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

Precision: 0.9883720930232558
Recall: 0.9883720930232558

**Random Forest classifier**

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

Precision: 0.9822222222222222
Recall: 0.355877616747182

**SVM classifier**

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

Precision: 0.9833333333333333
Recall: 0.38003220611916266

**Gradient Boosting classifier**

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

Precision: 0.9768518518518519
Recall: 0.3397745571658615

**MLP classifier**

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

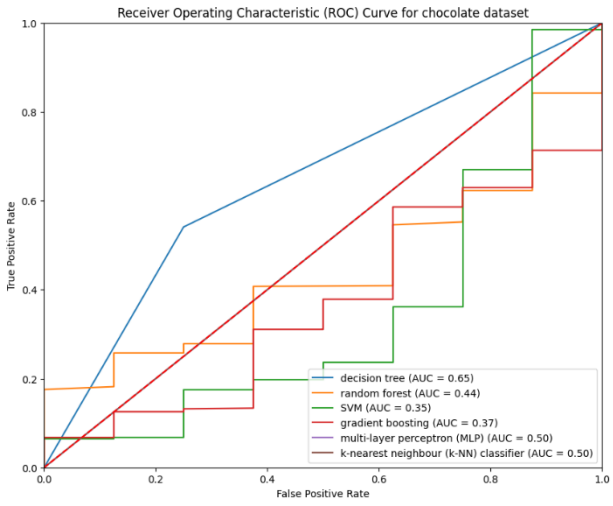
Precision: 0.9823008849557522
Recall: 0.357487922705314

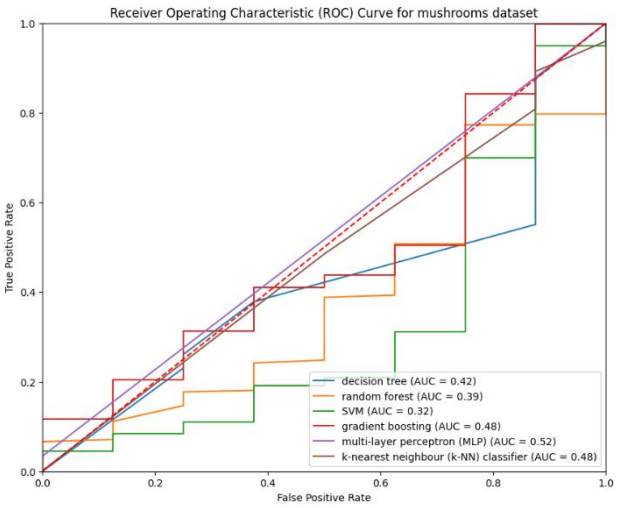
**KNN classifier**

| Predicted | user | non-user | All |
|---|---|---|---|
| Actual | | | |
| 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 |

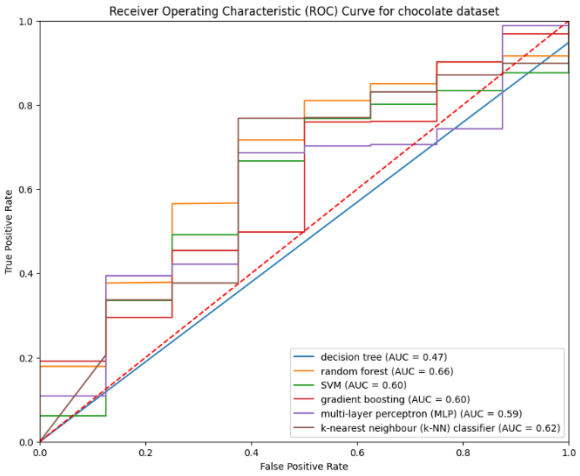
Precision: 0.9860279441117764
Recall: 0.7954911433172303

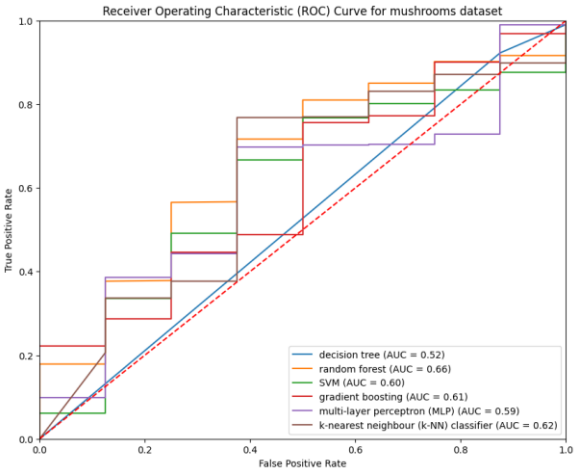
**KNN classifier**

| 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 |

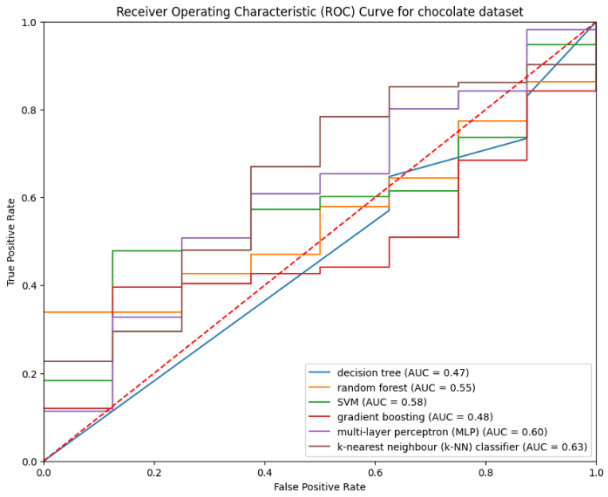
Precision: 0.989406779661017
Recall: 0.7520128824476651

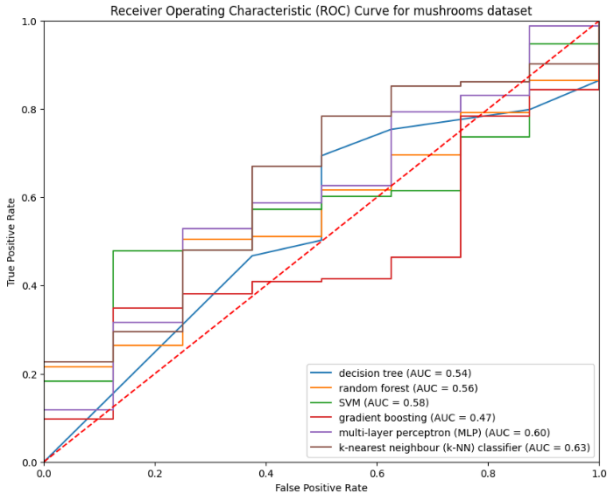
**KNN classifier**

| 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 |

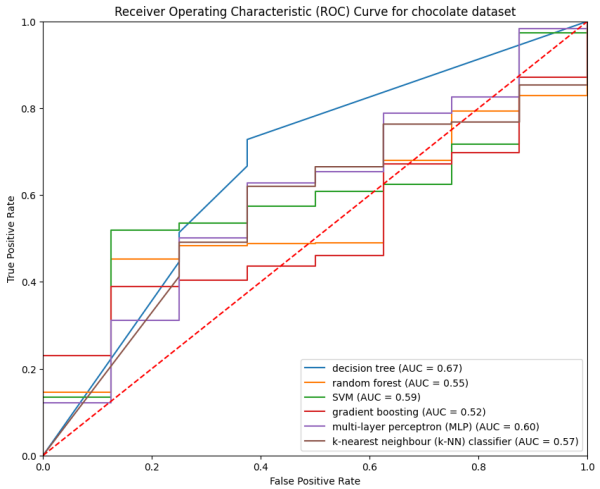
Precision: 0.9895615866388309
Recall: 0.7632850241545893

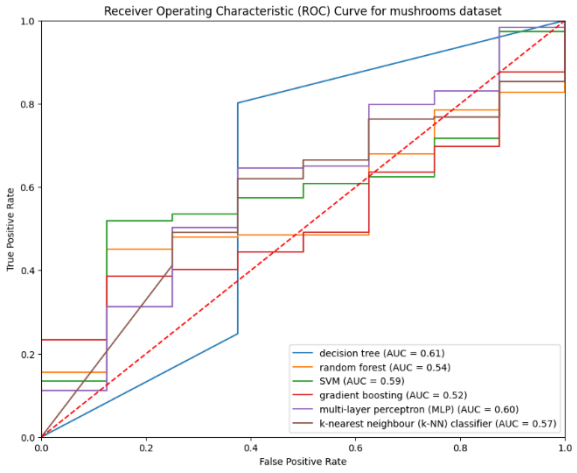
**KNN classifier**

| 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*



## Lesson Learnt

**Submit a 400-word to 500-word summary discussing the results you obtained and the lessons you learned when analyzing this data.**

**- Your answer should focus on the behaviour of the algorithms, the results obtained, and the impact of rebalancing.**

**- Your answer should also highlight the differences between the models constructed against the two datasets and the differences between the rebalancing processes and results for these two datasets.**

Behavior of algorithms:

- Pipeline class which contains 6 models for training at the same time.
- Identify overfitting / underfitting issues on the original classification by the comparison of precisions and recalls.
- State whichever