University of Ottawa

CSI 5155 Machine Learning

Assignment 1 – Report

Lecturer: Herna Viktor (hviktor@uottawa.ca)

Student: Kelvin Mock (kmock073@uOttawa.ca)

Student ID: 300453668

GitHub: https://github.com/kmock930/Drug-Consumption-Machine-Learning-analysis.git

Contents

Introduction

In the given classification problem, I applied a machine learning pipeline with 6 different classifiers on 2 datasets respectively based on the following approaches: classifying with fined tuned parameters (from the *Random Search* algorithm), classifying with under-sampled data, classifying with over-sampled data, and classifying with a combination of data from both sampling methods.

In this report, I am going to show their performances with *confusion matrices*, some insightful metrics (including the *recall* and the *precision*) of each classifier, as well as *Receiver-Operating Characteristic curves (ROCs)* which show the performances of each classifier in terms of an *Area Under Curve (AUC)*.

Models Evaluation

Classification without sampling

Chocolate Dataset

Decision Tree classifier

Predicted	user	non-user	All
Actual			
user	621.0	0.0	621.0
non-user	8.0	0.0	8.0
All	629.0	0.0	629.0

Precision: 0.9872813990461049
Recall: 1.0

Random Forest classifier

Predicted	user	non-user	All
Actual			
user	621.0	0.0	621.0
non-user	8.0	0.0	8.0
All	629.0	0.0	629.0

Precision: 0.9872813990461049
Recall: 1.0

SVM classifier

Predicted	user	non-user	All
Actual			
user	621.0	0.0	621.0
non-user	8.0	0.0	8.0
All	629.0	0.0	629.0

Precision: 0.9872813990461049
Recall: 1.0

Gradient Boosting classifier

Predicted	user	non-user	All
Actual			
user	618	3	621
non-user	8	0	8
All	626	3	629

Precision: 0.987220447284345
Recall: 0.9951690821256038

MLP classifier

Predicted	user	non-user	All
Actual			
user	617	4	621
non-user	8	0	8
All	625	4	629

Precision: 0.9872
Recall: 0.9935587761674718

KNN classifier

Predicted	user	non-user	All
Actual			
user	621.0	0.0	621.0
non-user	8.0	0.0	8.0
All	629.0	0.0	629.0

Precision: 0.9872813990461049
Recall: 1.0

Mushrooms Dataset

Decision Tree classifier

Predicted Actual	user	non-user	All
user	170	451	621
non-user	2	6	8
All	172	457	629

Precision: 0.9883720930232558
Recall: 0.9883720930232558

Random Forest classifier

Predicted Actual	user	non-user	All
user	221	400	621
non-user	4	4	8
All	225	404	629

Precision: 0.9822222222222222
Recall: 0.355877616747182

SVM classifier

Predicted Actual	user	non-user	All
user	236	385	621
non-user	4	4	8
All	240	389	629

Precision: 0.9833333333333333
Recall: 0.38003220611916266

Gradient Boosting classifier

Predicted Actual	user	non-user	All
user	211	410	621
non-user	5	3	8
All	216	413	629

Precision: 0.9768518518518519
Recall: 0.3397745571658615

MLP classifier

Predicted Actual	user	non-user	All
user	222	399	621
non-user	4	4	8
All	226	403	629

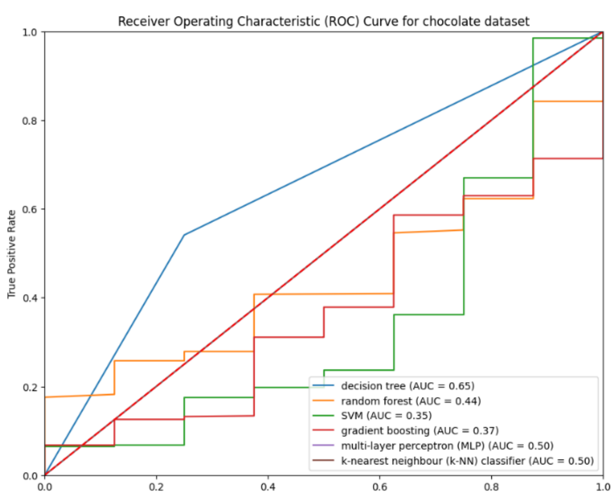
Precision: 0.9823008849557522
Recall: 0.357487922705314

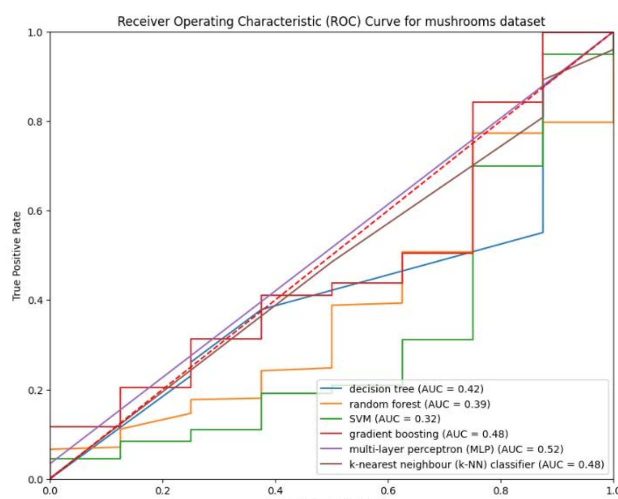
KNN classifier

Predicted Actual	user	non-user	All
user	239	382	621
non-user	4	4	8
All	243	386	629

Precision: 0.9835390946502057
Recall: 0.38486312399355876

ROC curve as a Summary





Classification with the Under-sampling method

Chocolate Dataset

Decision Tree classifier

Predicted	user	non-user	All
Actual			
user	589	32	621
non-user	8	0	8
All	597	32	629

Precision: 0.9865996649916248
Recall: 0.9484702093397746

Random Forest classifier

Predicted	user	non-user	All
Actual			
user	481	140	621
non-user	4	4	8
All	485	144	629

Precision: 0.9917525773195877
Recall: 0.7745571658615137

SVM classifier

Predicted	user	non-user	All
Actual			
user	553	68	621
non-user	8	0	8
All	561	68	629

Precision: 0.9857397504456328
Recall: 0.8904991948470209

Gradient Boosting classifier

Predicted	user	non-user	All
Actual			
user	472	149	621
non-user	5	3	8
All	477	152	629

Precision: 0.989517819706499
Recall: 0.7600644122383253

MLP classifier

Predicted	user	non-user	All
Actual			
user	493	128	621
non-user	7	1	8
All	500	129	629

Precision: 0.986
Recall: 0.7938808373590982

KNN classifier

Predicted	user	non-user	All
Actual			
user	568	53	621
non-user	8	0	8
All	576	53	629

Precision: 0.9861111111111112
Recall: 0.9146537842190016

Mushrooms Dataset

Decision Tree classifier

Predicted	user	non-user	All
Actual			
user	573	48	621
non-user	7	1	8
All	580	49	629

Precision: 0.9879310344827587
Recall: 0.9227053140096618

Random Forest classifier

Predicted	user	non-user	All
Actual			
user	481	140	621
non-user	4	4	8
All	485	144	629

Precision: 0.9917525773195877
Recall: 0.7745571658615137

SVM classifier

Predicted	user	non-user	All
Actual			
user	553	68	621
non-user	8	0	8
All	561	68	629

Precision: 0.9857397504456328
Recall: 0.8904991948470209

Gradient Boosting

Predicted	user	non-user	All
Actual			
user	470	151	621
non-user	5	3	8
All	475	154	629

Precision: 0.9894736842105263
Recall: 0.7568438003220612

MLP classifier

Predicted	user	non-user	All
Actual			
user	494	127	621
non-user	7	1	8
All	501	128	629

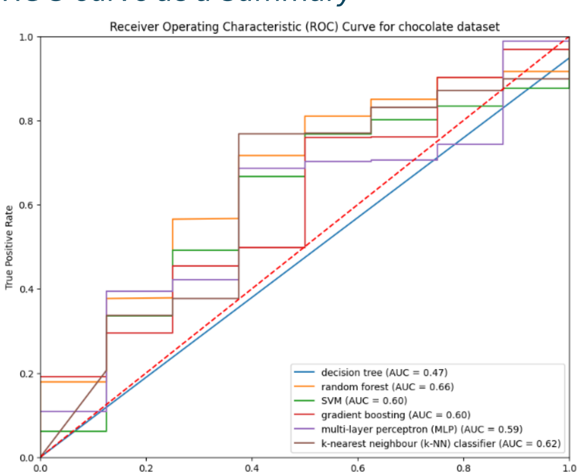
Precision: 0.9860279441117764
Recall: 0.7954911433172303

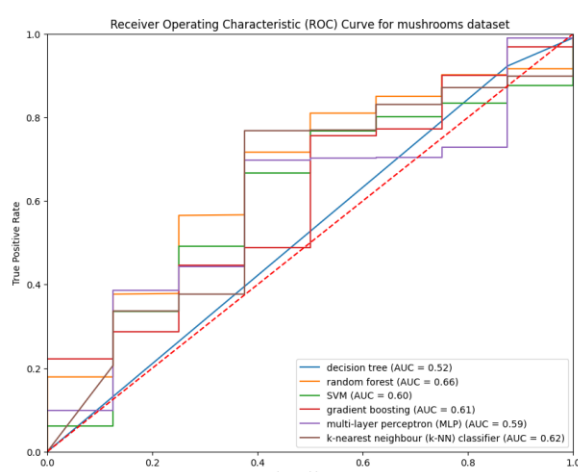
KNN classifier

Predicted	user	non-user	All
Actual			
user	568	53	621
non-user	8	0	8
All	576	53	629

Precision: 0.9861111111111112
Recall: 0.9146537842190016

ROC curve as a Summary





Classification with the Over-sampling method

Chocolate Dataset

Decision Tree classifier

Predicted	user	non-user	All
Actual			
user	456	165	621
non-user	7	1	8
All	463	166	629

Precision: 0.9848812095032398
Recall: 0.7342995169082126

Random Forest classifier

Predicted	user	non-user	All
Actual			
user	427	194	621
non-user	6	2	8
All	433	196	629

Precision: 0.9861431870669746
Recall: 0.6876006441223832

SVM classifier

Predicted	user	non-user	All
Actual			
user	481	140	621
non-user	7	1	8
All	488	141	629

Precision: 0.985655737704918
Recall: 0.7745571658615137

Gradient Boosting

Predicted	user	non-user	All
Actual			
user	415	206	621
non-user	6	2	8
All	421	208	629

Precision: 0.9857482185273159
Recall: 0.6682769726247987

MLP classifier

Predicted	user	non-user	All
Actual			
user	466	155	621
non-user	5	3	8
All	471	158	629

Precision: 0.9893842887473461
Recall: 0.750402576489533

KNN classifier

Predicted	user	non-user	All
Actual			
user	421	200	621
non-user	4	4	8
All	425	204	629

Precision: 0.9905882352941177
Recall: 0.677938808373591

Mushrooms Dataset

Decision Tree classifier

Predicted	user	non-user	All
Actual			
user	431	190	621
non-user	4	4	8
All	435	194	629

Precision: 0.9908045977011494
Recall: 0.6940418679549114

Random Forest classifier

Predicted	user	non-user	All
Actual			
user	428	193	621
non-user	5	3	8
All	433	196	629

Precision: 0.9884526558891455
Recall: 0.6892109500805152

SVM classifier

Predicted	user	non-user	All
Actual			
user	481	140	621
non-user	7	1	8
All	488	141	629

Precision: 0.985655737704918
Recall: 0.7745571658615137

Gradient Boosting classifier

Predicted	user	non-user	All
Actual			
user	411	210	621
non-user	6	2	8
All	417	212	629

Precision: 0.9856115107913669
Recall: 0.6618357487922706

MLP classifier

Predicted	user	non-user	All
Actual			
user	467	154	621
non-user	5	3	8
All	472	157	629

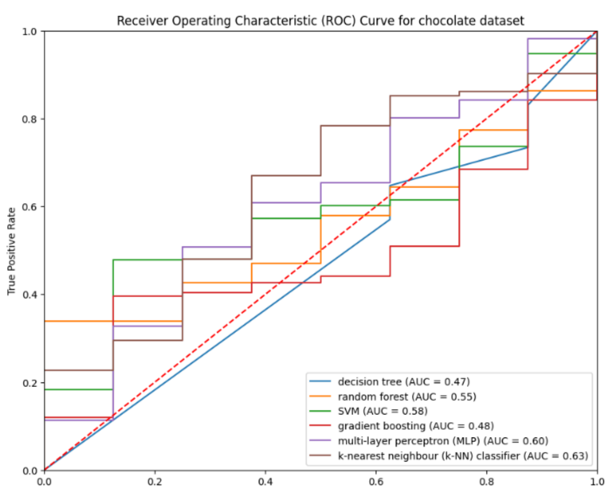
Precision: 0.989406779661017
Recall: 0.7520128824476651

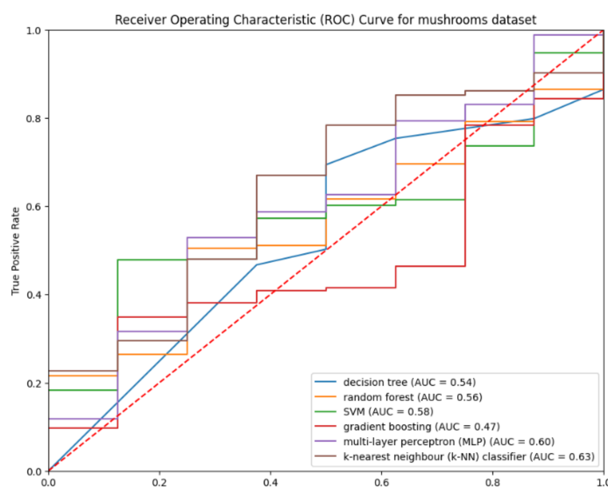
KNN classifier

Predicted	user	non-user	All
Actual			
user	421	200	621
non-user	4	4	8
All	425	204	629

Precision: 0.9905882352941177
Recall: 0.677938808373591

ROC curve as a Summary





Classification with a combination of sampling methods

Chocolate Dataset

Decision Tree classifier

Predicted	user	non-user	All
Actual			
user	319	302	621
non-user	2	6	8
All	321	308	629

Precision: 0.9937694704049844
Recall: 0.5136876006441223

Random Forest classifier

Predicted	user	non-user	All
Actual			
user	435	186	621
non-user	6	2	8
All	441	188	629

Precision: 0.9863945578231292
Recall: 0.7004830917874396

SVM classifier

Predicted	user	non-user	All
Actual			
user	469	152	621
non-user	7	1	8
All	476	153	629

Precision: 0.9852941176470589
Recall: 0.7552334943639292

Gradient Boosting classifier

Predicted	user	non-user	All
Actual			
user	427	194	621
non-user	6	2	8
All	433	196	629

Precision: 0.9861431870669746
Recall: 0.6876006441223832

MLP classifier

Predicted	user	non-user	All
Actual			
user	473	148	621
non-user	5	3	8
All	478	151	629

Precision: 0.9895397489539749
Recall: 0.7616747181964574

KNN classifier

Predicted	user	non-user	All
Actual			
user	448	173	621
non-user	5	3	8
All	453	176	629

Precision: 0.9889624724061811
Recall: 0.7214170692431562

Mushrooms Dataset

Decision Tree classifier

Predicted	user	non-user	All
Actual			
user	388	233	621
non-user	3	5	8
All	391	238	629

Precision: 0.9923273657289002
Recall: 0.6247987117552335

Random Forest classifier

Predicted	user	non-user	All
Actual			
user	435	186	621
non-user	6	2	8
All	441	188	629

Precision: 0.9863945578231292
Recall: 0.7004830917874396

SVM classifier

Predicted	user	non-user	All
Actual			
user	469	152	621
non-user	7	1	8
All	476	153	629

Precision: 0.9852941176470589
Recall: 0.7552334943639292

Gradient Boosting classifier

Predicted	user	non-user	All
Actual			
user	427	194	621
non-user	6	2	8
All	433	196	629

Precision: 0.9861431870669746
Recall: 0.9861431870669746

MLP classifier

Predicted	user	non-user	All
Actual			
user	474	147	621
non-user	5	3	8
All	479	150	629

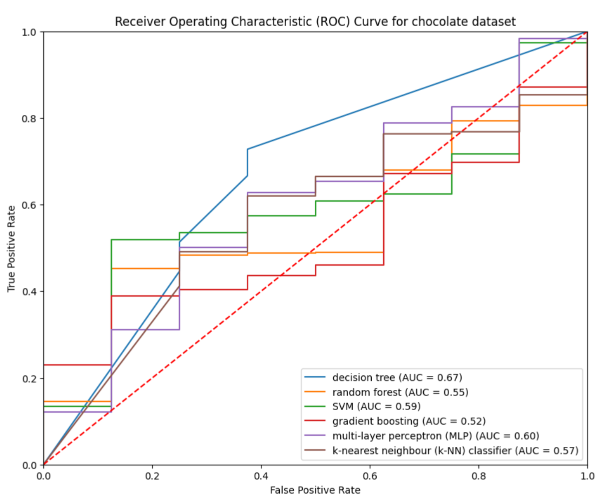
Precision: 0.9895615866388309
Recall: 0.7632850241545893

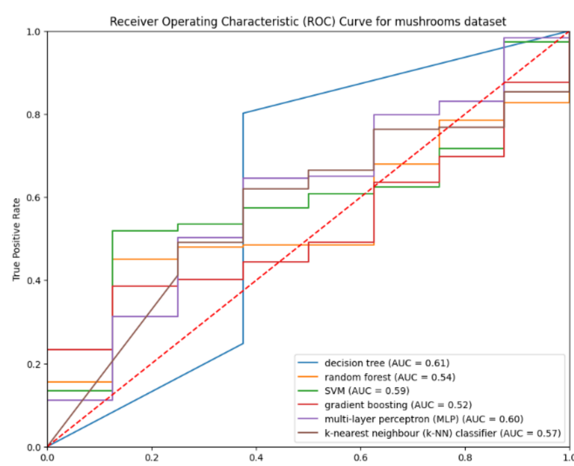
KNN classifier

Predicted	user	non-user	All
Actual			
user	448	173	621
non-user	5	3	8
All	453	176	629

Precision: 0.9889624724061811
Recall: 0.7214170692431562

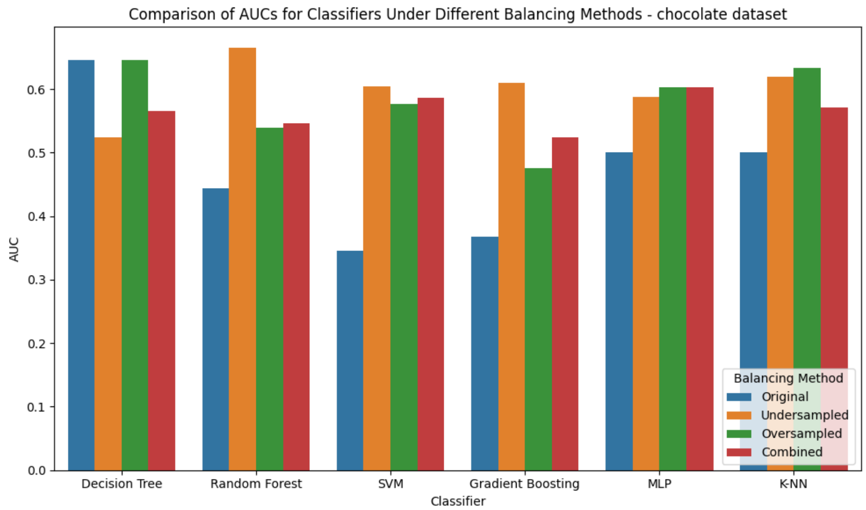
ROC curve as a Summary

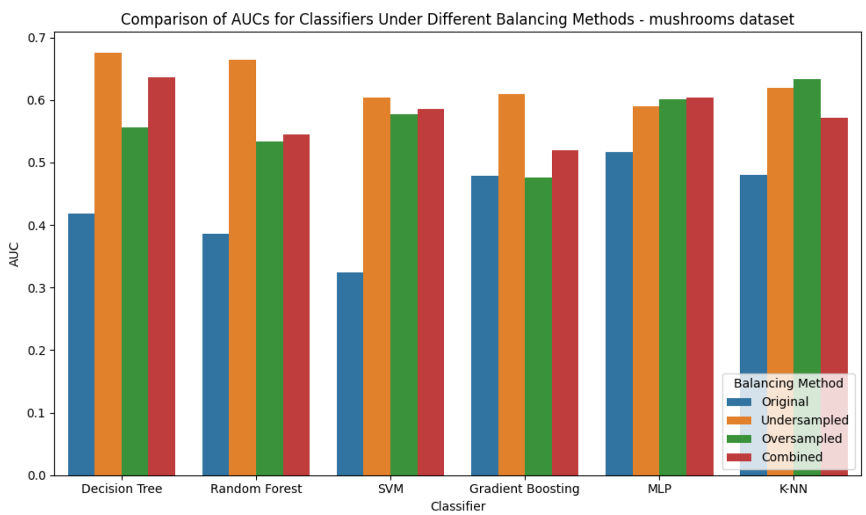




Further Analysis

Based on the above estimations, essentially the ROC curves, we have the AUCs which compares the performance of different classifiers under different sampling methods. Therefore, they are summarized in the following bar plots.





Lesson Learnt

Processes in the Pipeline

The pipeline begins with data preprocessing, followed by feature extraction. Six classifiers are then instantiated with custom parameters. *Random Search* was used for hyperparameter optimization. After splitting the datasets into training and test sets, classifiers were trained and evaluated using precision, recall, and AUC.

Analysis from the Evaluation Metrics

The primary analysis is mainly based on the *precision*, *recall*, and *AUC* scores from confusion matrices. In the original classification (without resampling), the Decision Tree achieved the highest scores across both datasets. With *under-sampling*, the Random Forest showed the highest precision in both datasets, while the Decision Tree had the highest recall and AUC in the chocolate dataset. For *oversampling*, KNN gave the highest precision and AUC in the chocolate dataset, and the SVM had the highest recall. Using *combined sampling*, the Decision Tree excelled in precision and AUC in both datasets, while MLP and Gradient Boosting had the highest recall. Overall, the **Decision Tree** is the best classifier which is also suggested in the bar plot regarding AUC comparisons across different sampling methods.

Issue in the Datasets

Why different sampling methods suggests the best classifier differently is because of the *class imbalance* issue. In the Chocolate dataset, we can clearly see that the number of "user" category sample (i.e., 621) is far more than that of "non-user" category samples (i.e., 8). In the Mushrooms dataset, we can also clearly see that the number of magic mushroom non-users (i.e., 386) is slightly more than that of users (i.e., 243). The imbalance of classes obviously leads to a biased conclusion. In other words, all techniques generally suggest **decision tree** the best classifier.

Addressing the Issue

Regarding the *class imbalance*, *Random Under Sampler* is used to under-sample the majority classes in both datasets; *SMOTE* is used to over-sample the minority classes; and a combination of techniques is used to make a fair sampling approach to the data. The advantage of that is obviously balancing the number of samples in each class. However, this may lead to other potential problems. For instance, after under-sampling the dataset, some data from the majority class are trimmed. This causes data loss. On the other hand, after over-sampling the minority class, some unnecessary data are added. Affirmatively, it reduces biases in the results, but it increases the variance of data.

Conclusion

According to the notion of "No Free Lunch" principle, no single algorithm is always the most accurate one. Therefore, the analysis here also considers using a combination of sampling methods to address the *class imbalance* issue fairly. The AUC comparison graphs suggests that:

The combination of methods has nearly an average AUC between both sampling methods; and,

By using any sampling techniques (including the combination of methods), it generally gives a higher AUC score from any classifier based on any dataset.

In conclusion, by using a sampling technique reduces biases from the class imbalance issue; and by using a combination of sampling techniques, a fair conclusion is drawn.