Drug Activity Prediction

Project Report

Project By: -

Ashfaque Jamal

Beauty Atara

Table Of Content:

Introduction:

Drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. The first step in the discovery of a new drug is usually to identify and isolate the receptor to which it should bind, followed by testing many small molecules for their ability to bind to the target site. This leaves researchers with the task of determining what separates the active (binding) compounds from the inactive (non-binding) ones. Such a determination can then be used in the design of new compounds that not only bind, but also have all the other properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc.).

The goal of this project is to allow us to develop predictive models that can determine given a particular compound whether it is active (1) or not (0). As such, the goal would be developing the best binary classification model.

A molecule can be represented by 100000 binary features which represent their topological shapes and other characteristics important for binding.

Objective:

The objective of this project is to allow us to develop predictive models that can determine given a particular compound whether it is active (1) or not (0). As such, the goal would be developing the best binary classification model.

Approach:

1. **Data Collection:**

   The collected data was in two parts, training data and testing data. The data was in .dat format, so first of all we converted it into .csv format. The training data was containing 800 rows and the testing data was containing 350 rows. The data was then expanded to a dimension of 800*1000000 and 350*1000000 for training and testing data respectively.

2. **Reading the data:**

   The data was read with the help of pandas library.

3. **EDA:**

   In Exploratory Data Analysis, we checked whether the data is having any null values or not. After that we also checked for any outliers present in the data.

4. **Data Pre-processing:**

   The first step in data pre-processing was to deal with the null values. Our data contained many null values, so we kept only those columns for which the number of null values is less than 15 % of the total values in column. Then after, the null values were replaced with the median of the specific feature. After dealing with null values, the next step was to scale the data. The data was then standardised using standard scaler. The last step in data pre-processing was applying One Hot Encoding in order to expand the data and convert it into sparse matrix.

5. **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.

6. **Dimensionality Reduction:**

   As the data uses was high-dimensional in nature, we tried to reduce the dimensionality of the data using the algorithm of PCA to facilitate faster development.

7. **Model Building:**

   The data was now ready to be fed to the classification model. The following classification models were used to make the predictions:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Perceptron
- Naive Bayes Classifier
- K Nearest Neighbours
- Support Vector Classifier

8. **Hyperparameter tuning**:

Using GridSearchCV the hyperparameters for each of the model were tuned, and those parameters for which the best results were obtained, were used for final model building.

9. **Model Evaluation**:

All the models were then evaluated on the test data. Since the data was imbalanced, the main focus was on the precision, recall and the f1 score for evaluating the model.

10. **Finalizing Model:**

After trying all the experiments with different combinations of all 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.

Deployment: Finally, the project was deployed using streamlit

Observation Table:

Classification algorithm	One Hot Encoding	Dimensionality Reduction technique	Resampling technique	Components of DR	F1 Score on test data
Decision Tree	NA	PCA	SMOTE	100	0.579
Gaussian NB	NA	PCA	SMOTE	350	0.532
Gaussian Naive Bayes	NA	PCA	NA	100	0.545
Decision Tree	NA	NA	SMOTE	100	0.549
SVC	YES	PCA	SMOTE	100	0.529
Random Forest	YES	PCA	SMOTE	350	0.504

.

Conclusion:

Maximum f1 score was 0.579, obtained from Decision Tree Classifier with PCA (n_components=100) and with resampling.