# Addressing Class Imbalance in Tabular Data Using KMeans-Based Undersampling

Devora Siminovsky and Tamar Michelson
devorasimi@gmail.com, t0584107102@gmail.com

March 2025

**Abstract**

Class imbalance is a common issue in classification tasks involving tabular data. We propose an approach that applies KMeans clustering to the majority class, replacing it with cluster centers to preserve its structure while reducing imbalance. This method aims to improve minority class classification while maintaining feature integrity. We evaluate it on multiple datasets, comparing it with standard balancing techniques using precision, recall, F1, and balanced accuracy. Our results provide insight into its potential benefits and limitations as a structured alternative to traditional resampling methods.

## 1 Introduction

### 1.1 Problem description

Class imbalance, where one class has significantly more samples than another, leads to biased models that favor the majority class. Many standard machine learning algorithms optimize for overall accuracy, resulting in poor performance in the minority class. Our goal is to explore alternative ways to address class imbalance while preserving the original feature distribution.

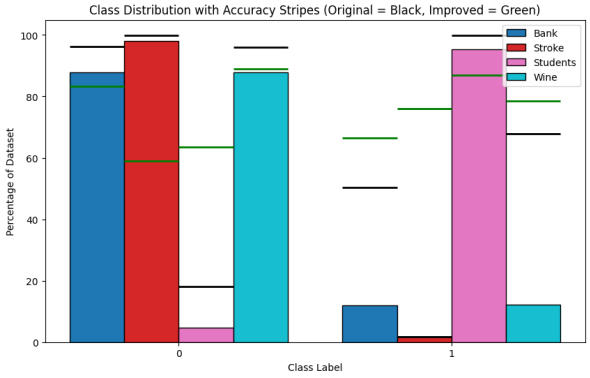


Figure 1: Comparison of classification results before and after applying our technique. The black and green lines represent the percentage of correct samples with (green) and without (black) using our method. When XGBoost is applied directly to the imbalanced data, the majority class is classified with high accuracy, while the minority class suffers from poor classification performance.

### 1.1.1 Element in the DS Pipeline to Improve

We focus on handling skewed class distributions in supervised learning, particularly classification models using tabular data (Fig. 1). We want to improve the element of balancing the data, by improving the Undersampling technique. Imbalanced datasets lead to misleading evaluation metrics, ineffective decision boundaries, and poor generalization. Current techniques, such as random undersampling, often discard critical information. Oversampling methods like SMOTE introduce synthetic data, which may not fully capture real-world distributions.

## 1.2 Connection to Course Material

Our project applies key course concepts, particularly understanding data rather than blindly applying models, decision trees, XGBoost, and SHAP. XGBoost, based on decision trees, is sensitive to class imbalance, which can skew predictions. To address this, we use clustering to balance the data while maintaining feature distributions, unlike traditional undersampling that may distort the data. SHAP values help us interpret model predictions, showing that after balancing, the most important features become clearer, reducing bias and preserving meaningful patterns in the data.

# 2 Solution overview

In our approach, we begin by applying KMeans clustering to the majority class, reducing the number of samples while preserving the structure of the data by selecting the cluster centers as

representative points. All minority class samples are retained, resulting in a balanced training dataset. Then we train an XGBoost model on the newly balanced dataset. Finally, we evaluate and compare the model's performance before and after balancing, using various metrics to assess improvements.

### 2.1 KMeans algorithm

KMeans is an unsupervised machine learning algorithm that partitions data into a predefined number of clusters. It assigns each data point to the cluster with the nearest centroid, iteratively updating the centroids to minimize the variance within each cluster. This process helps to uncover inherent patterns and groupings in the data. When applying our method, we observed that the data was not separated well. To address this issue, we enhanced the KMeans algorithm by incorporating a mechanism that assigns the most frequent value for binary or categorical features, rather than averaging them. This modification helps maintain the categorical integrity of the features while still benefiting from the clustering process.

### 2.2 Oversampling VS Undersampling

Oversampling increases the minority class samples, while undersampling reduces the majority class samples to balance the dataset. Oversampling can cause overfitting, while undersampling may lead to loss of valuable data. Both techniques aim to improve model performance by addressing class imbalance. Our method is an undersampling approach, where we apply KMeans clustering to reduce the majority class while preserving the integrity of the data and avoiding the loss of important features.

## 3 Experimental evaluation

### 3.1 Data

We used the following datasets for our analysis, with the class sizes and imbalance ratios summarized in Table 1:

1. Bank: A dataset containing marketing campaign data from a Portuguese bank, focusing on customer response to subscription offers.

2. Stroke: A dataset aimed at predicting the likelihood of a stroke based on various health indicators and lifestyle factors.

3. Students: A dataset on student performance, which includes information on demographics and academic scores in different subjects.

4. Wine: A dataset consisting of physicochemical attributes and quality ratings for red and white wines.

| Dataset | Class Imbalance Ratio | Majority Class | Minority Class |
|---|---|---|---|
| Bank | 0.1161 | 12,594 | 4,198 |
| Stroke | 0.0189 | 2,200 | 440 |
| Students | 0.0444 | 255 | 85 |
| Wine | 0.1433 | 393 | 131 |

Table 1: Size of both classes and imbalance ratio in each dataset.

### 3.2 Results

Balancing the dataset with k-means shifts class distributions, lowering accuracy but improving balanced accuracy and F1-score. Accuracy drops because the model misclassifies more majority-class instances, but this trade-off enhances minority-class detection. This trade-off is shown in (Fig. 1). In imbalanced datasets, accuracy alone is unreliable, as it can remain high even if the model favors the majority class. By balancing the dataset, the model prioritizes correctly classifying the minority class, which improves balanced accuracy. Precision tends to decrease after balancing, as the model increases its predictions for the minority class, resulting in more false positives. Conversely, recall improves because the model becomes better at identifying true positives for the minority class. This trade-off between precision and recall is expected, as the model sacrifices some precision in favor of better recall for the minority class. Therefore, metrics like balanced accuracy, precision, recall, and F1-score provide a more meaningful assessment of model performance in imbalanced datasets, as they reflect the model's ability to handle both classes effectively 2, 3.

| Method | Dataset 1 (Bank) | | | | | Dataset 2 (Stroke) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Bal Acc | Prec | Recall | F1 | Acc | Bal Acc | Prec | Recall | F1 |
| Baseline XGBoost | 0.906 | 0.733 | 0.643 | 0.504 | 0.565 | 0.981 | 0.509 | 0.333 | 0.019 | 0.035 |
| Our Balanced (KMeans) | 0.812 | 0.748 | 0.351 | 0.665 | 0.460 | 0.594 | 0.675 | 0.034 | 0.759 | 0.065 |
| Random Undersampling | 0.847 | 0.864 | 0.434 | 0.886 | 0.583 | 0.694 | 0.748 | 0.047 | 0.806 | 0.089 |
| SMOTE | 0.894 | 0.754 | 0.560 | 0.570 | 0.565 | 0.961 | 0.508 | 0.031 | 0.037 | 0.034 |

Table 2: Comparison of different data balancing techniques on model performance on Bank and Stroke Datasets.

| Method | Dataset 3 (Students) | | | | | Dataset 4 (Wine) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | Bal Acc | Prec | Recall | F1 | Acc | Bal Acc | Prec | Recall | F1 |
| Baseline XGBoost | 0.960 | 0.590 | 0.962 | 0.998 | 0.980 | 0.926 | 0.819 | 0.704 | 0.679 | 0.691 |
| Our Balanced (KMeans) | 0.858 | 0.753 | 0.980 | 0.869 | 0.921 | 0.878 | 0.838 | 0.500 | 0.786 | 0.611 |
| Random Undersampling | 0.768 | 0.727 | 0.981 | 0.772 | 0.864 | 0.742 | 0.838 | 0.318 | 0.964 | 0.478 |
| SMOTE | 0.916 | 0.632 | 0.966 | 0.945 | 0.956 | 0.878 | 0.823 | 0.500 | 0.750 | 0.600 |

Table 3: Comparison of different data balancing techniques on model performance on Students and Wine Datasets.

## 3.3 Explainability

### 3.3.1 PCA

To better understand the structure of the dataset and how well the K-Means clustering captures the underlying patterns, we performed a three-dimensional Principal Component Analysis (PCA). PCA reduces the dimensionality of the data while retaining its variance, allowing us to visualize the clustering results and assess how effectively the cluster centers represent the data's distribution. The following figure illustrates this analysis for the Students' dataset (Fig. 2). Visualizations for the other datasets can be found in the Git repository.

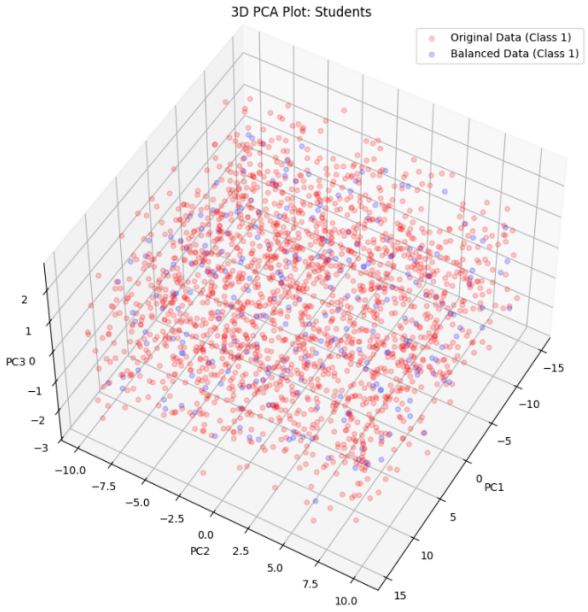


Figure 2: The scatterplot visualizes the data from the large group after performing a three-dimensional Principal Component Analysis (PCA). In the plot, the red dots represent all the original data points in the large group, while the blue dots indicate the centers obtained from the K-Means clustering. From the image, it is evident that these cluster centers effectively capture the overall structure of the data, providing a reliable representation of its distribution.

### 3.3.2 SHAP (SHapley Additive exPlanations)

We tested SHAP to understand the feature importance and how each feature influences model predictions. The analysis reveals notable differences between the balanced and imbalanced datasets, highlighting how balancing can shift feature importance and model behavior. This is shown for the Wine dataset (Fig. 3 right side).

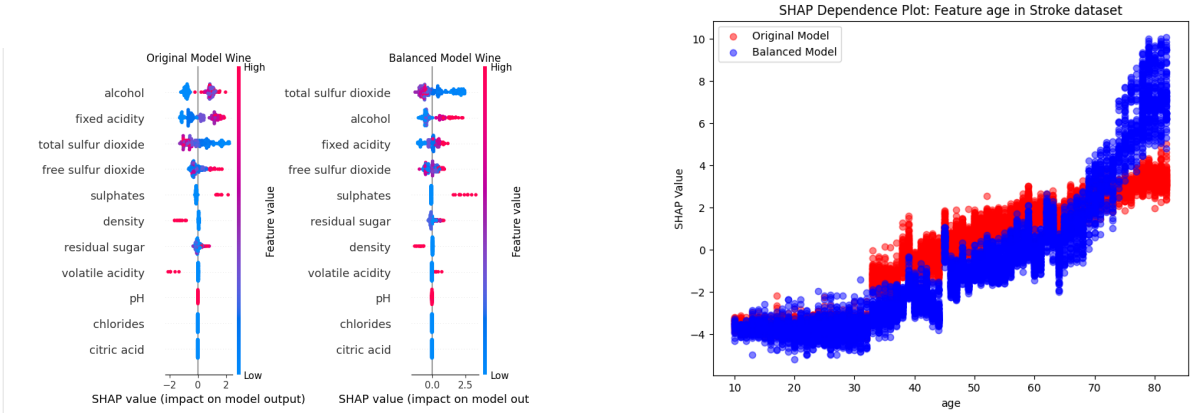


Figure 3: On the left, SHAP analysis for all features in the wine quality prediction dataset. On the right, SHAP analysis of age in the stroke prediction dataset.

The SHAP analysis in Fig. 3 left side, illustrates the impact of age on stroke prediction for models trained on both the original and balanced datasets. In the model built with the original data, age is given less importance, particularly for younger and older individuals. In contrast, the model trained on the balanced data assigns higher importance to these age extremes. This adjustment aligns more logically with medical understanding, as age is a critical factor in stroke risk. By emphasizing these variations, the balanced model likely improves predictive performance and better captures the relationship between age and stroke occurrence. Visualizations for the other datasets can be found in the Git repository.

### 3.3.3 Statistical tests

We performed 5-fold cross-validation and applied balanced accuracy to evaluate the model's performance. To statistically assess the results, we conducted several tests. The Wilcoxon signed-rank test was used to compare paired samples, and in each case, the p-value was below 0.05 (p-value = 0.0312), suggesting a significant difference between the balanced and imbalanced datasets. This implies that there was a consistent difference in performance when switching between the two, supporting the observed changes in accuracy. The Kolmogorov-Smirnov test was used to assess if the data followed a normal distribution. The p-values for this test were above 0.05 (ranging from 0.0595 to 0.1040), which means we fail to reject the null hypothesis, suggesting that the data may indeed follow a normal distribution in most cases. Finally, the paired t-test was applied to compare the means of the paired samples. The results showed highly significant p-values (ranging from 0.0003 to 0.0139), indicating that the differences between the balanced and imbalanced datasets were statistically significant. In conclusion, these statistical tests suggest that balancing the dataset has a significant impact on the model's performance, especially in terms of balanced accuracy, as reflected by the significant p-values from both the Wilcoxon signed-rank test and paired t-test.

## 4 Related Work

Class imbalance is a well-known challenge in machine learning, particularly in classification tasks involving tabular data. Several techniques have been proposed to address this issue, including Synthetic Minority Over-sampling Technique (SMOTE) [1], random undersampling, and cost-sensitive learning (takes into account the different misclassification costs of various classes) [2], [3]. SMOTE generates synthetic samples for the minority class by interpolating between existing samples, helping balance the dataset. However, it can introduce noise and may not always preserve the underlying feature distribution. Random undersampling, on the

other hand, reduces the size of the majority class by randomly removing samples, which may lead to loss of important information.

Other works have explored clustering-based approaches for data balancing. Yong et al. [4] proposed a hybrid oversampling method that combines K-means clustering and a genetic algorithm to enhance the representation of the minority class in imbalanced datasets. Their approach first applies K-means clustering to group the minority class samples, followed by the generation of new synthetic samples within each cluster using a genetic algorithm. Inspired by this, our approach applies KMeans clustering to replace the majority class with cluster centers, ensuring that key patterns are retained while balancing the dataset. In contrast, our method does not generate synthetic data for oversampling but instead applies undersampling by restructuring the majority class to better represent its distribution.

## 5 Conclusion

In the project, we explored the impact of balancing imbalanced datasets through K-means clustering on model performance. The results showed that balancing improved balanced accuracy and F1-score, making the model more effective in detecting minority classes, though it reduced accuracy. Statistical tests confirmed that these changes were statistically significant. This work highlights the importance of using alternative metrics, like balanced accuracy, over simple accuracy when dealing with imbalanced data. Through this project, we learned a lot about the intricacies of dataset balancing and its effects on model performance across different metrics.

## References


[1] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[2] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[3] Vimalraj S Spelmen and R Porkodi. A review on handling imbalanced data. In *2018 international conference on current trends towards converging technologies (ICCTCT)*, pages 1–11. IEEE, 2018.

[4] Yang Yong. The research of imbalanced data set of sample sampling method based on k-means cluster and genetic algorithm. *Energy Procedia*, 17:164–170, 2012.