0.89	1924
Random forest with SMOTE	0	0.93	0.96	0.95	1712
	1	0.96	0.93	0.95	1666
	avg / total	0.95	0.95	0.95	3378
SVM without SMOTE	0	0.91	0.94	0.92	1698
	1	0.41	0.32	0.36	226
	avg / total	0.85	0.86	0.86	1924
SVM with SMOTE	0	0.90	0.78	0.83	1712
	1	0.80	0.91	0.85	1666
	avg / total	0.85	0.84	0.84	3378
Neural network without SMOTE	0	0.88	1.00	0.94	1698
	1	0.00	0.00	0.00	226
	avg / total	0.78	0.88	0.83	1924
Neural network with SMOTE	0	0.84	0.81	0.83	1712
	1	0.81	0.84	0.83	1666
	avg / total	0.83	0.83	0.83	3378

ECFP4 features:

Model	**AUC w/o SMOTE**	**Accuracy w/o SMOTE**	**AUC with SMOTE**	**Accuracy with SMOTE**
Random Forest	0.65	96.59	0.95	98.58
SVM	0.64	88.22	0.85	84.48
Neural Network	0.62	91.2	0.87	86.8

Model	**Class**	**Precision**	**Recall**	**f1-score**	**Support**
Random Forest without SMOTE	0	0.92	0.98	0.95	1698
	1	0.72	0.32	0.45	226
	avg / total	0.89	0.91	0.89	1924
Random Forest with SMOTE	0	0.94	0.97	0.95	1712
	1	0.97	0.93	0.95	1665
	avg / total	0.95	0.95	0.95	3377
SVM without SMOTE	0	0.91	0.94	0.93	1698
	1	0.43	0.34	0.38	226
	avg / total	0.86	0.87	0.86	1924
SVM with SMOTE	0	0.90	0.79	0.84	1712
	1	0.81	0.91	0.85	1665
	avg / total	0.85	0.85	0.85	3377
Neural Network with SMOTE	0	0.91	0.97	0.94	1698
	1	0.53	0.28	0.36	226
	avg / total	0.86	0.89	0.87	1924
Neural Network without SMOTE	0	0.90	0.82	0.86	1712
	1	0.83	0.91	0.87	1665
	avg / total	0.87	0.87	0.87	3377

NRERLBD

ECFP6 features:

Model	**AUC w/o SMOTE**	**Accuracy w/o SMOTE**	**AUC with SMOTE**	**Accuracy with SMOTE**
Random Forest	0.72	96.5	0.98	98.46
SVM	0.71	94.74	0.96	96.52
Neural Network	0.54	96.6	0.95	97.6

Model	Class	Precision	Recall	f1-score	Support
Random forest without SMOTE	0	0.97	0.99	0.98	2085
	1	0.80	0.44	0.57	102
	avg / total	0.97	0.97	0.96	2187
Random forest with SMOTE	0	0.97	0.99	0.98	2031
	1	0.99	0.97	0.98	2120
	avg / total	0.98	0.98	0.98	4151
SVM without SMOTE	0	0.97	0.96	0.97	2085
	1	0.38	0.45	0.41	102
	avg / total	0.95	0.94	0.94	2187
SVM with SMOTE	0	1.00	0.93	0.96	2031
	1	0.94	1.00	0.97	2120
	avg / total	0.97	0.96	0.96	4151
Neural network without SMOTE	0	0.96	1.00	0.98	2085
	1	0.73	0.08	0.14	102
	avg / total	0.95	0.96	0.94	2187
Neural network with SMOTE	0	0.97	0.93	0.95	2031
	1	0.94	0.97	0.96	2120
	avg / total	0.95	0.95	0.95	4151

ECFP4 features:

Model	AUC w/o SMOTE	Accuracy w/o SMOTE	AUC with SMOTE	Accuracy with SMOTE
Random Forest	0.72	96.50	0.98	98.46
SVM	0.71	95.43	0.96	96.27
Neural Network	0.50	96.6	0.96	96.6

Model	Class	Precision	Recall	f1-score	Support
Random Forest without SMOTE	0	0.97	0.99	0.98	2085
	1	0.78	0.44	0.56	102
	avg / total	0.96	0.97	0.96	2187
Random Forest with SMOTE	0	0.97	0.99	0.98	2086
	1	0.99	0.97	0.98	2065
	avg / total	0.98	0.98	0.98	4151
SVM without SMOTE	0	0.97	0.98	0.98	2085
	1	0.51	0.43	0.47	102
	avg / total	0.95	0.95	0.95	2187
SVM with SMOTE	0	1.00	0.93	0.96	2086
	1	0.93	1.00	0.96	2065
	avg / total	0.96	0.96	0.96	4151
Neural Network with SMOTE	0	0.95	1.00	0.98	2085
	1	0.00	0.00	0.00	102
	avg / total	0.91	0.95	0.93	2187
Neural Network without SMOTE	0	0.98	0.95	0.96	2086
	1	0.95	0.98	0.96	2065
	avg / total	0.96	0.96	0.96	4151

NRPPARGAMMA

ECFP6 features:

Model	AUC w/o SMOTE	Accuracy w/o SMOTE	AUC with SMOTE	Accuracy with SMOTE
Random Forest	0.65	97.81	0.99	99.19
SVM	0.72	96.72	0.98	98.46
Neural Network	0.5	97.2	0.98	50.0

Model	Class	Precision	Recall	f1-score	Support
Random forest without SMOTE	0	0.98	1.00	0.99	1999
	1	0.74	0.30	0.43	46
	avg / total	0.98	0.98	0.98	2045

Random forest with SMOTE	0	0.99	1.00	0.99	2002
	1	1.00	0.98	0.99	1977
	avg / total	0.99	0.99	0.99	3979
SVM without SMOTE	0	0.99	0.98	0.98	1999
	1	0.34	0.46	0.39	46
	avg / total	0.97	0.97	0.97	2045
SVM with SMOTE	0	1.00	0.97	0.98	2002
	1	0.97	1.00	0.98	1977
	avg / total	0.98	0.98	0.98	3979
Neural network without SMOTE	0	0.98	1.00	0.99	1999
	1	0.00	0.00	0.00	46
	avg / total	0.96	0.98	0.97	2045
Neural network with SMOTE	0	1.00	0.96	0.98	2002
	1	0.96	1.00	0.98	1977
	avg / total	0.98	0.98	0.98	3979

ECFP4 features:

Model	**AUC w/o SMOTE**	**Accuracy w/o SMOTE**	**AUC with SMOTE**	**Accuracy with SMOTE**
Random Forest	0.67	97.84	0.99	99.2
SVM	0.70	96.82	0.98	98.27
Neural Network	0.50	97.2	0.98	50.0

Model	**Class**	**Precision**	**Recall**	**f1-score**	**Support**
Random Forest without SMOTE	0	0.99	1.00	0.99	1999
	1	0.76	0.35	0.48	46
	avg / total	0.98	0.98	0.98	2045
Random Forest with SMOTE	0	0.99	1.00	0.99	2002
	1	1.00	0.99	0.99	1977

	avg / total	0.99	0.99	0.99	3979
SVM without SMOTE	0	0.99	0.99	0.99	1999
	1	0.40	0.41	0.41	46
	avg / total	0.97	0.97	0.97	2045
SVM with SMOTE	0	1.00	0.97	0.98	2002
	1	0.97	1.00	0.98	1977
	avg / total	0.98	0.98	0.98	3979
Neural Network with SMOTE	0	0.98	1.00	0.99	1999
	1	0.00	0.00	0.00	46
	avg / total	0.96	0.98	0.97	2045
Neural Network without SMOTE	0	1.00	0.97	0.98	2002
	1	0.97	1.00	0.98	1977
	avg / total	0.98	0.98	0.98	3979

Inferences:

The classifiers were chosen based on the AUC of ROC and the accuracy obtained by 5-fold cross-validation on the training set. So the classifier that gave highest AUC and accuracy compared to others across both the feature extraction methods, is chosen to be the appropriate one and that feature extraction method is chosen to be the preferred choice for pre-processing. The results are tabulated below.

Assay	**Classifier selected**	**Feature Extraction**	**Accuracy (%)**
NRAHR	Random forest with SMOTE	ECFP6	96.43
NRAR	Random forest with SMOTE	ECFP6	98.74
NRARLBD	Random forest with SMOTE	ECFP4	99.31
NRARAROMATASE	Random forest with SMOTE	ECFP6	98.58
NRER	Random forest with SMOTE	ECFP4	98.58
NRERLBD	Random forest with SMOTE	ECFP4 / ECFP6	98.46
NRPPARGAMMA	Random forest with SMOTE	ECFP4	99.2

It is observed that SMOTING helps in improving accuracy of classifiers in most of the instances here. This is largely because SMOTE artificially creates samples of the minority class by sampling in the neighbourhood of points in minority class. But in some cases there is only a marginal improvement in the accuracy, as calculated by 5-fold cross-validation.

The are cases in which the accuracy of classifier has deteriorated after SMOTING. These have been marked in red. The accuracy of SVM has deteriorated after SMOTING for the assay NRER (with both ECFP4 and ECFP6 features). Same is the case with neural network in the assays NRPPARGAMMA (with both ECFP4 and ECFP6 features), NRER (with ECFP4 features), NRARAROMATASE (with ECFP4 features), NRARLBD (with ECFP4 features) and NRAR (with ECFP6 features). The same trend is shown by random forest for the assay NRAR (with ECFP4 features). The neural network is shallow so because of which the network is unable to capture the representation of the compounds. Cleverly crafted deep neural networks can do this better.

SMOTE improves the recall and precision of all the classifiers. The classification report for each of the assays and for each of the feature extraction methods proves this aspect. Improving the recall also improves the f1-score and AUC when compared to using classifiers without SMOTE. Since SMOTE balances instances from the minority class, the classifer has now learnt the actual representation of the minority class which in our case is class 1. So because of this the model is able to reduce the false negatives and false positives thereby increasing both recall and precision of class 1. This leads to an increase in AUC of ROC for all the classifiers which is consistent with the observations.

Conclusions:

Random forest with SMOTE is the best classifier on the basis of AUC of ROC and cross-validation, for the various assays. But it turns out that the feature extraction methods that give good accuracy (along with the classifer) is different for different assays. After studying 3 classifers along with 2 feature extraction methods, the best model consisting of classifier and feature extraction method, for each of the dataset has been tabulated in the ‘inferences’ section based on AUC of ROC and accuracy.

Future work could consist of studying ECFP4 and ECFP6 features in depth and their relation to the structure of the compound. Alternative feature extraction methods that have a close relation to the reactivity of the compounds could also be explored. The neural network used here is shallow. A study of applying deep learning on this area is a good way ahead.

Contributions:

Varshaneya V has defined and run baseline models (3 classifiers) and checked for the accuracies and ROC of baseline models for both feaure extraction methods on all 7 assays. Nikhil Rai introduced SMOTE into the baseline models to tackle class imbalance and checked for accuracies and ROC of these models with SMOTE for both feaure extraction methods on all 7 assays. Both of them shared the results and collectively wrote the inference and conclusion.