Overview of Methodology :

I used the following method across all models:

1. Split the columns to separate input and output features.
2. Select 6 features (drugs) that are to be analyzed.
3. Combine classes as suggested, "CL0" and "CL1" as "*Non-Users*" (0) and the rest as "*Users*" (1).
4. Apply feature selection method, SelectKBest(), for extracting best features.
5. Split the dataset into training (67%) and testing (33%) data.
6. Train the model and predict classes for the testing data.
7. Compare the predicted classes with the actual classes using the specified evaluation techniques.
8. Plot the confusion matrix and ROC curve.

The following table shows the difference between the models, as in the original paper and my study.

Differences ——— Models	Original Paper	Current Study
KNN	**Metrics**: Euclidian distance, Fisher's transformed distance, and the adaptive distance **Additional**: Weighted voting kernel	**Metrics**: Minkowski distance **Additional**: GridSearchCV() to calculate the optimum value for '*k*', to set the *n_neighbors* parameter.
Decision Tree	**Metrics**: Information gain, Gini gain, and DKM gain **Additional**: Fisher's discriminant and Pruning techniques to improve the tree.	**Metrics**: Gini gain **Additional:** Removed feature selection method SelectKBest() to use all available features for better decision making.
SVM	N/A	**Metrics**: '*rbf* ' kernel selected with default degree set as '*3*'. **Additional**: N/A
Random Forest	**Metrics**: 2,048 RF models per drug	**Metrics:** Default values of '*100*' used in n_estimators.

Results :

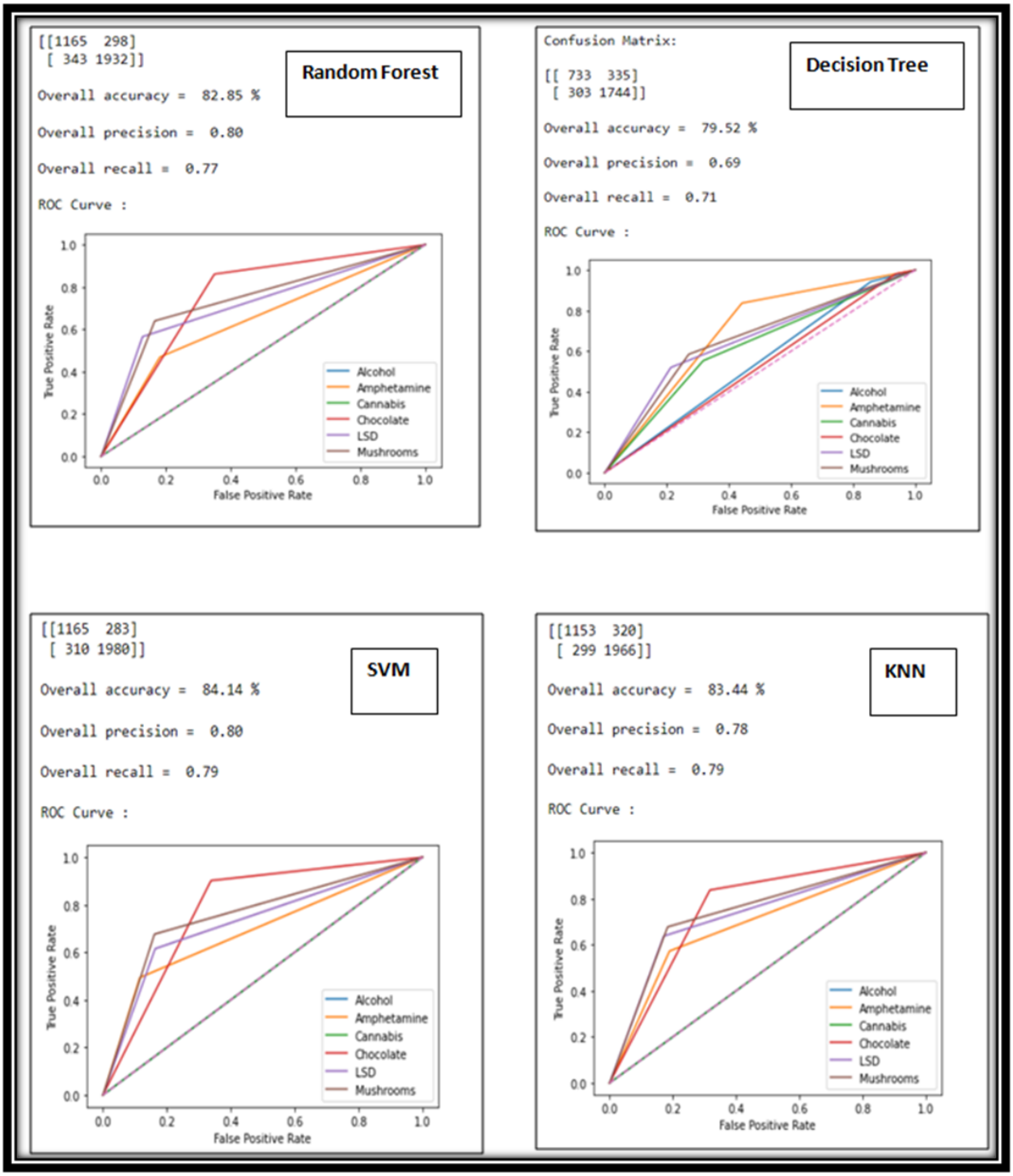
Classifiers / Drugs	Decision Trees	Random Forests	SVM	KNN
Alcohol	Precision = 0.97 Recall = 0.94 Accuracy = 92.45%	Precision = 0.96 Recall = 0.99 Accuracy = 95.82%	Precision = 0.96 Recall = 1.0 Accuracy = 96.62%	Precision = 0.96 Recall = 1.0 Accuracy = 96.95%
Amphetamines	Precision = 0.80 Recall = 0.83 Accuracy = 74.79%	Precision = 0.61 Recall = 0.56 Accuracy = 72.07%	Precision = 0.66 Recall = 0.54 Accuracy = 71.42%	Precision = 0.59 Recall = 0.54 Accuracy = 70.3%
Cannabis	Precision = 0.49 Recall = 0.55 Accuracy = 63.56%	Precision = 0.98 Recall = 1.0 Accuracy = 98.71%	Precision = 0.97 Recall = 1.0 Accuracy = 97.75%	Precision = 0.97 Recall = 1.0 Accuracy = 97.59%
Chocolate	Precision = 0.97 Recall = 0.98 Accuracy = 95.82%	Precision = 0.86 Recall = 0.86 Accuracy = 81.21%	Precision = 0.83 Recall = 0.89 Accuracy = 81.38%	Precision = 0.84 Recall = 0.85 Accuracy = 79.45%
LSD	Precision = 0.5 Recall = 0.51 Accuracy = 70.94%	Precision = 0.68 Recall = 0.59 Accuracy = 80.09%	Precision = 0.69 Recall = 0.57 Accuracy = 78.65%	Precision = 0.63 Recall = 0.63 Accuracy = 78.33%
Mushrooms	Precision = 0.54 Recall = 0.58 Accuracy = 67.73%	Precision = 0.70 Recall = 0.59 Accuracy = 74.95%	Precision = 0.68 Recall = 0.69 Accuracy = 77.04%	Precision = 0.64 Recall = 0.67 Accuracy = 75.12%

After looking at the results generated by the models, I will take a few examples to understand the performance and the values of the metrics to highlight the learnings from this study.

1. Common substances such as Alcohol and Cannabis have high values for all the metrics. The high values are due to the imbalance in the training data, since people have higher tendencies to use these drugs most of the data points were classified as *Users* and the same trend is followed in test data as well.

2. Less common drugs such as LSD and Amphetamines have lower values of precision & recall, this is due to the balance between both classes in the training data.

3. The original paper uses its results to find out patterns & correlations between character traits, age, ethnicity, etc (input features) and the tendency of drug usage. While I intend to see the quality of the models only.



4. Based on the overall analysis of each of the 4 models, Decision Trees seems to be the least accurate which might be due to the lack of input features, skewness of the data (for categories where individual accuracies are high) and can be improved by hyper-tuning the parameters and analyzing metrics such as gini index, entropy or information gain.

5. The other 3 models show similar results overall and for individual drugs as well. These accuracies can be further optimized with better values for parameters and feature selection methods. The scope of this study does not intend to understand the effects that fine-tuning these parameters might have on the accuracy of the model.