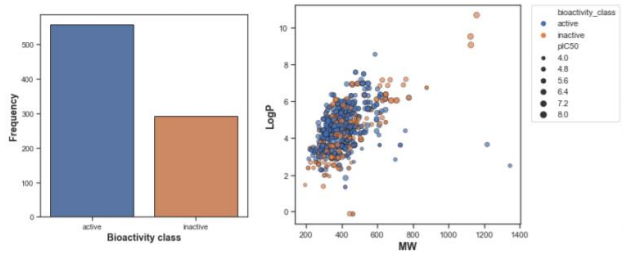
Project Report

Topic: Supervised machine learning for Quantitative Structure-Activity Relationship modeling (QSAR)

Introduction: QSAR is a technique that uses machine learning in order to learn the relationship between the chemical structure and the biological activity. Structure of these molecules are used to calculate molecular descriptors which essentially describe the physical & chemical properties that distinguish one molecule from the other in the dataset as binary representation showing absence or presence of the particular molecular feature. A collection of the molecular descriptors for all of the molecules constitutes the dataset(described below).

Dataset: The raw data was obtained from the ChEMBL Database(https://www.ebi.ac.uk/chembl/) with search keywords "Breast cancer (in all targets)" ; and filtered for organism type (Homo Sapiens) ; and target type(single protein), finally selecting, **Target ID:** CHEMBL5393. Bioactivity data for molecules that were reported as IC50 values in nM (nanomolar) unit were retrieved and saved to a raw file. The data was then cleaned for null values and labelling molecules as active (IC50 < 1000 nM), inactive (>5000 nM) and Intermediate (1000-5000 nM) under bioactivity_class. Combined the 3 columns from raw data (molecule_chembl_id,canonical_smiles,standard_value(IC50 values)) and bioactivity_class into a DataFrame [1357, 4] that was saved to a file. Molecular descriptors known as Lipinski descriptors (Molecular weight (MW) Dalton, Octanol-water partition coefficient (LogP), Hydrogen bond donors (NumHDonors) and Hydrogen bond acceptors (NumHAcceptors) were calculated as new features for the molecules using the rdkit package and combined with the above preprocessed data [1357, 8]. The IC50 values were normalized and converted to a negative logarithmic scale and relabelled as "pIC50" feature. The records for "Intermediate" labels were removed [850, 8].

Data Exploration: Data was explored using the Lipinski descriptors. Scatterplot of MW vs LogP confirmed balanced distribution of "Actives" and "Inactives". Among the 4 Lipinski's descriptors(MW, LogP, NumHDonors and NumHAcceptors), and pIC50 feature, only LogP and pIC50 showed ***statistically significant difference*** between **actives** and **inactives**, rest exhibited ***no difference*** between the **actives** and **inactives (Figure 1).**



Descriptor	Statistics	p	alpha	Interpretation
LogP	74720.5	0.023579	0.05	Different distribution (reject H0)
MW	80292.5	0.364795	0.05	Same distribution (fail to reject H0)
NumH-Acceptors	77293	0.106799	0.05	Same distribution (fail to reject H0)
NumHDonors	78172	0.146315	0.05	Same distribution (fail to reject H0)
pIC50	66904.5	9.17E-06	0.05	Different distribution (reject H0)

Figure 1

Model Building:

(a) Data Preparation: Non-essential and redundant features (molecule_chembl_id, canonical_smiles) were dropped from the dataset. Transformed bioactivity_class feature labels as 0(Inactive) and 1(Active). The features with object data types were converted to integer type. The final dataset had

6 features, namely, MW, LogP, NumHDonors, NumHAcceptors, pIC50 and the target (bioactivity_class). Highly correlated features with respect to the target were identified.

(b) Model complexity: ML agorithms considered were: Logistic Regression ● Random Forest ● Support Vector Machine ● KNeighbors ● Decision Tree ● XGBoost. 36 models were compared using five-fold cross-validation for different feature combinations in 6 categories (all features, Lipinski descriptors, top 4, top 3, and top 2 highly correlated features with respect to the target and the 2 statistically significant feature combination) for the six ML agorithms.

Model Evaluation: From the 5-fold cross validation, the average accuracy scores of train and test sets for each of the 36 models were compared **(Figure 2).** The accuracies for the 36 models were in the range 63.3-72.3%. As the number of features decreased, the accuracies of the models increased. Models with top3, Lipinski and 2 statistically combination gave better balance of accuracy and not being overfitted. In particular, Decision Tree models had lowest accuracy in comparison to other class of models.

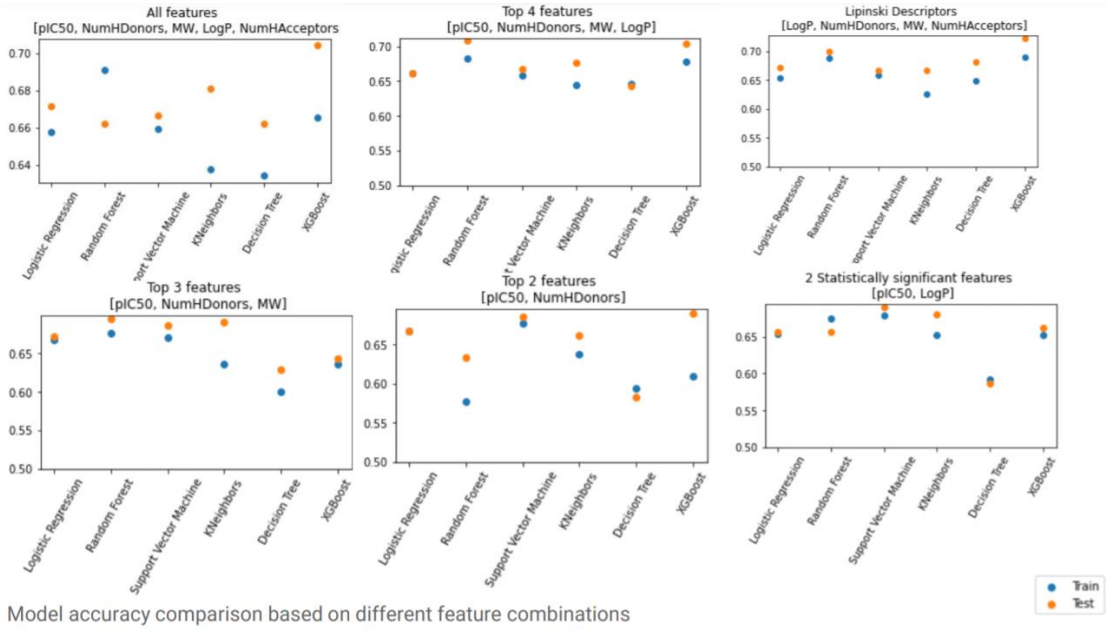


Figure 2

The models (all features) were also evaluated using the confusion matrices and ROC curves at different thresholds, which describe the performance of a classification model on a set of test data for which the true values are known. These were further used to calculate the metrics such as accuracy, precision, recall, F1-score and AUC (Area under the curve) for assessing the efficiency of classification by the models(**(Figure 3).** Based on these metrics, **XGB and LR** model performed the best among all models under consideration. Thus, it can further be used to predict bioactivity classification for a given set of records.

	LR	RFC	SVM	KNN	DT	XGB
Accuracy	0.6714	0.662	0.667	0.6808	0.6526	0.7042
Precision	0.6667	0.7329	0.6635	0.7338	0.75	0.773
Recall	1.0	0.7643	1.0	0.8071	0.7071	0.7786
F1-score	0.8	0.7483	0.7977	0.7687	0.7279	0.7758
AUC	0.83	0.62	0.83	0.64	0.62	0.67

Figure 3

Scope for Improvement: Neural network models could be tried as an alternative to model this dataset with improved accuracy.