CS 584: THEORY AND APPLICATIONS OF DATA MINING
HW2: DRUG ACTIVITY PREDICTION USING VRAIOUS CLASSIFIERS

NAME: SRIPATH CHERUKURI
MINER2 USER-ID: MAVERICK
F1-SCORE SCORE: 0.64
RANK: 440
G-NO: G01395231

PROBLEM STATEMENT
Using various classifiers predict whether the drug is active (1) or in-active (0) based on the given set of features in the sparse data.

1. INTRODUCTION
The classifiers that were chosen are perceptron and decision tree classifier. By experimenting with the two classifiers mentioned above we are going to find out which is better than the other.

2. APPROACH FOLLOWED IN THE IMPLEMENTATION OF THIS SOLUTION
There are various steps to be followed in this implementation. The first step is data cleaning followed by test, train split and then dimensionality reduction. Then after we perform over-sampling and then finally predict the labels for test data.

2.1 CLEANING DATA
As the data is sparse data, we first separate the class labels from the features and store the values respectively. Also, we remove the escape characters to make data free from any unwanted white spaces. Then, we create a mapping table for all the features given in the data set.

2.2. TRAIN, TEST SPLIT
In this step we split the data into train and test parts using an 80/20 split ratio keeping more part of data for training the classifier and the remaining for the test part and evaluation of the model.

2.3 DIMENSIONALITY REDUCTION
Here we pass the data after split for dimensionality reduction as the training data contains nearly 1,00,000 features. So, for reducing the complexity of the program and quick execution we reduce the data using PCA dimensionality reduction technique. PCA is principal component analysis which helps reduce the dimensions of the data to given number of components.

2.4 OVER-SAMPLING
The data set being imbalanced performing the over-sampling on the data is a must, otherwise the classifier performs bad and ignores the minority class. So, using over-sampling balances both the majority and minority classes. Imbalanced data here means that one class label dominates the other class label. SMOTE is used here to over-sample the imbalanced data.

3. PARAMETERS

There are various parameters that affect the result of the classifier, such parameters are considered mentioned below:

- **train, test split = 80/20** ratio, helps to keep more part of data for training, so classifier could train well on more samples.
- **n-components in dimensionality reduction = 500**, randomly chosen number to reduce the data, this shouldn't be either too large or too small.

4. ANALYSIS & METRICS

4.1 EXECUTION TIMES

Execution times for perceptron and decision tree classifier for both train and test data can be seen below:

CLASSIFIER/TIME IN SECONDS	Training Time	Predictions Time	Test Data (predictions) Time
Perceptron	0.02 s	0.003 s	0.002 s
Decision Tree Classifier	0.461 s	0.001 s	0.001 s

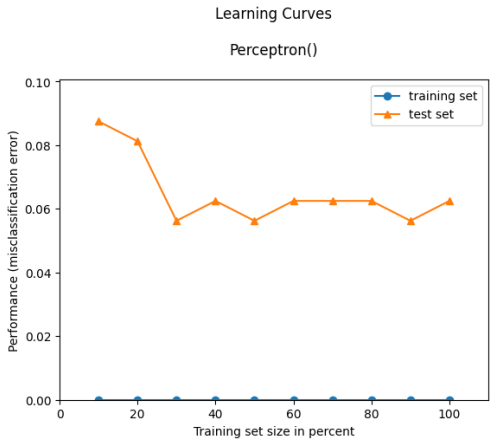
4.2 F1-SCORE, PRECISION, RECALL

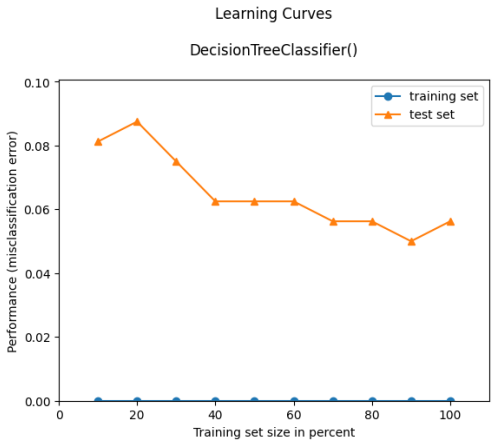
The F1-score, precision and recall values for both perceptron and decision tree classifier can be seen below:

CLASSIFIER	F1-score	Precision	Recall
Perceptron	0.44	0.29	1.00
Decision Tree Classifier	0.67	0.71	0.62

4.3 LEARNING CURVE

Learning curves for perceptron and decision tree classifier can be seen below:





5. CONCLUSION

As decision tree classifier achieved more F1-score and miner score than the perceptron from the above results, we can say that decision tree is better than perceptron.