Employing Random Sampling Techniques on Machine Learning Models: Performance Comparison

Team members:

Jinghao Chen

Roxanne Alvarez

Ananta Arora

Teshani Jayasinghe

Data Analysis and Statistical Inference Course (DANA 4800)

Teamwork assessment

Instructor: Dr. Monica Nguyen

April 2024

Table of Contents

Background

Data imbalance is one of the most common issues in machine learning (ML) classification tasks. It refers to the unequal representation of classes in the training dataset. In other words, there is a minority class that has significantly fewer instances than the majority class; ideally, all classes should be equally distributed. Because of this imbalance, the minority class becomes more difficult to predict as there is less information for the machine learning model to learn from during training.[1] One good example of an imbalanced dataset is the proportion of emails that are spam and not spam. If the model is trained on this dataset, it may exhibit bias towards predicting incoming emails as not spam. Data imbalance could potentially lead to problems such as bias towards the majority class [2], poor generalization to unseen data, and misleading evaluation of the machine learning model's accuracy. In this context, two random sampling techniques, oversampling and undersampling, are explored to address the issue.

[1] edX, "What Is Undersampling?," Master's in Data Science, April 2022, https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/.
[2] Priyanka Dave, "From Bias to Balance: Solving Imbalanced Data Issues," Medium, September 20, 2023, https://priyanka-ddit.medium.com/how-to-deal-with-imbalanced-dataset-86de86c49#:~:text=Bias%20Toward%20Majority%20Class%3A%20The.

Introduction

This study presents three ML models (Support Vector Machine – SVM, Random Forest – RF, eXtreme Gradient Boosting – XGBoost) tasked with predicting the survival outcome of mechanically ventilated patients admitted to an Intensive Care Unit (ICU). In order to accurately predict patient outcomes, a preprocessing step is necessary to handle missing and invalid values. Additionally, two random sampling techniques, such as undersampling and oversampling, are employed to address the heavily imbalanced data. Subsequently, the performances of the three ML models utilizing both undersampling and oversampling methods are evaluated and compared to determine which technique is resulted in more accurate predictions of patient survival outcomes.

Dataset

The Medical Information Mart for Intensive Care (MIMIC-III) dataset is a comprehensive health-related dataset that primarily focuses on patients admitted to the Intensive Care Unit at the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, USA. It consists of 18,883 observations and 70 variables. After removing missing values and invalid ranges, a subset of 12,489 patients and 68 variables (including the response variable) is used to train and test the ML models. During the data exploration process, it is observed that the dataset is heavily skewed, with a significant disparity between the number of survived and deceased patients. Specifically, there are 10,331 survivors, while a substantial 2,158 patients did not survive (Fig. 1).

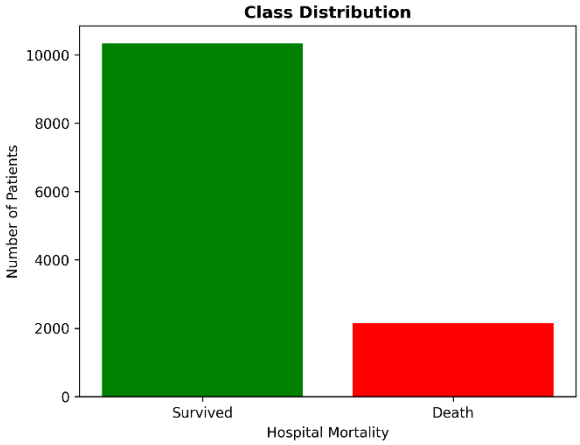

Figure 1. Class distribution of the dataset (MIMIC-III)

Data Preprocessing

Dealing with missing values is a crucial task in the exploratory data analysis process. Removing the missing values without proper evaluation can lead to issues that could significantly impact the results. These issues include loss of information, analysis bias, and statistical power reduction.

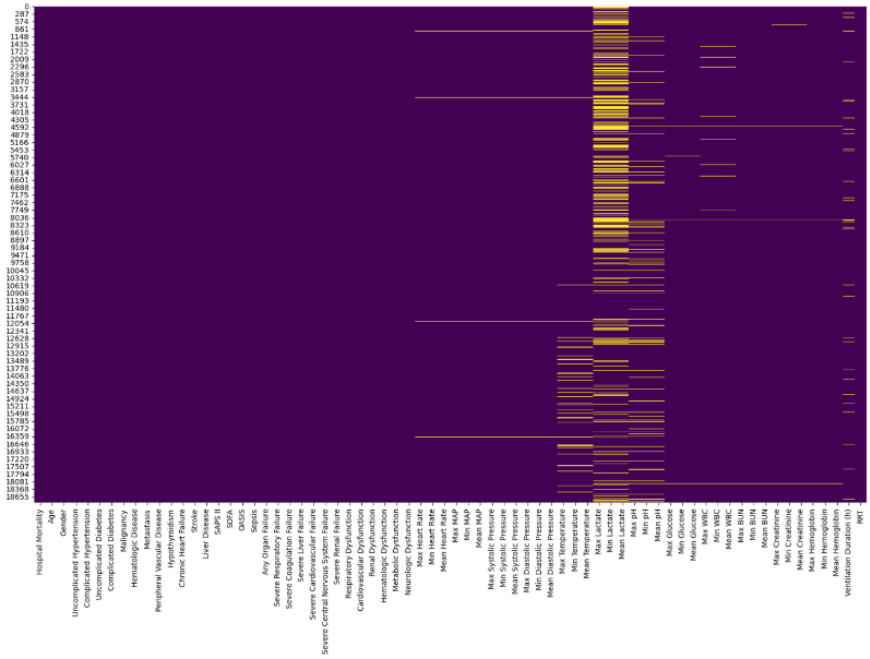

Figure 2. Heatmap for the visualization of the missing values

The heatmap in Fig. 2 illustrates the proportion of missing values, represented by yellow bars, in the dataset. The vital signs data of forty-one patients and the laboratory results of fifty patients are completely missed. This results in 6,084 rows with at least one missing value across all features, which have to be removed. Additionally, the information on renal replacement therapy and ventilation duration is not available on the first day of patient admission, therefore these two columns are excluded in the analysis. Furthermore, M A Papadakis et al. (1993) provided valuable information about the physiologically valid ranges for vital signs and laboratory results[3]. Using this as a reference, 310 observations are removed where at least one of their values falls outside the valid range. The remaining dataset contained 12,799 entries and 68 variables after removing all missing and invalid values.

Random Sampling Methods

Two random sampling methods are employed to balance the training data and optimize the ML models' capability in learning the patterns of both classes, survivors and non-survivors.

Undersampling

After splitting the dataset into training and test subsets, the undersampling process is performed. The training subset, which includes 3,016 observations is use to train the model, while the test subset (1,300 observations) is used to verify the model's predictions. The undersampling technique, as illustrated in Fig. 3, randomly removes entries from the majority class (survivors) until a balanced dataset is achieved, ensuring a more accurate data representation.

[3] M A Papadakis et al., "Prognosis of Mechanically Ventilated Patients," *The Western Journal of Medicine* 159, no. 6 (1993): 659–64, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1022451/.

Two most prevalent disadvantages of undersampling are loss of potentially crucial information and inaccurate representation of the real-world scenario.

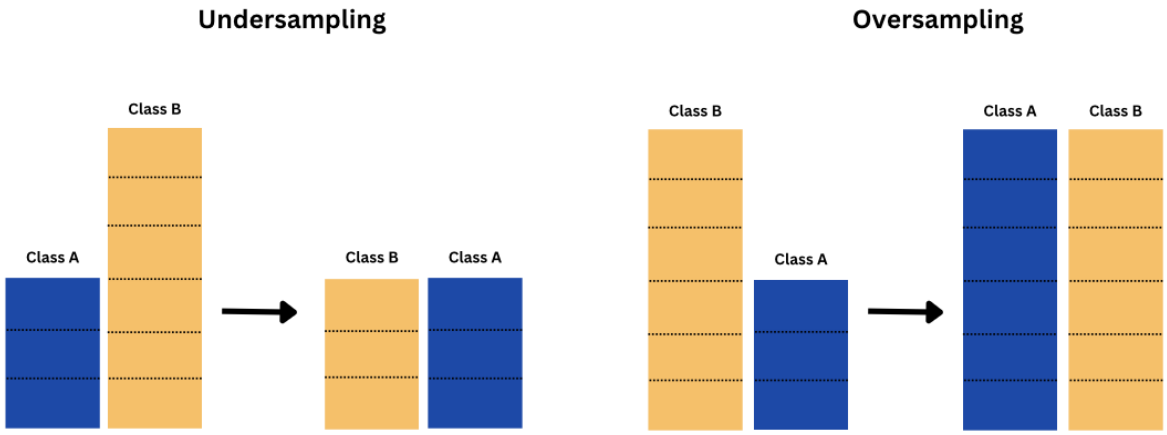

Figure 3. Illustration of the undersampling and oversampling process.

Oversampling

Oversampling is another random sampling technique used to address class imbalance issues where the training data (14,468 observations) is rebalanced by increasing the number of instances in the minority class (non-survivors) through replication of existing instances until a balanced data between survivors and non-survivors is achieved. The test set is now increased to 6,914. This is illustrated in Fig. 3.

Similar to undersampling, oversampling is a straightforward process that doesn't require complex algorithms. However, it is prone to overfitting since it only replicates the existing samples, therefore limiting the model from capturing new observations that may provide additional information about the minority class.

Performance Evaluation

This section discusses various evaluation metrics used to assess the effectiveness of the two random sampling techniques.

Confusion Matrix

The Confusion matrix is a 2X2 table that visualizes the performance of a ML model used for classification problems. It contains metrics including true positive, true negative, false negative, and false positive. True positives (TP) are described as instances correctly classified by the model as positive. Likewise, true negatives (TN) are correctly classified as negative. False negatives (FN) are the instances that are incorrectly classified as negative, while false positives (FP) are incorrectly classified as positive.

The following is an example of a confusion matrix.

Table 1. Sample confusion matrix.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Undersampling

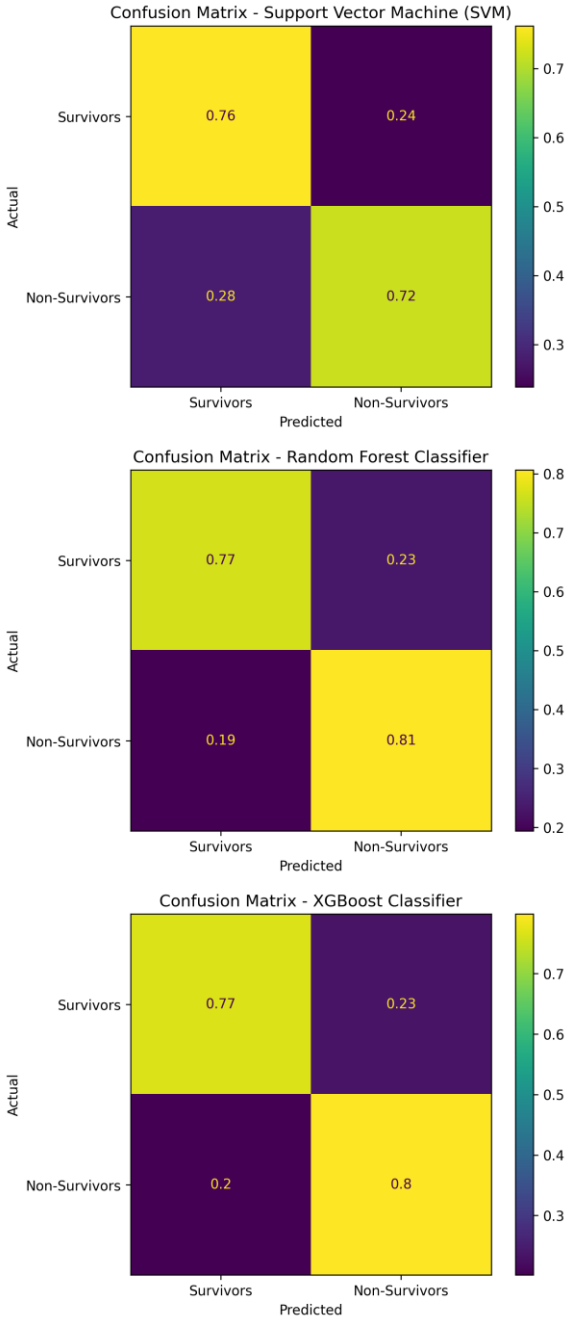

Figure 4. Confusion Matrix of 3 models fit with undersampling dataset.

Oversampling

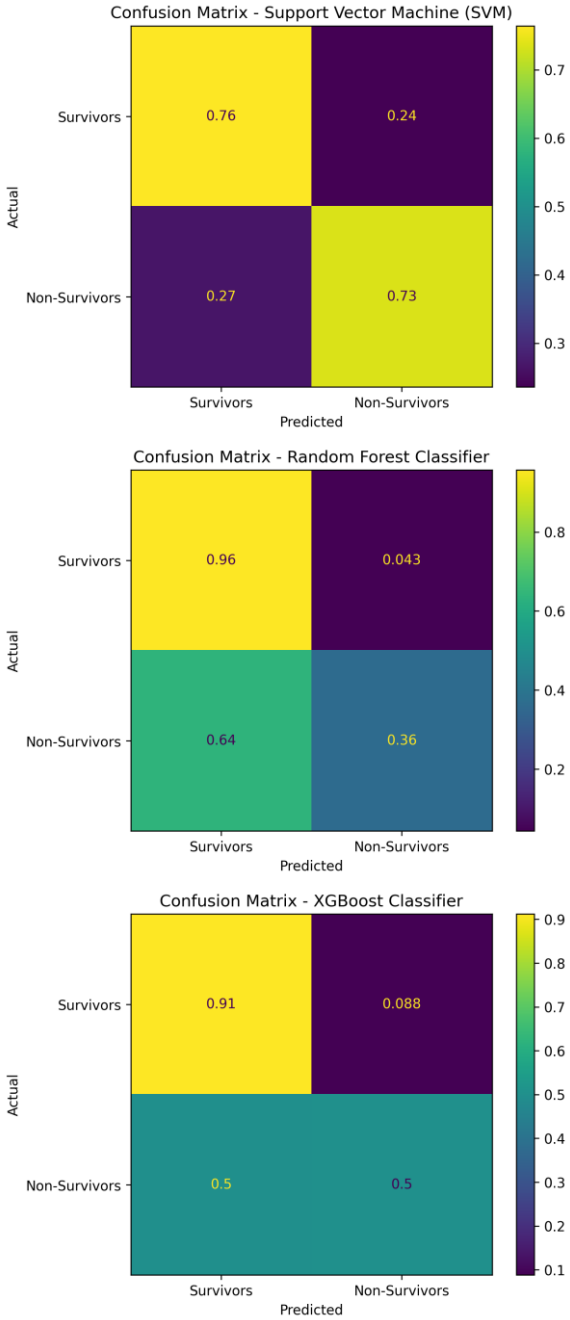

Figure 5. Confusion Matrix of 3 models fit with oversampling dataset.

In a confusion matrix, the columns represent the distribution of the predicted classes while rows represent the distribution of the actual classes. This provides a graphical representation of the model's accuracy in predicting classes by measuring true positives, true negatives, false positives, and false negatives.

The ML model performance comparison between undersampling and oversampling using the confusion matrix as metrics is illustrated in Fig. 4 and Fig. 5. Based on the two groups of confusion matrices, certain models are better suited to either the undersampled or the oversampled dataset. In particular, the SVM models performed equally well with both techniques, as evidenced by the identical confusion matrices produced. This suggests that the choice of techniques does not significantly impact the performance of some models.

On the other hand, the RF and XGBoost models show a clear preference for the undersampling technique. The TP and TN values in both groups support this finding. For instance, in the RF model, the TP and TN values are almost equal for the undersampled dataset, while for the oversampled dataset, the TP value is 0.96, and the TN value is 0.36. This suggests that the model may be biased towards the TP and may not be accurately predicting the patient's death.

Classification Report

One other approach used to evaluate the model is the use of sklearn.metrics module, which provides the classification report method. This generates a tabular summary displaying the primary classification metrics for each class, including precision, recall, F1-score, and accuracy. These four metrics are all derived from the confusion matrix.

Precision is the proportion of correct positive predictions out of all positive predictions made by the model. The precision value ranges from 0 to 1, with 1 meaning that the model produces zero false positives (FP). The formula as

$$Precision = \frac{TP}{TP + FP}$$

Recall measures the proportion of actual positives that are correctly identified by the model, with a value ranging from 0 to 1. A score of 1 indicates that the model produces no false negatives (FN). The formula as

$$Recall = \frac{TP}{TP + FN}$$

F1 score is a metric that considers both precision and recall. It is calculated as the mean of both metrics and assigns equal importance. The score ranges from 0 to 1, where 1 indicates perfect precision and recall, meaning the model produces zero errors. On the other hand, a score of 0 indicates that either precision or recall is 0, which implies that the model incorrectly predicts everything. The formula as

$$F1\ Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$$

The F1 score comes in handy when the user has difficulty choosing between high precision and low recall or vice versa.[4]

Support is the number of samples that are used in that class.

Accuracy is the proportion of correct predictions to total predictions. However, it cannot be treated as only a metric to measure the performance of a model.

[4] Vaibhav Jayaswal, "Performance Metrics: Confusion Matrix, Precision, Recall, and F1 Score," Medium, September 15, 2020, https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Macro average is the average of all classes without considering the proportion of each class in the dataset. This means it treats each class equally.

Weighted average is the average of each class, considering each class's impact on the metric, which is proportional to its size.

Table 2. Classification report of 3 models fit with undersampling and oversampling dataset.

Classification Report: SVM - Undersampling				
	Precision	Recall	F1 Score	Support
Survival	0.73	0.76	0.76	650
Death	0.75	0.72	0.74	650
Accuracy			0.74	1300
Macro avg.	0.74	0.74	0.74	1300
Weighted avg.	0.74	0.74	0.74	1300

Classification Report: SVM - Oversampling				
	Precision	Recall	F1 Score	Support
Survival	0.74	0.76	0.75	3097
Death	0.76	0.73	0.75	3097
Accuracy			0.75	6194
Macro avg.	0.75	0.75	0.75	6194
Weighted avg.	0.75	0.75	0.75	6194

Classification Report: RF Classifier - Undersampling				
	Precision	Recall	F1 Score	Support
Survival	0.8	0.77	0.78	650
Death	0.78	0.81	0.79	650
Accuracy			0.79	1300
Macro avg.	0.79	0.79	0.79	1300
Weighted avg.	0.79	0.79	0.79	1300

Classification Report: RF Classifier - Oversampling				
	Precision	Recall	F1 Score	Support
Survival	0.60	0.96	0.74	3097
Death	0.89	0.36	0.52	3097
Accuracy			0.66	6194
Macro avg.	0.75	0.66	0.63	6194
Weighted avg.	0.75	0.66	0.63	6194

Classification Report: XGBoost Classifier - Undersampling				
	Precision	Recall	F1 Score	Support
Survival	0.79	0.77	0.78	650
Death	0.77	0.80	0.79	650
Accuracy			0.78	1300
Macro avg.	0.78	0.78	0.78	1300
Weighted avg.	0.78	0.78	0.78	1300

Classification Report: XGBoost Classifier - Oversampling				
	Precision	Recall	F1 Score	Support
Survival	0.65	0.91	0.76	3097
Death	0.85	0.50	0.63	3097
Accuracy			0.71	6194
Macro avg.	0.75	0.71	0.69	6194
Weighted avg.	0.75	0.71	0.69	6194

It's important to note that all the metrics in the classification report are calculated from the confusion matrix. This revealed that the SVM model had a similar accuracy rate, just like they have a similar confusion matrix, whereas the other two models showed different results (Table 2). For instance, when the RF classifier is trained with an undersampling dataset, its accuracy rate is 0.79. However, when trained with an oversampling dataset, its accuracy rate dropped to 0.66. This

means that the undersampling model outperformed the oversampling model, as suggested by the confusion matrix.

When the RF classifier is trained with the oversampling dataset, it indicated that the model is good at identifying survival cases (high recall). However, it struggled with accurately identifying death cases (lower recall). The model seems to be more cautious in predicting death, resulting in a high precision but low recall for the death class.

ROC-AUC

The ROC curve is a graphical representation of a binary classifier's performance as the discrimination threshold changes. The curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different threshold settings to show how well the classifier distinguishes between positive and negative samples.

The AUC (Area Under the ROC Curve) measures the area under the curve as a whole. It evaluates the classifier's performance across all possible classification thresholds, providing a comprehensive performance measure. The AUC ranges from 0 to 1, with higher values indicating better performance.

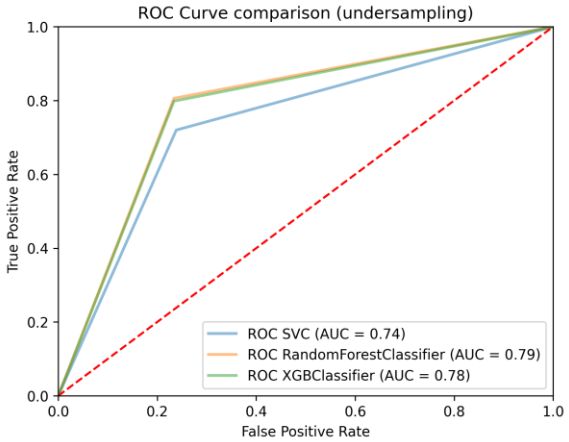

Figure 6. ROC curve undersampling

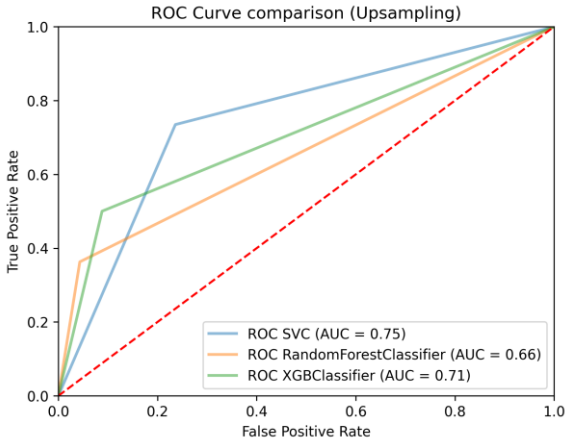

Figure 7. ROC curve oversampling

The true positive and false positive rates in the x-axis and y-axis are calculated from the confusion matrix, so similar results are presented. The SVM model still achieves similar AUC values in both techniques. Also, the RF classifier and XGBoost with the undersampling dataset outperformed the same model fit with the oversampling dataset. This can be illustrated in Fig. 6 and 7.

Conclusion

This study aims to assess the impact of undersampling and oversampling techniques on the ML model's performance in predicting patient survival outcomes. The results indicates that both methods are effective in addressing imbalanced data. However, undersampling is more efficient in achieving a balanced dataset and improving model precision and recall performance for certain models, such as the RF and XGBoost classifiers.

In general, it is advisable to use undersampling as it reduces the size of the data, which results in shorter training time and less computer power consumption. When it comes to selecting an evaluation method, a confusion matrix provides a more detailed breakdown of where your classifier is making mistakes, whereas a classification report gives you important metrics to quickly assess your classifier's performance. Moreover, the ROC-AUC curve allows you to visualize the balance between the true positive rate and the false positive rate of your classifier. It provides a more direct visualization way to evaluate the performance of the model.

Data availability

To conduct this study, the names of the repository can be found below:
https://mimic.physionet.org. The certification ID obtained for this study is 13273317.

Bibliography

Ashraf, Abdallah. “Oversampling — Handling Imbalanced Data.” Medium, December 23, 2023. https://medium.com/@abdallahashraf90x/oversampling-for-better-machine-learning-with-imbalanced-data-68f9b5ac2696#:~:text=Oversampling%20is%20a%20data%20augmentation.

Brownlee, Jason. “8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset.” Machine Learning Mastery, June 7, 2016. https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/.

Dave, Priyanka. “From Bias to Balance: Solving Imbalanced Data Issues.” Medium, September 20, 2023. https://priyanka-ddit.medium.com/how-to-deal-with-imbalanced-dataset-86de86c49#:~:text=Bias%20Toward%20Majority%20Class%3A%20The.

edX. “What Is Undersampling?” Master’s in Data Science, April 2022. https://www.mastersindatascience.org/learning/statistics-data-science/undersampling/.

Jayaswal, Vaibhav. “Performance Metrics: Confusion Matrix, Precision, Recall, and F1 Score.” Medium, September 15, 2020. https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262.

Papadakis, M A, K K Lee, W S Browner, D L Kent, D B Matchar, M K Kagawa, J Hallenbeck, D Lee, R Onishi, and G Charles. “Prognosis of Mechanically Ventilated Patients.” *The Western Journal of Medicine* 159, no. 6 (1993): 659–64. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1022451/.