Edwin Thomas (Student Number: 300278402)

CSI 5155: Machine Learning

Assignment 1: Drug Consumption Dataset Report

In this assignment, the topic of binary classification on the drug consumption dataset is explored for six different drug types using grouping (Decision Tree(DT), Random Forests) and grading (KNN, SVM) classifiers. To assess the impact of features on the classification task, a feature reduction technique that uses ANOVA f-score is employed, which is suitable for continuous features and categorical outputs. Experimental evaluation shows that reduction of feature set size reduced performance by ~2.5% (Appendix A,B).

Reduction of the multi-class classification problem to two classes (users/non-users) resulted in highly imbalanced datasets. Models trained on default sklearn hyperparameters yielded biased scores, with significantly higher sensitivity when proportion of users are higher and vice-versa for the opposite scenario. To reduce this impact of overfitting on the imbalanced distribution, an exhaustive grid search for hyper-parameters (HP) is performed for each model. For the grouping models, increase in minimum-samples-per-leaf and class-frequency-weighted impurity scores reduced overfitting and yielded superior results on the test set. Similar trends are observed in SVM when regularization factor is (>1.0) and nearest neighbors (k>15) for KNN. The best performing model is DT with Sensitivity and Specificity scores above 75% on average, like [1].

Methodologies used in [1] are similar to the ones followed in this assignment, in terms of the HP search techniques. However, they differ w.r.t evaluation (Leave-one-out-cross-validation as opposed to stratified hold out test sets) and feature selection methods (best model among all feature subsets vs feature selection prior model optimization), and additional HP considerations such as Adaptive distances for KNN. Table 1, summarizes and compares the results. The performance gap can be attributed to the inherent imbalance in datasets. Features such as Age, Gender, OScore, Imp and SS are found to have the most intersections with chosen features in [1] on average for the six classes (Table 2).

Edwin Thomas (Student Number: 300278402)

Results: When top six feature are chosen (k=6)

	KNN		Decision Tree		Random Forest		SVC		Reference [1]	
Drug Type	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
Amphet	52.68	80.15	74.11	67.59	61.61	67.09	59.38	72.86	81.3	71.48
Cannabis	86.84	67.16	77.51	79.9	75.6	81.37	86.12	71.57	79.29	80
Ecstacy	70.56	77.01	74.6	75.13	76.61	71.66	70.16	78.88	76.17	77.16
LSD	69.02	84.7	79.35	79.22	78.26	76.71	72.83	84.02	85.46	77.56
Mushrooms	72.05	83.97	75.98	78.12	80.79	74.55	75.98	81.42	65.56	94.79
VSA	28.95	88.83	73.68	73.81	77.63	65.93	23.68	90.29	83.48	77.64
Avg	63.35	80.30	75.87	75.63	75.08	72.89	64.69	79.84	78.54	79.77

Table 1. Comparison of grouping and grading model performances and previous approaches. The colors in the cells indicate relative performance with other algorithms in the same row (from RED (least) to GREEN (best)).

Drug Type	Chosen Features	Best Hyperparameters	Best Model
Amphet	**Age Gender** Country Oscore Impulsive **SS**	{'class_weight': 'balanced', 'criterion': 'gini', 'max_features': 'sqrt', 'min_samples_leaf': 18, 'splitter': 'best'}	DT
Cannabis	**Age** Country **Oscore Cscore Impulsive** SS	{'class_weight': 'balanced', 'criterion': 'gini', 'max_features': None, 'min_samples_leaf': 17, 'splitter': 'random'}	DT
Ecstacy	**Age Gender** Country Oscore Impulsive **SS**	{'class_weight': 'balanced', 'criterion': 'gini', 'max_features': None, 'min_samples_leaf': 23, 'splitter': 'best'}	DT
LSD	**Age** Gender Country **Oscore Impulsive** SS	{'class_weight': 'balanced', 'criterion': 'gini', 'max_features': None, 'min_samples_leaf': 12, 'splitter': 'random'}	DT
Mushrooms	Age **Gender** Country Oscore Impulsive SS	{'class_weight': 'balanced', 'criterion': 'entropy', 'max_features': 'log2', 'min_samples_leaf': 26, 'splitter': 'best'}	DT
VSA	**Age Education** Country **Cscore** Impulsive **SS**	{'class_weight': 'balanced', 'criterion': 'entropy', 'max_features': 'sqrt', 'min_samples_leaf': 27, 'splitter': 'best'}	DT

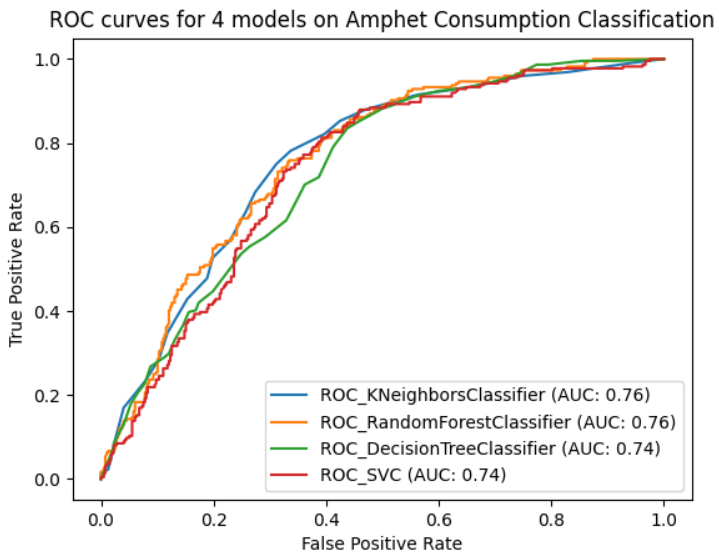
Table 2. Details of the best model for each dataset and best set of hyper-parameters that balances performance in each class. Features highlighted in BOLD correspond to intersection of chosen features in [1].

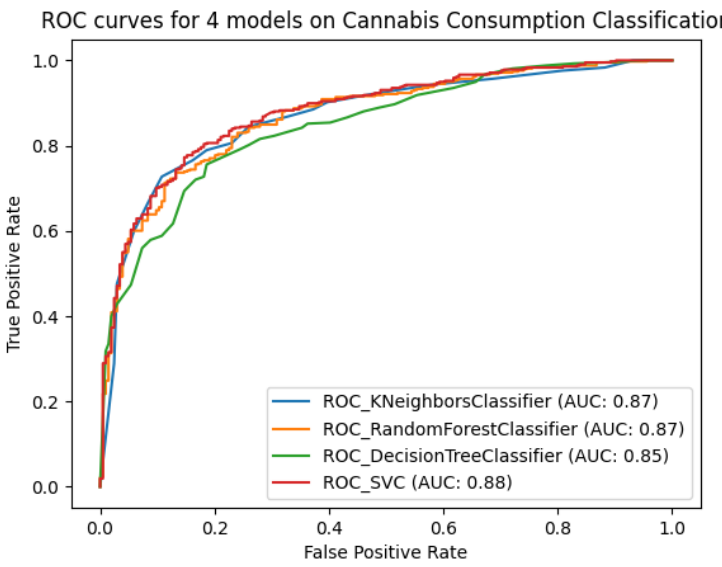
Edwin Thomas (Student Number: 300278402)

	KNN			Decision Tree			Random Forest			SVC		
Drug Type	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Amphet	59.9	52.68	56.06	56.27	74.11	63.97	51.3	61.61	55.98	55.19	59.38	57.2
Cannabis	84.42	86.84	85.61	88.77	77.51	82.76	89.27	75.6	81.87	86.12	86.12	86.12
Ecstacy	67.05	70.56	68.76	66.55	74.6	70.34	64.19	76.61	69.85	68.77	70.16	69.46
LSD	65.46	69.02	67.2	61.6	79.35	69.36	58.54	78.26	66.98	65.69	72.83	69.07
Mushrooms	72.37	72.05	72.21	66.92	75.98	71.17	64.91	80.79	71.98	70.45	75.98	73.11
VSA	26.51	28.95	27.67	28.14	73.68	40.73	24.08	77.63	36.76	25.35	23.68	24.49
Avg	62.62	63.35	62.92	61.38	75.87	66.39	58.72	75.08	63.90	61.93	64.69	63.24

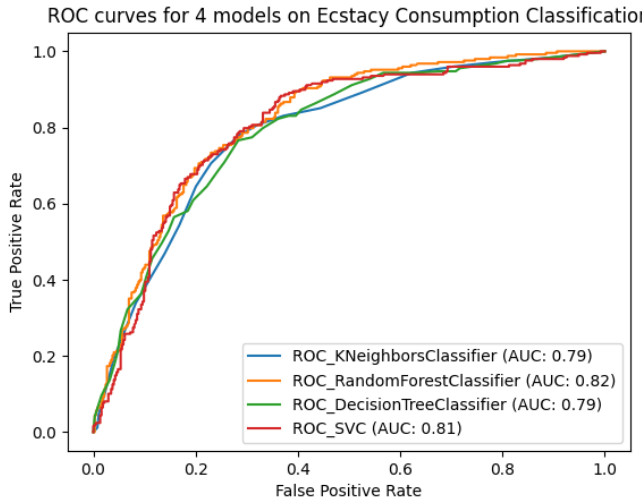
Table 3. Precision, Recall and F1 measures in (%) corresponding to the models under consideration.

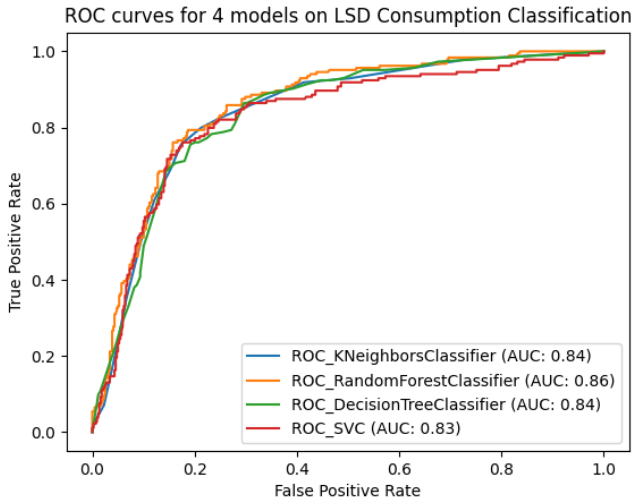
Table 4. ROC Curves obtained on the best models for the 6 Drug Types

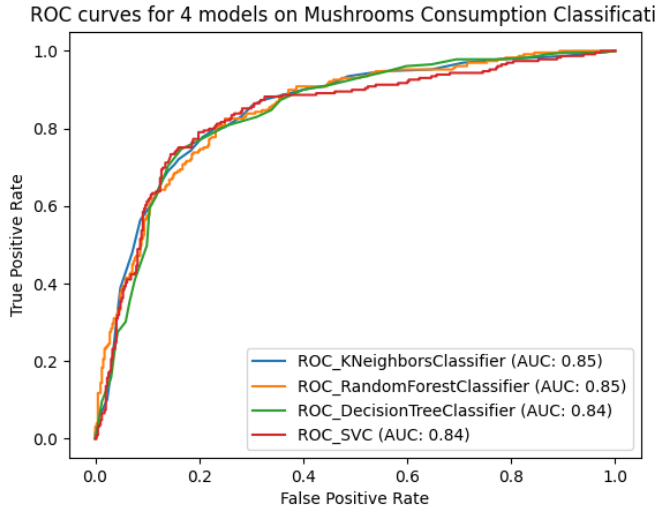


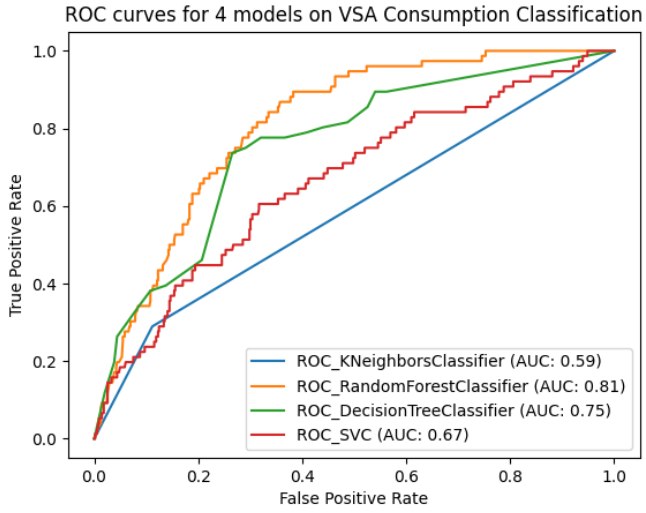


Edwin Thomas (Student Number: 300278402)









Edwin Thomas (Student Number: 300278402)

Appendix-A: Results obtained when when (k=12) or 12 features

	KNN		Decision Tree		Random Forest		SVC		Reference [1]	
	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
Amphet	57.14	79.4	73.66	69.35	78.57	59.8	55.36	77.39	81.3	71.48
Cannabis	86.12	67.16	82.06	75.49	67.7	84.8	87.32	69.61	79.29	80
Ecstacy	70.16	78.07	70.16	78.34	75.81	65.24	68.15	78.07	76.17	77.16
LSD	64.67	85.84	76.09	78.54	74.46	76.94	62.5	86.99	85.46	77.56
Mushrooms	74.24	80.41	80.79	76.08	80.79	69.72	72.49	78.88	65.56	94.79
VSA	34.21	90.48	84.21	74.54	68.42	74.36	32.89	92.12	83.48	77.64
Avg	64.42	80.23	77.83	75.39	74.29	71.81	63.12	80.51	78.54	79.77

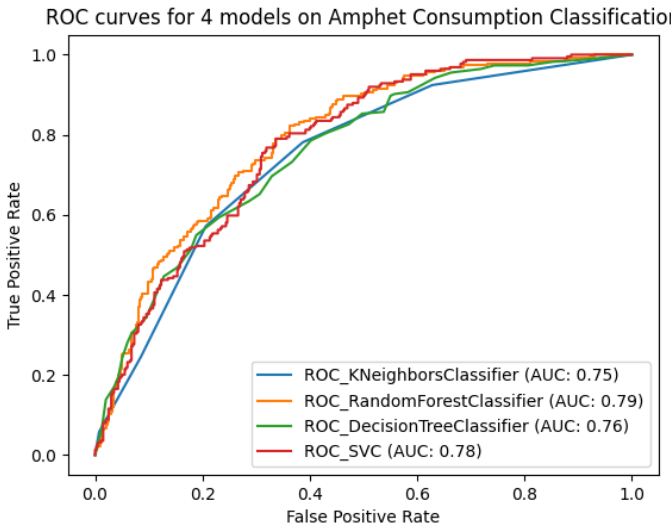
Table 5. Comparison of grouping and grading model performances and previous approaches. The colors in the cells indicate relative performance with other algorithms in the same row (from RED (least) to GREEN (best)).

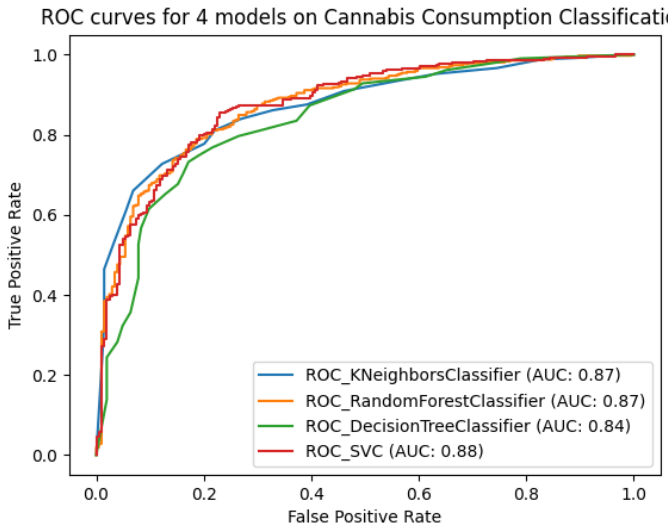
	KNN			Decision Tree			Random Forest			SVC		
Drug Type	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Amphet	60.95	57.14	58.99	57.49	73.66	64.58	52.38	78.57	62.86	57.94	55.36	56.62
Cannabis	84.31	86.12	85.21	87.28	82.06	84.59	90.13	67.7	77.32	85.48	87.32	86.39
Ecstacy	67.97	70.16	69.05	68.24	70.16	69.18	59.12	75.81	66.43	67.33	68.15	67.74
LSD	65.75	64.67	65.21	59.83	76.09	66.99	57.56	74.46	64.93	66.86	62.5	64.61
Mushrooms	68.83	74.24	71.43	66.31	80.79	72.83	60.86	80.79	69.42	66.67	72.49	69.46
VSA	33.33	34.21	33.77	31.53	84.21	45.88	27.08	68.42	38.81	36.76	32.89	34.72
Avg	63.52	64.42	63.94	61.78	77.83	67.34	57.86	74.29	63.30	63.51	63.12	63.26

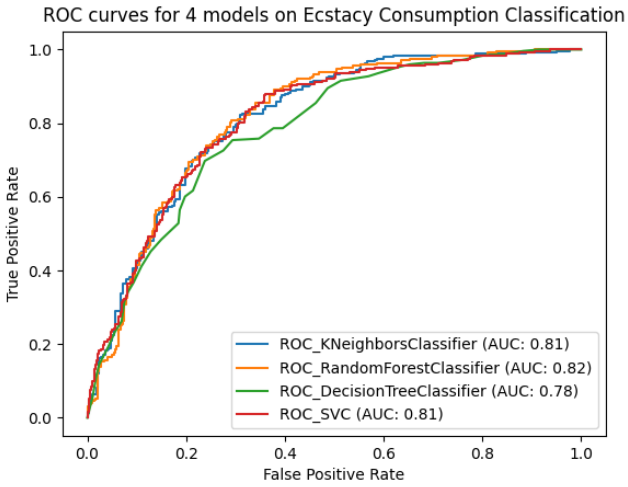
Table 6. Precision, Recall and F1 measures in (%) corresponding to the models under consideration.

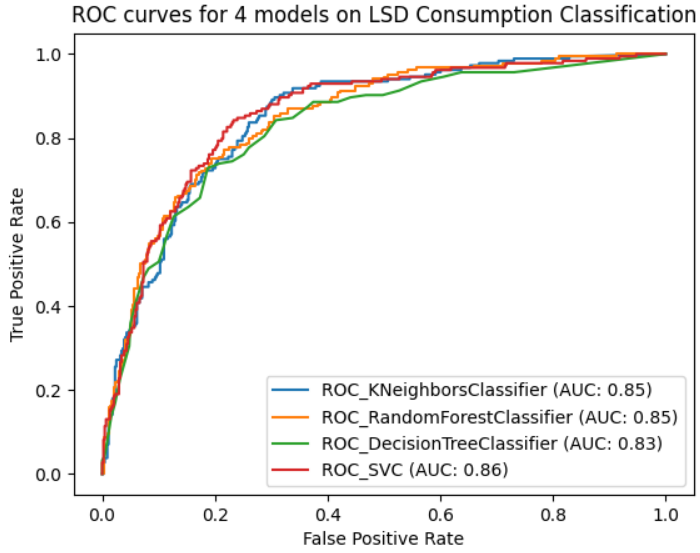
Edwin Thomas (Student Number: 300278402)

Table 7. ROC Curves obtained on the best models for the 6 Drug Types

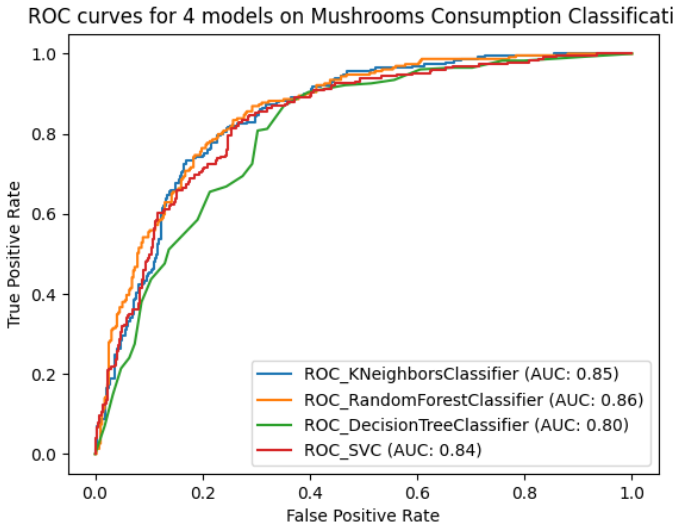


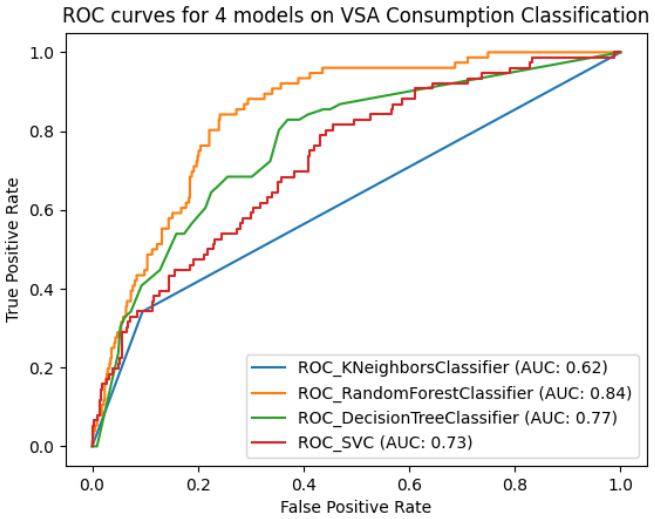






Edwin Thomas (Student Number: 300278402)

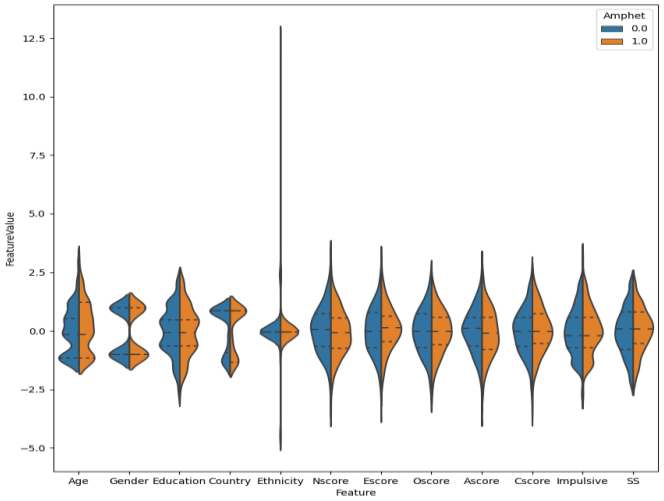


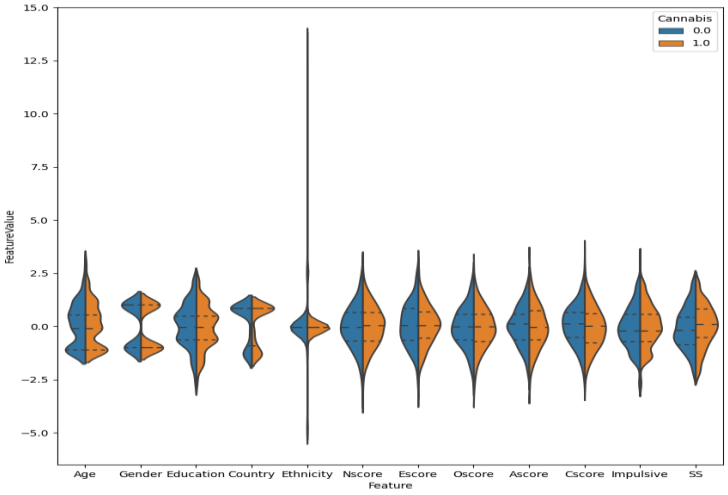


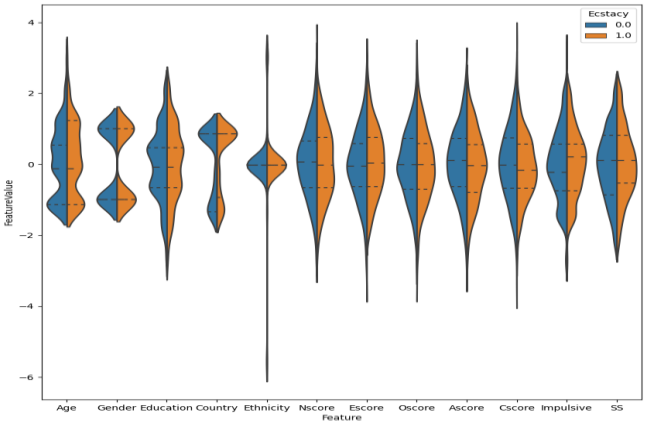
Appendix B (Feature Selection Analysis with Violin Plots: choosing value of k (features))

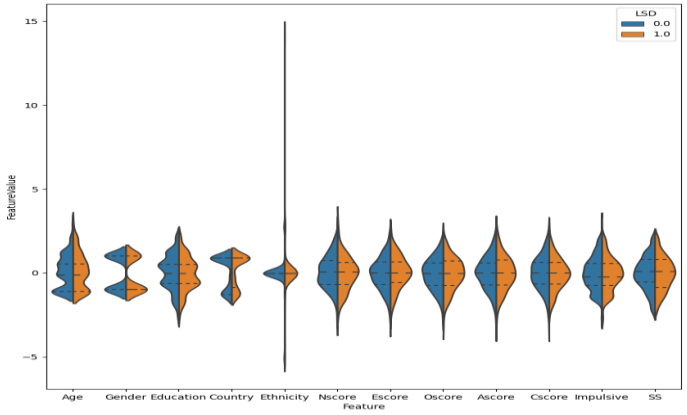
Table 8. Violin Plots with Feature Analysis of each drug type. The results indicate a similar per-class distribution of feature values, with overlapping medians and interquartile ranges. This makes it difficult to predict which feature has larger impact on the target variable. However, some features such as Ethnicity can be ruled out due to large spread and low peak which results in 3 feature groups to select from namely {Age, Gender, Education, Country}, {N, E, O, A, C scores}, {Impulsivity, SS}. When k=6 (~2-3 features from each bracket) we can get a good approximation (trade-off) in performance w.r.t to k=12 as observed in Table 1 and 5.

Edwin Thomas (Student Number: 300278402)

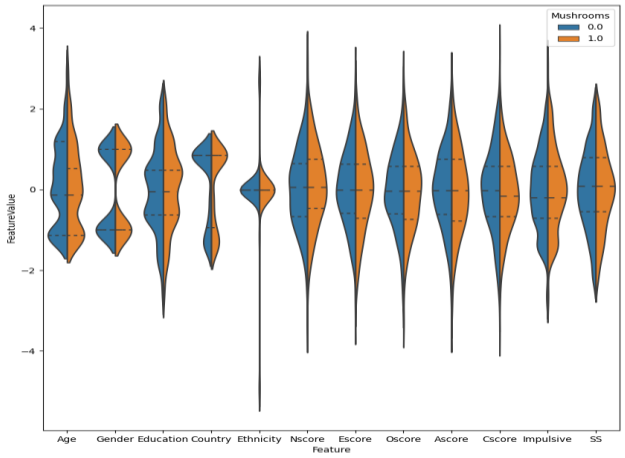


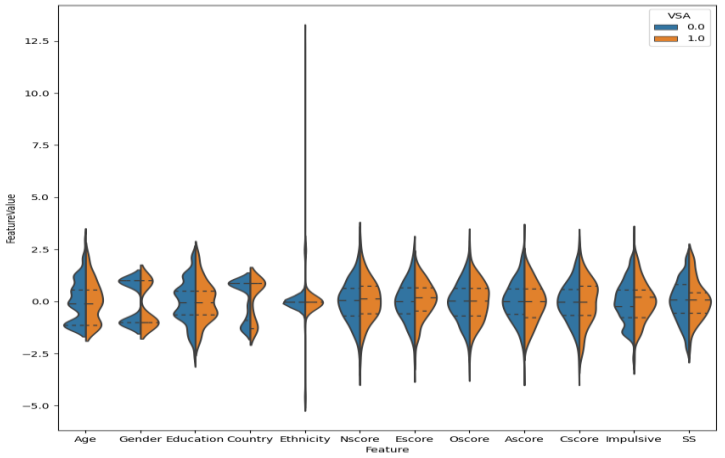






Edwin Thomas (Student Number: 300278402)





References

[1] E. Fehrman, V. Egan, A. N. Gorban, J. Levesley, E. M. Mirkes, A. K. Muhammad,"Personality Traits and Drug Consumption. A Story Told by Data." Springer, Cham, 2019. ISBN 978-3-030-10441-2