Combating Class Imbalance for Infant Mortality Risk Modeling: Resampling Strategies in Brazil's Unified Health System

Ricardo Morsoleto[1], Vinícius A. Silva[1], Juliano de S. Caliari[1], Simone Mara F. Miranda[1],
Hiran Nonato M. Ferreira[1]

Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais
(IFSULDEMINAS) - Campus Passos, Brasil
`ricardo.morsoleto@alunos.ifsuldeminas.edu.br`
`{vinicius.silva, juliano.caliari}@ifsuldeminas.edu.br`
`sisimaramiranda@gmail.com`
`hiran.ferreira@ifsuldeminas.edu.br`

Abstract. This study addresses the global public health challenge of infant mortality (IM) by developing predictive machine learning models using nationally representative data from Brazil's DATASUS system. Leveraging birth (SINASC, 2018–2022) and death records (SIM, 2018–2023), the research rigorously preprocesses data, applies probabilistic entity matching, and enriches features via Principal Component Analysis (PCA). To counteract severe class imbalance (infant deaths vs. non-deaths), multiple sampling techniques are evaluated: undersampling (Random Undersampler, Edited Nearest Neighbours), oversampling (Random Oversampler, ADASYN), and hybrid methods (SMOTETomek, SMOTEENN), both with and without PCA. The XGBoost classifier is employed with default hyperparameters. Key findings indicate that Random Undersampler (RU) achieves the highest recall (0.8031), critical for identifying true IM cases, while Random Oversampling (ROS) yields the best precision (0.6451), minimizing false positives. Edited Nearest Neighbours (ENN) with PCA achieves the optimal F1-score (0.4488), balancing precision and recall. The study concludes that sampling strategy selection should align with clinical priorities: RU for maximizing detection, ROS for reliability, and ENN for balanced performance. PCA integration showed negligible impact on results.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: Data Imbalance, Infant Mortality, Machine Learning, Resampling Techniques

1. INTRODUCTION

Infant mortality (IM) persists as a serious global public health problem, with WHO estimates indicating approximately 5.2 million deaths of children under five years of age in 2019, including 2.4 million among newborns [Organization et al. 2020]. Reducing IM is a critical target of the Sustainable Development Goals (SDGs), yet progress has been uneven and insufficient in various regions. This persistence is observed even in developed nations, such as the United States [Crawford 2025], the United Kingdom [Pillay et al. 2025] and France [Trinh et al. 2022], where socioeconomic factors and care gaps perpetuate challenges.

Early risk identification is fundamental for effective interventions and resource targeting [Batista et al. 2021]. In this context, Machine Learning (ML)-based predictive models emerge as promising tools, capable of identifying complex patterns associated with IM risk [Chivardi et al. 2023]. To make this possible however, the predictive model must be trained with clinical, demographic and socioeconomic data, thus ensuring its applicability in real-world scenarios [Batista et al. 2021].

However, an obstacle in building these models is the inherently rare nature of IM events within studied populations [Organization et al. 2020]. This scarcity results in severely imbalanced datasets, where the class of interest (`"death"`) is drastically underrepresented compared to the majority class (`"no death"`). This imbalance tends to bias conventional ML algorithms, optimized to maximize overall accuracy, leading them to favor the majority class [Kumar et al. 2021]. Consequently, models applied to imbalanced data struggle to identify rare IM cases (low sensitivity/recall), becoming clinically useless for prevention [He and Ma 2013].

Although the use of ML in child health is growing, studies that systematically investigate and compare the effectiveness of different imbalance treatment techniques specifically for IM prediction using data from the unified health system remain limited. Our contribution lies in providing practical evidence about optimal strategies for handling the rarity of IM events in predictive modeling, aiming to increase the clinical utility of these tools.

This work aims to fill this gap by investigating resampling techniques for predicting infant deaths based on national DATASUS data. We used records from SINASC (2018–2022) and SIM (2018–2023), applying rigorous preprocessing, probabilistic entity matching and enrichment via PCA. We evaluated undersampling methods (RU, ENN), oversampling methods (ROS, ADASYN), hybrid sampling methods (SMOTETomek, SMOTEENN) and their combinations with PCA, using the XGBoost algorithm for classification. Our results compare critical metrics (recall, precision, F1-score) and offer guidelines for technique selection according to clinical priorities, contributing to more equitable and effective models for preventing infant mortality.

The remainder of this article is organized as follows. Section 2 presents related works, highlighting previous studies on class imbalance handling in predictive modeling, especially within the healthcare domain. Section 3 details the materials and methods adopted, including data sources, preprocessing, entity matching, sampling strategies, and classification procedures. The experimental results and their analysis are presented in Section 4, focusing on the performance of different resampling techniques. Finally, Section 5 concludes the article, summarizing the findings and suggesting directions for future research.

2. RELATED WORK

In real-world application contexts, datasets will often have imbalanced classes relative to each other. The ratio between them affects the choice of technique, which is not always straightforward to determine. Due to this, there arises a need to compare models to ensure correct selection.

In [Silva et al. 2025], data from the SINASC database were explored for predicting premature births in Pernambuco state. The majority class (full-term births) represented 88.97% of records, with the minority class comprising the remainder. Three techniques were applied—oversampling, undersampling, and hybrid sampling—though the specific algorithm for each technique was not specified. Results show strong performance from the hybrid sampling technique, particularly when tripling the database size.

In parallel, [Wongvorachan et al. 2023] explored comparisons between oversampling, undersampling, and hybrid sampling more deeply by testing Random Oversampler (ROS), Random Undersampler (RUS), and SMOTE-NC + RU algorithms across two scenarios: moderate imbalance and extreme imbalance. The hybrid technique performed best, dramatically improving precision and recall for the chosen model. It was recommended by the authors for cases of extreme imbalance.

In [Khushi et al. 2021], various balancing algorithms were explored, with Random Oversampling producing the best results. Oversampling algorithms achieved higher performance than others, increasing the area under the curve (AUC) metric. Conversely, undersampling algorithms yielded poorer results.

In parallel, [Chakraborty et al. 2021] explored the efficacy of resampling techniques for bias mitigation, comparing the Fair-SMOTE algorithm against traditional approaches like SMOTE, Random Undersampling (RUS), and Random Oversampling (ROS), alongside state-of-the-art methods including Optimized Pre-processing (OP) and Fairway. Across 10 datasets with structural bias, Fair-SMOTE not only balanced class and sensitive attribute distributions but also eliminated biased labels via situation testing. Consequently, the technique outperformed all others: it reduced bias metrics (e.g., EOD by up to 97%) and increased Recall and F1 (in 29/36 scenarios) without compromising performance. The authors recommend Fair-SMOTE as an integrated solution for fairness and accuracy, especially in high social-sensitivity contexts.

3. MATERIALS AND METHODS

Figure 1 shows the steps followed during the development of this work. The databases used belong to the Live Birth Information System (SINASC), corresponding to the years 2018 to 2022, and the Mortality Information System (SIM), corresponding to the years 2018 to 2023. Both SIM and SINASC are public databases that can be accessed and downloaded via the DataSUS website[1]. SINASC records contain 49 attributes, including demographic data of the mother and newborn, gestational data, and data on delivery and birth. SIM records contain 59 attributes, including demographic data of the deceased and mother, as well as data on death.

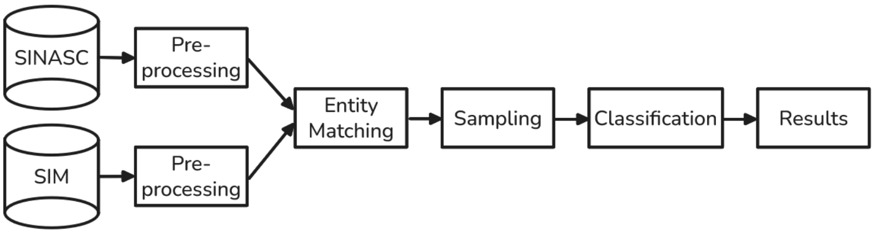

Fig. 1. General workflow followed in this work for predicting infant deaths under 1 year

3.1 Pre-processing

The data pre-processing stage is fundamental for improving data quality, thereby increasing the performance of supervised Machine Learning models [Maharana et al. 2022]. This is also an important step for the Entity Matching process as it aims to ensure that both databases used have consistent structures and similar formatting [Barlaug and Gulla 2021].

The first step in pre-processing was the removal of irrelevant records and attributes that lack useful information for predictive modeling. Records belonging to individuals aged one year or older were removed from the SIM database, reducing the number of rows from 7,749,150 to 162,848. Duplicate records were also removed, further reducing the total records in SIM to 162,734; SINASC had no duplicate records. Finally, attributes containing information about the mother, newborn/deceased, or delivery were selected. The selected attributes are shown in Table 0**??** and Table 0**??** for the SIM and SINASC databases, respectively.

Next, the databases were standardized to make their structures consistent. Among the standardized attributes were DTNASC, CODMUNRES, PESO, PARTO, GRAVIDEZ, GESTACAO, and ESCMAE. This process searched for categorical values outside the domain defined by the documentation and for dates (present in the DTNASC column) with typos (e.g., 31-12-2120). Additionally, it was decided to eliminate records with null values due to the use of techniques that do not accept such values. After removal, the SIM database had 102,220 records, while the SINASC database had 13,566,672.

Finally, data augmentation was applied using principal component analysis (PCA) to create new attributes based on semantic groupings. The sets are: IDADEMAE, RACACORMAE, ESTCIVMAE,

---

[1] https://datasus.saude.gov.br/

Table I. Selected SINASC attributes for use

Attribute	**Description**
DTNASC	Deceased's date of birth
IDADE	Deceased's age
SEXO	Deceased's sex
PESO	Birth weight (grams)
RACACOR	Race/Skin color
CODMUNRES	Municipality of residence code
PARTO	Delivery type
IDADEMAE	Mother's age
ESCMAE	Mother's education (years)
OCUPMAE	Mother's occupation type
QTDFILVIVO	Number of mother's living children
QTDFILMORT	Number of deceased children
GESTACAO	Gestational age ranges
GRAVIDEZ	Pregnancy type

Table II. Selected SINASC attributes for use

Attribute	**Description**
IDADEMAE	Mother's age in years
RACACORMAE	Mother's race/color
ESTCIVMAE	Mother's marital status
QTDFILVIVO	Number of living children
QTDFILMORT	Number of deceased children
PARIDADE	Defines if it's the first pregnancy
ESCMAE	Education level in years
CODMUNRES	Municipality code of residence
CODOCUPMAE	Mother's occupation code
GESTACAO	Gestational weeks classification
IDANOMAL	Congenital anomaly detected
GRAVIDEZ	Type of pregnancy
CONSPRENAT	Number of prenatal visits
CONSULTAS	Prenatal visits classification
MESPRENAT	Month of gestation when prenatal care began
PARTO	Type of delivery
DTNASC	Date of birth
APGAR5	Apgar indicator at 5th minute
PESO	Birth weight in grams
SEXO	Newborn's sex
RACACOR	Newborn's race/color

QTDFILVIVO, QTDFILMORT, PARIDADE, ESCMAE, CODMUNRES, and CODOCUPMAE to form the attribute MÃE (MOTHER); GESTACAO, GRAVIDEZ, CONSPRENAT, CONSULTAS, and MESPRENAT for the attribute PRENATAL (PRENATAL CARE); and IDANOMAL, PARTO, APGAR5, PESO, SEXO, and RACACOR for the attribute NEWBORN. This dimensionality reduction technique consolidates correlated features into semantically meaningful composite variables, enhancing both computational efficiency and clinical interpretability. By capturing shared variance within each thematic cluster, the PCA-derived attributes provide more robust representations of maternal health factors, prenatal care patterns, and neonatal outcomes while mitigating multicollinearity issues.

3.2 Entity Matching

Entity matching (EM) involves searching for records in different data sources that belong to the same person, with many different techniques available. Due to the high data quality and attribute consistency, the use of state-of-the-art Deep Learning techniques for EM was unnecessary, opting instead for more straightforward blocking and exact comparison methods between records [Barlaug and Gulla 2021].

The chosen technique, Probabilistic Entity Matching implemented by recordlinkage[2] Python library, consists of exact comparison of equivalent attributes between records from different databases. It works in three steps. The first, called blocking, groups records by each unique value in the selected attributes. The attributes used in this step were DTNASC and CODMUNRES. This step reduces computational complexity by pairing only records that agree on all attributes. Next, each previously generated pair has additional attributes compared: PESO, PARTO, GRAVIDEZ, SEMAGESTAC, and ESCMAE. Finally, each pair is classified as match or non-match according to the number of identical attributes found in the previous step. Only records with 5 identical attributes were considered matches, meaning all 5 attributes in addition to the first 2 must be identical between databases. However, this technique does not guarantee that found matches truly belong to the same person; this guarantee is only provided by Deterministic Entity Matching, which requires primary attributes not found in public databases. After this stage, a new attribute called OBITO (DEATH) was created in the SINASC database, where successfully paired records received value 1, and others 0.

While this probabilistic approach carries a risk of label noise (i.e., some matched records may not truly represent the same infant), this is an inherent challenge when using de-identified public health records lacking unique identifiers. Requiring exact matches on 7 attributes (DTNASC, CODMUNRES, PESO, PARTO, GRAVIDEZ, SEMAGESTAC, ESCMAE) significantly reduces this risk compared to fewer attributes. Potential mismatches would likely occur in scenarios with highly common attribute combinations (e.g., multiple infants born on the same day in a large municipality with identical weight and delivery type) or data entry errors. While this noise could slightly attenuate model sensitivity (Recall), the rigorous 7-attribute requirement provides a pragmatic balance between matching accuracy and computational feasibility for national-scale data.

3.3 Sampling

Due to significant imbalance between the classes of interest (infant death vs. non-death), sampling techniques were applied to improve predictive model performance. Implementations are available in Python's imbalanced learning library[3]. All techniques were applied to both original data and PCA-enriched data, which generated three new attributes (MAE [MOTHER], PRENATAL, NEWBORN) from semantic variable groupings.

The undersampling techniques used were: Random Undersampler (RU), which randomly reduces the majority class (non-death) until class record counts are equal. Edited Nearest Neighbours (ENN), which selectively removes majority class examples near decision boundaries based on classification inconsistencies among neighbors using k-Nearest Neighbours. The oversampling techniques were: Random Oversampler (ROS), which randomly replicates minority class (death) records. ADASYN (Adaptive Synthetic Sampling), which synthetically generates minority class examples focusing on regions where classification is harder, using adaptive density distribution.

The hybrid techniques (combining oversampling and undersampling) used were: SMOTETomek (ST), which applies SMOTE (Synthetic Minority Oversampling Technique) to synthetically generate minority class samples then removes redundant/ambiguous majority class examples using Tomek Links to refine decision boundaries. SMOTEENN (SE), which combines SMOTE to create synthetic minority samples with Edited Nearest Neighbours (ENN) to remove misclassified majority examples, promoting aggressive cleaning of overlapping regions.

---

[2] https://recordlinkage.readthedocs.io/en/latest/

[3] https://imbalanced-learn.org/stable/

3.4 Classification

For classification, the Extreme Gradient Boost algorithm[4] (XGBoost) was employed. This choice was made due to the model's prior success in previous implementations and its efficiency with large datasets. It is a boosting algorithm based on decision trees that sequentially combines weak predictors to form a strong predictor. In each iteration, the model creates a new tree to correct previous errors. Given the greater focus on sampling metrics in this work, default hyperparameters were used.

3.5 Results

The metrics used were Accuracy, Precision, Recall and F1-score. These were calculated from confusion matrix values: True Positive (TP), representing a class 1 (death) record correctly classified; False Positive (FP), when a class 1 record is misclassified as class 0 (non-death); True Negative (TN), when a class 0 record is correctly classified; and False Negative (FN), when a class 0 record is misclassified.

Accuracy (AC), Formula 1, is the proportion of correct predictions. Precision (PR), Formula 2, indicates the proportion of predicted positive instances that are truly positive. Recall (RE), Formula 3, measures the proportion of actual positive instances correctly identified. F1-score, Formula 4, represents the harmonic mean of Precision and Recall, balancing both metrics. This is particularly useful for imbalanced class distributions.

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$PR = \frac{TP}{TP + FP} \tag{2}$$

$$RE = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times (PR \times RE)}{PR + RE} \tag{4}$$

The presence of substantial class imbalance in the dataset rendered conventional accuracy measurements inadequate for assessing model efficacy. To address this, precision (measuring relevance of positive predictions), recall (measuring coverage of actual positives), and their harmonic mean (F1-score) were prioritized as robust evaluation metrics. These calculations were uniformly performed using the scikit-learn Python library[5].

4. RESULTS

The experiments evaluated various sampling and data balancing techniques applied to predicting infant death under one year of age. The machine learning algorithm used to create the predictive model was XGBoost, without hyperparameter tuning. Principal component analysis (PCA) was used as a data augmentation technique, creating three new attributes. Table III summarizes the results for the main evaluation metrics (accuracy, precision, recall and F1), comparing undersampling, oversampling and combination methods. The best results for each metric are highlighted in bold. Metrics were calculated considering death classification as the positive class.

---

[4] Available at: https://xgboost.readthedocs.io/en/stable/

[5] https://scikit-learn.org/stable/

Table III. Results acquired after applying the algorithms

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Baseline	0.9925	0.5512	0.2429	0.3373
Baseline + PCA	0.9924	0.5400	0.2485	0.3403
Random Undersampler	0.9142	0.0694	**0.8031**	0.1278
Random Under Sampler + PCA	0.9143	0.0694	0.8025	0.1279
Edited Nearest Neighbours	0.9923	0.5108	0.3904	0.4425
Edited Nearest Neighbours + PCA	0.9924	0.5110	0.4001	**0.4488**
Random Oversampler	0.9931	0.6425	0.2807	0.3907
Random Oversampler + PCA	0.9931	**0.6451**	0.2797	0.3902
ADASYN	0.9918	0.4725	0.3792	0.4207
ADASYN + PCA	0.9920	0.4891	0.3656	0.4184
SMOTEENN	0.9917	0.5510	0.3120	0.3984
SMOTEENN + PCA	0.9915	0.5326	0.3120	0.3935
SMOTETomek	0.9913	0.5068	0.3147	0.3883
SMOTETomek + PCA	0.9914	0.5189	0.3	0.3802

All techniques achieved accuracy greater than 90%, but other metrics showed significant variation. Baseline refers to using only XGBoost without adding sampling techniques. The undersampling techniques, Random Undersampler (RU) and Edited Nearest Neightbours (ENN), achieved the best recall (0.8131) and best F1-score (0.4488) respectively. However, ENN's execution time was the longest among all techniques due to the large database size. These techniques used majority sampling strategy (technique modifies only the majority class), other hyperparameters remained unchanged.

For oversampling techniques, Random Oversampler stood out with the highest precision (0.6426), minimizing false positives, but with moderate recall (0.2808), being in complete opposition to RU which achieved the best recall but low precision. ADASYN showed balanced performance (F1-Score: 0.4208) but inferior to other methods. The sampling strategy was minority-focused (modifies only the minority class), other hyperparameters were unchanged.

Finally, for hybrid techniques, SMOTEENN showed moderate performance, achieving the best precision, 0.5510, and F1-score, 0.3984, among hybrid models. SMOTETomek also showed moderate performance, with the highest recall, 0.3147, in the group. The sampling strategy was minority-focused for the SMOTE algorithm, and majority-focused for ENN and Tomek. Combined techniques with PCA did not significantly alter results, maintaining the same trends observed in original methods.

These findings highlight key trade-offs inherent to each resampling strategy when dealing with highly imbalanced health data. The high recall observed with Random Undersampling reflects its ability to rebalance the dataset by reducing majority class instances, increasing sensitivity to infant death cases. However, this comes at the cost of much lower precision, due to a higher rate of false positives. In contrast, Random Oversampling improved precision by replicating minority instances, but increased the risk of overfitting, as indicated by its lower recall. ENN achieved the best F1-score, suggesting that selective removal of noisy majority examples near class boundaries helps balance recall and precision. Overall, the results reinforce that resampling technique choice should align with the specific clinical objective—whether maximizing detection (recall), minimizing false alarms (precision), or achieving balanced performance (F1-score).

5. CONCLUSION

This study explored the challenge of addressing class imbalance in predictive modeling for infant mortality (IM) risk using national data from Brazil's Unified Health System (DATASUS). Through the evaluation of multiple resampling techniques—undersampling, oversampling, and hybrid methods—our findings confirmed that resampling significantly impacts model performance, especially in terms of recall, precision, and F1-score. The results highlight that the choice of sampling strategy should be guided by clinical priorities: when the goal is to maximize the identification of true IM cases, Random Undersampling (RU) proved to be the most effective.

An additional contribution of this study was the incorporation of feature enrichment via Principal Component Analysis (PCA). Although PCA did not produce statistically significant improvements across the tested configurations, its use demonstrated potential for dimensionality reduction and variable grouping for future exploration.

However, this study has several limitations. First, the use of default XGBoost hyperparameters is a notable constraint, as tuning parameters such as learning_rate, max_depth, scale_pos_weight, and min_child_weight can significantly affect model performance, especially with imbalanced datasets. Our results thus represent baseline performance without tuning. Optimized hyperparameters would likely improve all metrics (Precision, Recall, F1) across sampling techniques and could alter the observed rankings. This decision aimed to isolate the effect of sampling methods, but future studies should include systematic hyperparameter optimization.

Future work should explore stratification methods, given the uneven distribution of mortality risk across socioeconomic and demographic groups, as well as advanced hyperparameter tuning and alternative classifiers, such as deep learning models or ensemble stacking.

ACKNOWLEDGMENT

The authors are grateful for the support of CNPq (funding: 445273/2023-2), FAPEMIG, Federal Institute of Southern Minas Gerais (IFSULDEMINAS) and Federal University of Alfenas (UNIFAL).

REFERENCES

Barlaug, N. and Gulla, J. A. Neural Networks for Entity Matching: A Survey. *ACM Trans. Knowl. Discov. Data* 15 (3): 52:1–52:37, Apr., 2021.

Batista, A. F., Diniz, C. S., Bonilha, E. A., Kawachi, I., and Chiavegatto Filho, A. D. Neonatal mortality prediction with routinely collected data: a machine learning approach. *BMC pediatrics* vol. 21, pp. 1–6, 2021.

Chakraborty, J., Majumder, S., and Menzies, T. Bias in machine learning software: why? how? what to do? In *Proc. of the 29th ACM Joint Meeting on European Software Engineering Conference and Symp. on the Foundations of Software Engineering.* ESEC/FSE 2021. New York, USA, pp. 429–440, 2021.

Chivardi, C., Zamudio Sosa, A., Cavalcanti, D. M., Ordoñez, J. A., Diaz, J. F., Zuluaga, D., Almeida, C., Serván-Mori, E., Hessel, P., Moncayo, A. L., et al. Understanding the social determinants of child mortality in latin america over the last two decades: a machine learning approach. *Scientific reports* 13 (1): 20839, 2023.

Crawford, L. *Impact of Ohio Senate Bill 265 on Infant Mortality Rate in Ohio.* Ph.D. thesis, Walden Univ., 2025.

He, H. and Ma, Y. Imbalanced learning: foundations, algorithms, and applications, 2013.

Khushi, M., Shaukat, K., Alam, T. M., Hameed, I. A., Uddin, S., Luo, S., Yang, X., and Reyes, M. C. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. *IEEE Access* vol. 9, pp. 109960–109975, 2021.

Kumar, P., Bhatnagar, R., Gaur, K., and Bhatnagar, A. Classification of imbalanced data: review of methods and applications. In *IOP conf. series: materials science and engineering.* Vol. 1099. IOP Pub., pp. 012077, 2021.

Maharana, K., Mondal, S., and Nemade, B. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings* 3 (1): 91–99, June, 2022.

Organization, W. H. et al. Infant mortality, 2020.

Pillay, T., Dawson, K., and Trenell, M. Infant mortality is rising in the uk—reducing modifiable risks can help reverse reverse the trend, 2025.

Silva, A., Rocha, E., and Endo, P. Evaluating how different balancing data techniques impact on prediction of premature birth - extended abstract – ctdgsi 2025. In *Anais Estendidos do XXI Simpósio Brasileiro de Sistemas de Informação.* SBC, Porto Alegre, RS, Brasil, pp. 111–114, 2025.

Trinh, N. T., de Visme, S., Cohen, J. F., Bruckner, T., Lelong, N., Adnot, P., Rozé, J.-C., Blondel, B., Goffinet, F., Rey, G., et al. Recent historic increase of infant mortality in france: A time-series analysis, 2001 to 2019. *The Lancet Regional Health–Europe* vol. 16, 2022.

Wongvorachan, T., He, S., and Bulut, O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information* 14 (1): 54, Jan., 2023. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.