Report for Machine Learning Project

Member 1: Musab bin Jamil - 29409

Member 2: Hussain Diwan - 29410

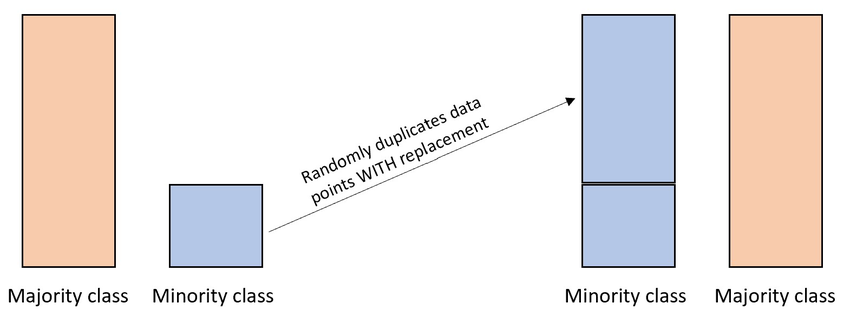
# Q1: Explain each selected CI technique with diagram, if necessary, in your own words.

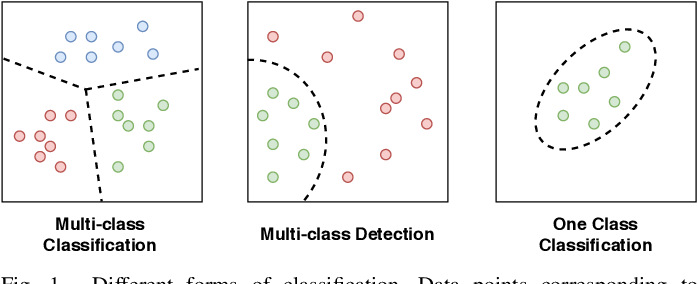
## Ans 1:

For our project we chose random over sampling, class weighting and one class learning.

## Random Over Sampling

For a class imbalance dataset, the core issue is the amount of data available. Random over sampling counters this issue by duplicating samples for the minority class at random. This balances the dataset to a common level which is same for all classes in the dataset.



Figure 1 Random Over Sampler

## One-Class Learning

If we train our data on only one class this can effectively resolve our imbalance issue. Since the data is curating itself based on only one label it will be able to learn most features without any sort of dwarfing effect or bias from other labels. This will result in a model with better accuracy for that specific label as the model is focused on only this part of the dataset. This can be applied on the minority class or any other class based on what our required preference is and which label is valued to us.

Figure 2 One-Class Classification

## Class Weighting

Another technique that can be used to tackle class imbalance is class weighting. This means we provide weights for each class in our dataset and the algorithm takes into account these values and learns on each class based on the provided weight or importance. Hence, if there is a minority class, it would be assigned a heavier weight and therefore, balanced learning. In Figure 3, the weights are applied in the majority voting phase.

# Q2: What is the impact of each CI solution on the classification performance (compare with baseline)?

## Ans 2:

The following table shows the accuracies for selected 5 machine learning models.

Table 1 Results of CI Solutions

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Class Imbalance Project ML1 | | | | | | | |
|  |
| **Algorithm** | **Sampling Technique** | **Dataset 1: Celestial Object Dataset** | | **Dataset 2: Fake News Dataset** | | **Dataset 3: Credit Record Dataset** | |  |
| **Accuracy** | **CV Score** | **Accuracy** | **CV Score** | **Accuracy** | **CV Score** |  |
| KNN | No Sampling | 0.921 | 0.918 | 0.653 | 0.532 | 0.66 | 0.32 |  |
| Random Over Sampler | 0.934 | 0.936 | 0.547 | 0.469 | 0.85 | 0.6 |  |
| Class Weighting | N/A | N/A | N/A | N/A | N/A | N/A |  |
| One-Class Learning | 1 | 1 | Not performed due to overfitting |  | 1 | 1 |  |
| Logistic Regression | No Sampling | 0.83 | 0.817 | 0.653 | 0.659 | 0.42 | 0.42 |  |
| Random Over Sampler | 0.727 | 0.749 | 0.531 | 0.537 | 0.213 | 0.21 |  |
| Class Weighting | 0.681 | 0.667 | 0.653 | 0.66 | 0.104 | 0.082 |  |
| One-Class Learning | N/A | N/A | N/A | N/A | N/A | N/A |  |
| Naive Bayes | No Sampling | 0.5 | 0.493 | 0.414 | 0.408 | 0.474 | 0.466 |  |
| Random Over Sampler | 0.598 | 0.623 | 0.477 | 0.482 | 0.247 | 0.244 |  |
| Class Weighting | 1 | 1 | Not performed due to overfitting |  | 0.4 | 0.4 |  |
| One-Class Learning | 1 | 1 | Not performed due to overfitting |  | 1 | 1 |  |
| Random Forest | No Sampling | 0.991 | 0.99 | 0.646 | 0.651 | 0.625 | 0.63 |  |
| Random Over Sampler | 0.996 | 0.997 | 0.611 | 0.617 | 0.856 | 0.76 |  |
| Class Weighting | 0.991 | 0.99 | Not performed due to ensemble technique |  | 0.604 | 0.611 |  |
| One-Class Learning | 0.998 | 1 | Not performed due to ensemble technique |  | 0.715 | 1 |  |
| XGBoost | No Sampling | 0.992 | 0.992 | 0.652 | 0.659 | 0.68 | 0.681 |  |
| Random Over Sampler | 0.992 | 0.997 | 0.583 | 0.594 | 0.641 | 0.728 |  |
| Class Weighting | 0.992 | 0.992 | Not performed due to ensemble technique |  | 0.68 | 0.681 |  |
| One-Class Learning | N/A | N/A | N/A | N/A | N/A | N/A |  |

Based on the figures presented on table 1, we can see the trend, that the simpler algorithms (KNN, Logistic regression and Naïve-Bayes algorithm) have better classification after the dataset has been balanced. While the ensemble methods (Random Forest and XGBoost) had a marginal change in accuracy as they handle class imbalance with multiple learners. But as is mentioned in the notebook, each technique had an impact on other metrics like precession, recall and more. (For further detail kindly refer to the notebook)

# Q3: Does the performance of a CI solution get impacted significantly by the choice of different algorithms (compare for both baseline and CI-based)?

## Ans 3:

For the simpler algorithms, the CI solution takes quite a hit and is adequately affected based on the model, but for the ensemble techniques there is marginal change in accuracy after the CI solution and that too is a worse accuracy than on the imbalanced dataset. This is understandable as ensemble techniques are capable of handling class imbalance with the help of multiple learners.

# References

1. Figure 1: <https://www.researchgate.net/figure/Random-oversampling-process_fig1_367177472>
2. Figure 2: <https://www.semanticscholar.org/paper/One-Class-Classification%3A-A-Survey-Perera-Oza/bd6262ebdd1a865e8e6859ab7dd8dc576d2a90e6/figure/0>