Drug Activity Prediction

Name: Meenakshi Paryani **SJSU ID:** 011819392
Rank: 3
F1-Score: 0.8571

Problem Statement: Develop predictive models that can determine, given a particular compound, whether it is active (1) or not (0). The goal of this exercise was to develop the best binary classification model.

Approach: For this activity, various experiments were conducted using various **Classification** methods, **Ensemble** methods with various combinations of **dimensionality reduction algorithms** along with **Resampling** since it is class imbalanced data. There was a training data which was fed to the model which contained 800 drugs along with their activity and features effecting the activity. After trying all the experiments with different combinations of above parameters, the best model was arrived at. The resulting model was then used to compute the activities of the drugs in test dataset into **active** or **inactive** classes. On a high level, following steps were followed:

Pre-Processing: From the dataset, the goal was to get a sparse matrix format of the data which could then be fed to the dimensionality reduction algorithm for further pre-processing. Following steps were performed to get the compressed sparse matrix of the data

- Create the Coo Matrix – sparse matrix in coordinate format
    - Advantage – facilitates duplicate values and fast conversion to other sparse formats
- Create the CSC Matrix – from the Coo matrix, create the compressed sparse column matrix
- Create the CSR Matrix – convert the CSC matrix to the compressed row matrix

Resampling: The dataset provided contains 800 drug activities out of which, there were only 78 actives (+1) and 722 inactive (0). This distribution is referred to as **Class-Imbalance**. To deal with the class imbalance, SMOTE (Over Sampling technique) was applied to increase the samples of the minority-class to have equal distribution of minority vs majority so that there is no bias between the classes while classification.

PS: It was noticed that the F1-Score reduced after applying SMOTE so further experiments were performed without oversampling.

Dimensionality Reduction: As the data uses was high-dimensional in nature, tried to reduce the dimensionality of the data using two algorithms **SparsePCA** and **TruncatedSVD** to facilitate faster development

Classification: As a final step, to classify the test data drug records as active/inactive, used following algorithms:

- KNN
- Perceptron
- Random Forest
- Extra Trees Classification

F-1 score for Combinations that worked well:

Classification Algorithm	Classification Parameters	Dimensionality Reduction Algorithm	DR parameters	F1-Score
Random Forest	NA	Truncated SVD	Comp=8, Random_state = 42	0.75
KNN	K = 3(best)	Truncated SVD	Comp=8, Random_state = 42	0.667
Extra Trees	NA	Truncated SVD	Comp=8, Random_state = 42	0.111
Perceptron	NA	Truncated SVD	Comp = 7, random_state = 42	0.711
Perceptron	NA	Truncated SVD	Comp=8/9, Random_state = 42	**0.8571**
Perceptron	NA	Truncated SVD	Comp = 10, Random_state = 42	0.73

Summary: Maximum F1-score of **0.8571** was achieved with Truncated SVD (number of components=8 and random_state=42) and Perceptron algorithm without resampling

Leaderboard:

Rank	F1 (on 50%)	User ID	Submission Count
1	0.8571	11545378	10
2	0.8571	11812788	30
3	0.8571	11819392	9