

Report

We transform the drug usage columns into a binary class where 0 represents a non-user and 1 represents a user. The source data has 12 features, and we select the first 6 ranking features for each drug through the Recursive Feature Elimination method, where according to the chosen drug we rank the features using the Logistic Regression model, and the least informative feature is NScore which is not considered in any drug, on the other hand, the most informative one is Age. The 4 models are trained and tuned using the GridSearchCV library according to the training data. After training SVC and Random Forest turn out to be better models in terms of accuracy and precision. We also use ROC curves to compare the 4 models for each drug. In the classes of alcohol and chocolate, the data has a sample imbalance problem which trains the model in a way where it gives a lot of false positives and specificity = 0 and does the opposite for VSA. Except for Coke, we get an accuracy of more than 70% for each drug.

A few of the features selected for each drug overlap with the paper that worked on the same dataset. But the paper uses the correlation of features to select a specific drug which is better than backward elimination as it uses the target variable for selection which fits the model more towards training data. Our SVC and random forest model give very close sensitivity and specificity percentages to the paper evaluations, and the Leave One Out Cross Validation method used in the paper affects the model to classify the users. Even though using Decision Tree for VSA drug we do not get near results to the paper, which means feature selection and tuning lacks for VSA.