Evaluating how different balancing data techniques impact on prediction of premature birth - Extended Abstract – CTDGSI 2025

Anna Beatriz Silva[1], **Elisson da Silva Rocha (coorientador)**[2],
Patricia Takako Endo (orientadora)[1]

[1]Universidade de Pernambuco (UPE)
Caruaru – PE – Brazil

{anna.beatrizs, elisson.rocha, patricia.endo}@upe.br

1. Motivation and goals

According to the World Health Organization (WHO) [WHO 2023], preterm births, defined as births before 37 weeks of gestation, are classified into three categories: extremely preterm (less than 28 weeks), very preterm (28-31 weeks), and moderate to late preterm (32-37 weeks). In 2020, approximately 13.4 million babies were born prematurely worldwide, accounting for 9.9% of live births, with one million deaths due to complications. Brazil has a preterm birth rate of 12%, twice that of European countries, with higher rates in the Northeast and Southeast regions. Preterm birth remains a significant public health challenge, necessitating effective prediction tools to mitigate risks and improve outcomes. Artificial intelligence (AI) models have shown promise in healthcare, including preterm birth prediction, but they often face the challenge of imbalanced datasets, where term births dominate over preterm cases [Akazawa and Hashimoto 2022]. This imbalance leads to biased learning and poor performance in identifying preterm births.

This study evaluates machine learning models using real data from the Sistema de Informação Sobre Nascidos Vivos (SINASC), focusing on sociodemographic and obstetric data from SUS. By avoiding reliance on biomarkers or genetic data, this research ensures practical applicability in low-resource settings. The study aims to assess the performance of machine learning models on imbalanced datasets and compare sampling techniques to improve prediction accuracy. It also highlights the practical implications of using SUS data for preterm birth prediction, aligning with Brazil's Grand Challenges in Information Systems Research (2016-2026). By transforming data into knowledge, this research contributes to addressing critical public health issues, offering tools for early identification of high-risk pregnancies. Ultimately, the study supports the strengthening of the SUS, promoting a more efficient, equitable, and evidence-based healthcare system for maternal and neonatal care.

2. Material and method

The dataset used in this study is sourced from SINASC, containing 526,368 records and 61 attributes related to pregnant women in Pernambuco, Brazil, from 2018 to 2021. SINASC, implemented by the Brazilian federal government in the 1990s, collects data on all live births nationwide, providing essential information for health system planning, policy development, and public health programs. The preprocessed dataset is publicly available on Mendeley Data[1].

[1] https://data.mendeley.com/datasets/z3ychcthm2/3

The data preprocessing comprised several steps: identifying relevant attributes, grouping age ranges, and removing empty data, duplicates, and outliers. The goal was to train models to predict premature birth, with two target classes: Preterm and Term. The first step involved manually excluding attributes irrelevant to the study, such as system attributes, medical or hospital data related to birth, fetal attributes, and certain maternal attributes [Lee and Ahn 2020]. Attributes considered critical for preterm birth prediction, based on existing literature, were retained. These included the mother's age during pregnancy, history of preterm birth, multiple pregnancies, chronic diseases, infections, genetic influences, nutritional factors, and lifestyle.

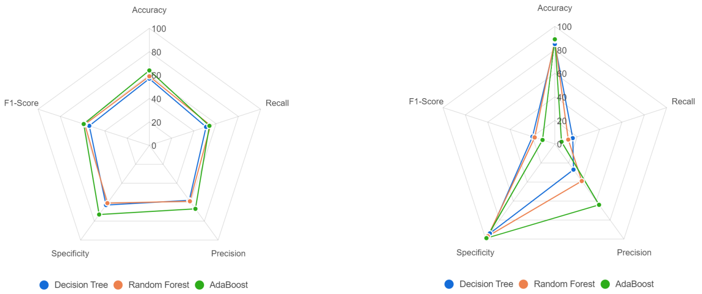
The mother's age distribution was analyzed due to its wide range of values. Age intervals were defined based on previous studies, which highlight age-related risks for preterm birth [Fuchs et al. 2018, Waldenström et al. 2017, Auger 2012, Schempf et al. 2007]. The selected intervals were 0-19, 20-24, 25-29, 30-34, and 35-54 years. Empty or duplicate records were removed to maintain data integrity. The final dataset included 15 sociodemographic and obstetric history attributes, as health history data was unavailable. The dataset was then divided into Term and Preterm classes. The Term class represented 88.97% of the data, while Preterm accounted for the remainder. To ensure realistic testing and avoid data leakage [Rosenblatt et al. 2024], 30% of the minority class (Preterm) and an equal number of records from the majority class (Term) were set aside, totaling 16,690 records. This approach ensured balanced and valid model evaluation across different scenarios.

The experiments involved creating distinct scenarios using different datasets based on data balancing techniques: Undersampling, Oversampling, and Hybridsampling. These datasets were used as input for machine learning models, including Decision Tree, Random Forest, and AdaBoost. Grid Search was employed to select the models' hyperparameters, with accuracy as the evaluation metric. Each scenario had its own set of hyperparameters. In the Undersampling approach, the majority class was reduced to match the minority class, resulting in 38,614 records per class. For Oversampling, the minority class was increased to equal the majority class, yielding 411,689 records per class. Hybridsampling involved creating synthetic data by doubling, tripling, or quadrupling the minority class size. The first scenario doubled the minority class and reduced the majority class to 77,228 records each. The second scenario tripled the minority class, balancing it at 115,842 records, while the third scenario quadrupled it, resulting in 154,456 records per class.

3. Results and discussion

Figure 1 presents a radar chart that distributes the evaluation metrics values on a scale of 0 to 100. The results obtained with Undersampling (Figure 1a) showed relatively low performance across all models, particularly in sensitivity, which measures the ability to correctly identify positive cases. Sensitivity values were below 55%, with Decision Tree at 51%, Random Forest at 55%, and AdaBoost at 54%. Specificity, however, showed slightly better results, reaching up to 73% for AdaBoost. This indicates that while Undersampling improved the detection of negative cases, it struggled with accurately identifying the minority (positive) class.

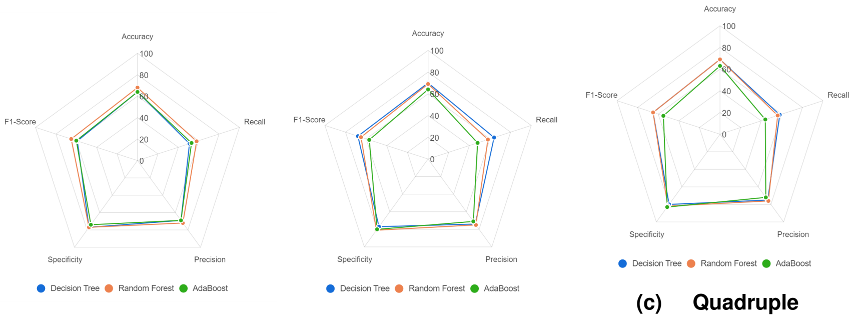
Figure 1b presents the models' performance with the Oversampling technique, which increased the Preterm class by 403,879 synthetic data points, resulted in high accuracy,bove 85%, but extremely low sensitivity, below 15%. Decision Tree achieved 16%, Random Forest 12%, and AdaBoost only 6%. In contrast, specificity exceeded 90% in all cases, reaching 99% for AdaBoost. This stark disparity highlights the inefficacy of Random Oversampling in improving positive class detection and suggests a strong bias toward the majority class.



(a) Undersampling **(b) Oversampling**

Figure 1: Models' performance when Undersampling and Oversampling techniques were used to balance the training dataset

Hybridsampling techniques yielded more balanced results. In the double size scenario, as showed in Figure 2a, accuracy ranged from 64% to 68%, with Random Forest achieving the highest sensitivity at 58%. The triple size scenario produced the best overall results, highlighted in Figure 2b, with Decision Tree achieving 70% accuracy, 64% sensitivity, and an F1-score of 68%. However, extending the balancing strategy to quadruple size led to a decline in performance, as showed in Figure 2c, with sensitivity dropping to 44% for AdaBoost. Hybridsampling triple size emerged as the most effective approach, with Decision Tree demonstrating the best balance between sensitivity (64% of sensitivity), specificity (77% of specificity), and precision (74% of precision), underscoring the benefits of strategic data augmentation while cautioning against excessive synthetic data generation.



(a) Double size **(b) Triple size** **(c) Quadruple size**

Figure 2: Models' performance when the Hybridsampling technique was used to balance the training dataset

Data balancing techniques significantly impact model performance. Undersampling reduced training data, degrading performance, while oversampling resulted in low recall and F1-scores, with a bias toward the majority class despite high accuracy. Hybrid-sampling, particularly when tripling the dataset, emerged as the most effective approach, yielding the best metrics and suggesting that models benefit from more extensive minority class data. The findings emphasize the importance of data balancing, especially in healthcare, where negative outcomes are less frequent.

4. Conclusion and future work

This work analyzed various data balancing techniques to evaluate machine learning models for preterm birth prediction, highlighting their distinct impacts on performance. While the results are promising, limitations include the dataset's restriction to Pernambuco, which may not fully represent Brazil's diverse regions, and delays in data collection, which could affect the relevance of predictions given evolving healthcare trends.

As a scientific contribution, this study was published by PLoS ONE (h-index 435) [Silva et al. 2025]. Future work includes expanding the analysis with additional data types, such as medical images, and exploring other boosting algorithms like CatBoost and LightGBM. Ultimately, the goal is to integrate these models into real-world SUS scenarios to enhance preterm birth prediction and healthcare outcomes.

References

Akazawa, M. and Hashimoto, K. (2022). Prediction of preterm birth using artificial intelligence: a systematic review. *Journal of Obstetrics and Gynaecology*, 42(6):1662–1668.

Auger, Nathaliea, b. c. D. P. H. S. P. R. W. (2012). Estimating gestational-age-specific and cause-specific associations. *Epidemiology 23(2)*.

Fuchs, F., Monet, B., Ducruet, T., Chaillet, N., and Audibert, F. (2018). Effect of maternal age on the risk of preterm birth: A large cohort study. *PLOS ONE*, 13(1):1–10.

Lee, K.-S. and Ahn, K. H. (2020). Application of artificial intelligence in early diagnosis of spontaneous preterm labor and birth. *Diagnostics*, 10(9).

Rosenblatt, M., Tejavibulya, L., Jiang, R., Noble, S., and Scheinost, D. (2024). Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications*, 15(1):1829.

Schempf, A. H., Branum, A. M., Lukacs, S. L., and Schoendorf, K. C. (2007). Maternal age and parity-associated risks of preterm birth: differences by race/ethnicity. *Paediatric and Perinatal Epidemiology*, 21(1):34–43.

Silva, A. B., Rocha, E. d. S., Lorenzato, J. F., and Endo, P. T. (2025). Evaluating how different balancing data techniques impact on prediction of premature birth using machine learning models. *PloS one*, 20(3):e0316574.

Waldenström, U., Cnattingius, S., Vixner, L., and Norman, M. (2017). Advanced maternal age increases the risk of very preterm birth, irrespective of parity: a population-based register study. *BJOG: An International Journal of Obstetrics & Gynaecology*, 124(8):1235–1244.

WHO (2023). Who. preterm birth.