CSI 5155 Assignment 1 Report

Name: Zhikun (Sam) Yuen Student ID: 300323972

The seven labels I picked: Legal highs (Legalh), Ecstasy, Cannabis, Nicotine, Caffeine (Caff), LSD, VSA.

Part A:

In data analysis, there are no missing value. In the 7 selected labels, all of them are not balance for the two classes and two of them are extremely imbalance, Caff and VSA (refer to appendix).

Although all the 12 features are categorial features but some of them are ordinal features (categories with ordering) and some of them are nominal features (no ordering). For the final result, I turn the nominal features, Country and Ethnicity, into one-hot encoding, and normalize the ordinal features to a range between 0 to 1.

Actually, I try to use 3 different ways to transform the data 1. one-hot encoding for all categorial features, 2. use the raw features and 3. One-hot encoding for the nominal features only and normalizing the ordinal features with min-max scaler. The results are similar for all these 3 transformation.

For all 7 labels. Decision tree is the worst among the 4 models in all metrics (precision, recall and AUC), except for VSA's recall. The AUC scores of random forest, SVM and KNN are similar if the data is not extreme imbalance.

I select two extremely imbalance classes, Caff and VSA among the 7 labels. If class 1 is the majority (Caff), both precision and recall is close to 1 but the AUC is low for all models. If class 1 is the minority (VSA), the recall is very low. There are an interesting observation for VSA in KNN. It predicts all test data as non-user (class 0), i.e. there is no true positive prediction, so both precision and recall of KNN are 0 in this case. However, its AUC is pretty high (0.81). This shows that AUC cannot reflect the results if the data is negative examples (class 0) dominate the majority of the data, i.e. high AUC in imbalance data refers to good ordering of the prediction scores/probability during AUC's computation but may not refer to good performance. For imbalance data, AUC score is more confident for low score but not high confident for high score.

Part B:

The raw dataset is cleaned already. After the quantification for both ordinal and nominal features, they didn't use any normalization approaches to normalize the range of these two kinds of features. Then, they kept the ordinal features and use dummy coding for the nominal features. In my feature engineering part, I transform the nominal features (country and Ethnicity) to one-hot encoding instead of dummy encoding and normalize the all ordinal features to [0, 1] with min-max scaler. The similar thing in here is that the paper and I keep both ordinal features as numeric instead of one-hot encoding.

In feature selection, they use sparse PCA and found that all the ordinal features have higher principal variable ranking and Double Kaiser's ranking. Only the two countries. UK and USA are informative in their sample, so they remove all nominal features. Instead of using PCA, I use the univariate feature selection in sklearn to do feature selection. The univariate feature selector will return the best k features based on univariate statistical tests. I use random search to find the optimal k between [10, 24]. For the score function of the univariate feature selector,

I try both chi2 stats and mutual information, but the results are similar. In my feature selection result, the selector will also select some nominal features for both country (not just USA and UK) and Ethnicity, which is different from the paper.

For the models, they tried KNN, DT, LDA, GM, PDFE, LR, NB and RF. They use sensitivity and specificity with LOOCV to qualify the classifiers. In my case, I use a train set and test set to quality DT, RF, SVM and KNN. In their metrics, DT achieves the best results in six out of seven in the seven selected labels. The other one best classifier in Caff is KNN. However, in my results, DT is usually the worst classifier among the four in precision, recall and AUC.

Appendix:

	Decision Tree	Random Forest	SVM	KNN	Class ratio: 0 vs 1
Legalh	P: 0.75 R: 0.61 A: 0.78	**P: 0.78** R: 0.67 **A: 0.86**	P: 0.80 **R: 0.68** A: 0.85	P: 0.78 R: 0.65 A: 0.85	Train: 1.58:1 Test: 1.28:1
Ecstasy	P: 0.60 R: 0.58 A: 0.69	P: 0.64 R: 0.65 **A: 0.81**	P: 0.65 **R: 0.67** **A: 0.81**	**P: 0.68** R: 0.65 A: 0.80	Train: 1.53:1 Test: 1.47:1
Cannabis	P: 0.85 R: 0.78 A: 0.77	P: 0.87 R: 0.87 **A: 0.89**	P: 0.88 **R: 0.88** **A: 0.89**	**P: 0.89** R: 0.84 A: 0.88	Train: 0.52:1 Test: 0.43:1
Nicotine	P: 0.73 R: 0.66 A: 0.61	P: 0.74 R: 0.89 **A: 0.75**	P: 0.66 R: 1.0 **A: 0.75**	P: 0.76 R: 0.84 A: 0.73	Train: 0.48:1 Test: 0.50:1
Caff	**P: 0.98** R: 0.97 A: 0.55	**P: 0.98** **R: 1.0** **A: 0.73**	**P: 0.98** **R: 1.0** A: 0.59	**P: 0.98** **R: 1.0** A: 0.62	Train: 0.02:1 Test: 0.02:1 **Extremely Imbalance**
LSD	P: 0.55 R: 0.54 A: 0.72	P: 0.60 R: 0.61 **A: 0.81**	**P: 0.61** R: 0.62 **A: 0.81**	**P: 0.61** **R: 0.66** A: 0.79	Train: 2.44:1 Test: 2.28:1
VSA	P: 0.36 **R: 0.33** A: 0.68	**P: 0.67** R: 0.07 A: 0.80	P: 0.33 R: 0.17 A: 0.72	P: 0 R: 0 **A: 0.81**	Train: 7.47:1 Test: 6.69:1 **Extremely Imbalance**