Predictive Signatures of Violence Against Women in Brazilian Health Records: A Comparative Study Using Classifier Committees and Logistic Regression

Rafaela M. Lucca[1], **Roseli A. F. Romero**[1]

[1]Institute of Mathematics and Computer Sciences (ICMC)
Graduate Program in Computer Science and Computational Mathematics
University of São Paulo (USP), São Carlos, SP, Brazil

rafaelalucca@usp.br, rafrance@icmc.usp.br

Abstract. *Violence against women is a serious public health problem in Brazil, yet predictive studies on its notified records remain scarce, and most existing work treats violence as a single phenomenon. This paper investigates whether the three main types of violence registered in the Brazilian Notifiable Diseases Information System (SINAN), namely physical, psychological, and sexual, have distinct predictive signatures. We work with 2.22 million notifications from women between 2018 and 2024 and compare a regularized logistic regression against two classifier committees, Random Forest and LightGBM, evaluated under a temporal train and test split. Our analysis prioritizes interpretability and calibration over raw performance. Results show that each type of violence has its own distinguishable signature, the non-linear gain over the linear model is marginal, and class balancing substantially degrades probability calibration in imbalanced outcomes. The model fails to detect roughly 44.9 percent of psychological violence cases, exposing how automated tools tend to mirror the underreporting already present in the source data.*

1. Introduction

Violence against women is a serious public health problem worldwide, and a particularly acute one in Brazil. The Brazilian Notifiable Diseases Information System, known by the acronym SINAN, registers cases of interpersonal and self-inflicted violence reported by health services across the country. Despite the richness of these records, the literature based on SINAN has been mostly descriptive or focused on temporal trends [Lima et al. 2025, Coelho et al. 2025], with limited attention to predictive approaches that combine multi-outcome comparison, statistical rigor, and interpretability.

The research problem addressed in this work is the absence of predictive, type-differentiated analyses of violence against women in SINAN. It remains unknown whether the distinct types share a common predictive structure or each carries its own signature. Concretely, this work investigates the following question: do the three main types of violence against women notified in SINAN, namely physical, psychological, and sexual, have distinct predictive signatures, or do they share the same underlying patterns? The answer carries direct policy implications. A positive finding would support the design of type-specific interventions, while a negative one would justify cross-cutting strategies that treat violence as a single phenomenon.

To address this question, we compare a regularized logistic regression, treated as the interpretable reference model, against two classifier committees: Random Forest and LightGBM. Both of these are themselves ensembles of decision trees. The choice

is methodologically motivated. The non-linear committee-based models are introduced as a hypothesis test on whether relevant non-linear structure exists in the data, rather than as candidates to beat the linear model. If the gain is small, the phenomenon can be considered essentially additive, and the interpretable model is preferable.

The contributions of this work are as follows. First, to the best of our knowledge, this is the first comparative study of predictive signatures across types of violence in SINAN, contrasting linear and ensemble classifiers. Second, we propose a defensible preprocessing pipeline with cardinality auditing, non-supervised grouping, deduplication of redundant columns, and regularized variable selection. Third, we report an explicit calibration analysis showing how class balancing degrades probability quality. Finally, we offer a critical reading of false negatives as a reflection of systemic underreporting, illustrated by the regional pattern of sexual violence in Northern Brazil.

2. Related Work

Studies on SINAN violence data have largely focused on epidemiological description and temporal trends. [Lima et al. 2025] analyzed a decade of notifications between 2014 and 2023, highlighting a sharp increase after 2020 and the overrepresentation of socially vulnerable women. [Coelho et al. 2025] used Prophet and Random Forest to project violence rates up to 2033, but treated violence as a single phenomenon rather than examining type-specific structure.

In the Brazilian context, recent machine learning applications focus mostly on individual-level prediction or auxiliary detection, rather than on SINAN structured data. [Souto et al. 2019] addressed acoustic scene classification for domestic violence using audio recordings, achieving 73 percent accuracy with a Support Vector Machine. More recently, municipal-level initiatives in Recife coupled electronic health records with SINAN data to flag women at risk of revictimization within a 90-day window. This is an individual-level early-warning approach, methodologically distinct from the descriptive and explanatory framing we adopt here.

Internationally, [González-Prieto et al. 2021] developed recidivism prediction models for intimate partner violence using Spanish data, emphasizing the ethical care required when deploying such systems. [Kozhevnikova et al. 2025], in a systematic review of machine learning for violence prediction, found that most studies emphasize discrimination metrics while neglecting calibration, a gap our work explicitly addresses. The pattern of comparing regularized linear models with tree-based ensembles is well established in broader healthcare prediction.

What distinguishes this study can be summarized in four points. First, we treat the three types of violence in SINAN as separate prediction problems and compare their predictive signatures. Second, we report a regularization comparison between L1, L2, and Elastic Net on SINAN data, which to our knowledge has not been done before. Third, we evaluate probability calibration and the trade-off induced by class balancing. Fourth, we frame the results in an explanatory perspective, rather than as an individual triage tool.

3. Materials and Methods

Figure 1 summarizes the full pipeline, from raw SINAN extraction to the comparative modeling stage. Each step is detailed in the subsections that follow.

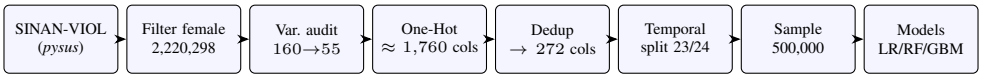


Figure 1. Overview of the preprocessing and modeling pipeline. Raw SINAN records are filtered, audited, encoded, and deduplicated before a temporal split and a stratified subsample feed the comparative modeling stage.

3.1. Data Source and Variable Selection

We use the SINAN-VIOL dataset, which corresponds to the Interpersonal and Self-inflicted Violence module, acquired through the pysus Python library. The study period covers 2018 to 2024, and therefore includes the pre-pandemic, pandemic, and post-pandemic intervals. After filtering for female sex, the dataset contains 2,220,298 notifications.

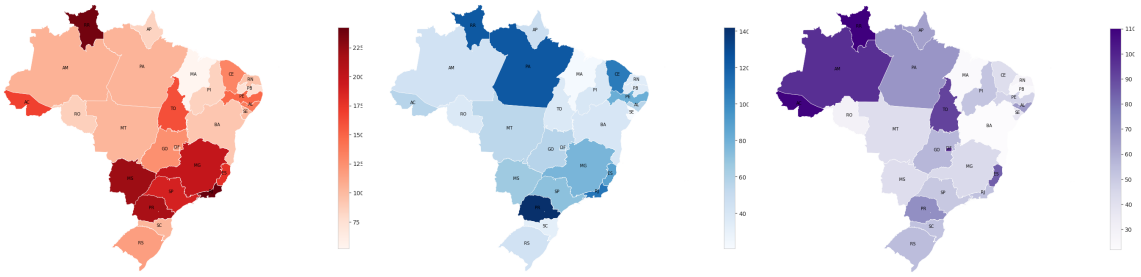
The original SINAN record contains 160 variables. We carried out a systematic audit to identify which of these variables could be used as predictors without compromising study validity. Four exclusion criteria were applied. We discarded post-event variables, such as referrals and case evolution, which would introduce causal leakage. We also discarded administrative identifiers, high-cardinality fields without additional information such as municipal codes with more than 5,000 unique values, and variables describing other types of violence or outcome details that would interfere with the three target outcomes. After this audit, 55 variables were retained, organized into nine thematic domains: victim profile, aggressor profile, occurrence context, means of aggression, target outcomes, recidivism, vulnerability, disability, and location.

3.2. Descriptive Profile

The notified population is predominantly young. The average age is 28.3 years and the median is 26 years, with a peak in the 15 to 20 age group. Brown-skinned women are slightly overrepresented (47.5 percent against 45.3 percent in the Brazilian population), while white women are underrepresented. Annual notifications show a clear upward trend that accelerates after 2020, a pattern that combines pandemic-related stress and improved notification coverage.

The aggressor profile by sex, computed as rates per 100,000 inhabitants of the same sex between 2020 and 2024, makes the gendered nature of the phenomenon explicit. Men are 2.6 times more likely than women to be the aggressor in physical violence cases, 6.7 times more likely in psychological cases, and 37.6 times more likely in sexual cases. Around 96 percent of sexual violence aggressors are men.

The geographic distribution of the three violence types reveals distinct spatial patterns. As shown in Figure 2, physical and psychological violence are comparatively diffuse, while sexual violence concentrates sharply in Northern states such as Roraima, Amazonas, Acre, and Tocantins, which report the highest notified rates per 100,000 women.



(a) Physical violence (b) Psychological violence (c) Sexual violence

Figure 2. Notified rate per 100,000 women per year by Brazilian state, 2020 to 2024, for (a) physical, (b) psychological, and (c) sexual violence.

This geographic pattern aligns with epidemiological evidence from the Amazon region. In the Marajo Archipelago, in the state of Para, 1,094 cases of sexual violence against children and adolescents were recorded between 2018 and 2022. Among the victims, 90 percent were female adolescents aged 12 to 17, and stepfathers and fathers account for more than 35 percent of the perpetrators [Bernardo and Rodrigues 2024]. The fact that our model identifies the Northern region as having the highest sexual violence rates is therefore not a statistical artifact. It reflects a documented social pattern of vulnerability tied to geographic isolation, limited access to protection services, and the prevalence of intrafamilial perpetration.

3.3. Preprocessing and Cardinality Audit

We carried out a cardinality audit on all categorical variables. Those with more than 20 unique categories were grouped without using the response variable. The hour of occurrence was grouped into four epidemiological time bands, and the state of occurrence was aggregated from 27 federative units into five Brazilian macroregions, preserving the geographic signal. The year of birth and the state of notification were removed as redundant with the derived age variable and the state of occurrence, respectively. After One-Hot Encoding, the pairwise correlation analysis revealed structural redundancy among missing-value flags. We applied union-find deduplication, keeping the highest-variance representative from each group of columns with absolute correlation above 0.95. The dimensionality was reduced from approximately 1,760 to 272 columns, with no loss of substantive information.

3.4. Experimental Setup

The dataset was partitioned temporally. The training set covers 2018 to 2023, and the test set is the year 2024. This setup simulates prospective application, evaluating a classifier on future data, and is more rigorous than random splitting for longitudinal data with potential trends. Due to computational constraints, training was done over a stratified sample of 500,000 records, preserving the prevalence of each outcome. The statistical adequacy of this sample is supported by three observations. The 95 percent confidence interval margin for any proportion is 0.14 percent. The Events-Per-Variable ratio is approximately 44 for the rarest outcome, namely sexual violence, which is well above the recommended minimum of 10 [Harrell 2001]. Finally, differences in outcome prevalence between the sample and the full population remain below one percentage point. These arguments establish statistical adequacy rather than empirical stability across sample sizes, so a formal sensitivity analysis that varies the training-set size is left as future work, and the present results should be read under the sample size reported here.

3.5. Modeling Pipeline

We adopt a simple-to-complex philosophy in which logistic regression is the interpretable reference model, not a baseline to be defeated. Its coefficients provide direct interpretation as odds ratios with confidence intervals. We compare three regularization schemes. L2 (Ridge) shrinks coefficients without zeroing them. L1 (Lasso) [Tibshirani 1996] performs automatic variable selection by zeroing irrelevant coefficients. Elastic Net [Zou and Hastie 2005] combines L1 and L2. This comparison is a statistically defensible alternative to stepwise selection, whose confidence intervals are invalidated by the sequential nature of the selection.

For the classifier committees, we compare two ensemble-based approaches that are dominant methods for heterogeneous tabular data and span the complementary bagging and boosting paradigms: Random Forest [Breiman 2001], based on bagging, and LightGBM [Ke et al. 2017], based on gradient boosting. We also report CatBoost and a Multi-Layer Perceptron for completeness. These models are introduced as a hypothesis test: do they reveal relevant non-linear or interaction structure that the linear reference model cannot capture? All models were trained in two configurations, with and without class balancing, in order to measure empirically the trade-off between recall, which tends to improve with balancing, and probability calibration, which tends to deteriorate.

Rather than an exhaustive search, we used a fixed configuration chosen a priori, which keeps the comparison transparent and reproducible. The logistic regression uses the saga solver with $C = 1.0$ and, for Elastic Net, ℓ_1 ratio 0.5. Random Forest uses 100 trees with maximum depth 15 and a minimum of 20 samples per leaf. LightGBM uses 200 trees, learning rate 0.1, 63 leaves, and maximum depth 8. All models use a fixed random seed and are run with and without balanced class weights. A systematic hyperparameter search is left as future work.

3.6. Evaluation Metrics

In a social problem, the choice of metrics is an ethical decision and not only a technical one. A false negative represents a victim that the system did not identify, and this asymmetry motivates prioritizing recall and calibration over raw accuracy. We report recall, PR-AUC, and Balanced Accuracy as primary metrics, along with the Brier score [Brier 1950] for probability calibration. F1-score and ROC-AUC are reported as secondary metrics. Accuracy is reported for completeness only, since it tends to be inflated by the majority class in imbalanced outcomes.

4. Results

4.1. Outcome prevalence

Physical violence shows near-equal class balance, with 51.4 percent of positive cases. Psychological violence is moderately imbalanced, with 23.8 percent of positives, and sexual violence is more imbalanced, with 16.1 percent. This gradient guides the metric prioritization discussed in Section 3.

4.2. Pre-modeling: distinct bivariate signatures

Even before training any model, the bivariate analysis already provides preliminary evidence for the central hypothesis. Among the top 15 variables most strongly associated with each outcome by Cramer's V, only six are universal, meaning that they appear in all three types. These universal variables refer to the sex of the aggressor, the presence of self-injury, poisoning as a means of aggression, and self-perpetration. Against these,

15 exclusive associations are distributed across the three outcomes. Three of them are specific to physical violence and relate to means of aggression, alcohol use by the aggressor, and conjugal relations. Four are specific to psychological violence and concern threats, recidivism, and ex-partners. Eight are specific to sexual violence and split into two groups: the relation with the aggressor (unknown, known, stepfather) and the vulnerability of the victim (low schooling, pregnancy, and sexual orientation). The signatures are therefore distinguishable even before any model is trained.

4.3. Comparative model performance

Table 1 reports the performance of all models on the 2024 test set, in the configuration without class balancing, which preserves probability calibration.

Table 1. Performance on the 2024 test set, without class balancing. Rec stands for recall (sensitivity), which is our primary metric since a false negative is an undetected victim. F1 is the F1-score, AUC is ROC-AUC, PR is PR-AUC, and Brier is the Brier score (lower is better).

Outcome	Model	Rec	F1	AUC	PR	Brier
Physical	LR (L1)	0.831	0.849	0.916	0.919	0.108
	Random Forest	0.819	0.857	0.926	0.933	0.101
	LightGBM	0.821	0.867	0.934	0.942	0.092
	CatBoost	0.819	0.865	0.933	0.941	0.093
	MLP	0.819	0.866	0.934	0.942	0.092
Psychological	LR (L1)	0.503	0.587	0.869	0.671	0.113
	Random Forest	0.504	0.595	0.883	0.713	0.108
	LightGBM	0.551	0.627	0.891	0.734	0.103
	CatBoost	0.546	0.624	0.890	0.731	0.104
	MLP	0.545	0.625	0.892	0.735	0.103
Sexual	LR (L1)	0.722	0.763	0.948	0.828	0.059
	Random Forest	0.695	0.765	0.957	0.850	0.058
	LightGBM	0.772	0.809	0.968	0.885	0.048
	CatBoost	0.766	0.807	0.967	0.883	0.048
	MLP	0.790	0.813	0.968	0.886	0.048

The recall column reveals the central asymmetry of this problem. Without class balancing, recall reaches around 0.83 for physical violence but drops to roughly 0.55 for psychological violence, meaning the model misses nearly half of the real cases. The recall and calibration trade-off is examined in Section 4.6.

4.4. The non-linear gain is marginal

The improvement of the best non-linear model over the regularized logistic regression is consistently small. In ROC-AUC, the gain is 0.018 for physical violence, 0.023 for psychological violence, and 0.019 for sexual violence, all below any practical-relevance threshold. The overfitting analysis shows train and test gaps between -0.006 and $+0.006$ for all models, confirming the absence of significant overfitting. These results support the interpretation that the phenomenon is essentially additive, so the interpretable linear model captures most of the predictive structure and validates the choice of prioritizing interpretability.

4.5. The three predictive signatures

Table 2 summarizes the key odds ratios from the L1 logistic regression. Confidence intervals were obtained with statsmodels for the top features.

Table 2. Key odds ratios with 95 percent confidence intervals from the multivariate logistic regression with L1 selection. Values come from a statsmodels fit on a 50,000-record stratified subsample, expanded beyond the top selected features to include socially relevant variables such as age and mental disability.

Outcome	Variable	OR (95% CI)
Physical	Bodily force	22.45 [16.15, 31.19]
	Sharp object	14.50 [12.21, 17.22]
	Firearm	8.42 [6.72, 10.55]
	Conjugal relation	4.00 [3.63, 4.41]
Psychological	Ex-spouse	2.39 [2.19, 2.60]
	Residence as location	2.53 [1.92, 3.35]
	Threat	2.10 [1.50, 2.93]
	Recidivism	2.02 [1.92, 2.13]
	Self-injury	0.12 [0.10, 0.14]
Sexual	Aggressor sex (male)	7.37 [6.49, 8.38]
	Mental disability	3.70 [2.73, 5.02]
	Unknown aggressor	2.07 [1.85, 2.33]
	Stepfather	1.57 [1.27, 1.94]
	Age (per unit)	0.24 [0.22, 0.25]

The three signatures emerge clearly from the results. Physical violence is dominated by the instruments of aggression. Bodily force alone yields an odds ratio above 22, and the presence of a sharp object or a firearm also increases the odds substantially. A conjugal relation with the aggressor adds another factor of four, consistent with the dominant role of intimate partner violence in this outcome.

Psychological violence shows a different pattern, marked by persistence and intimacy. Recidivism doubles the odds of the case being classified as psychological, and ex-spouses appear as a strong marker, with an odds ratio of 2.39. The residence is the typical location of occurrence, suggesting that this is a violence that unfolds inside the home and tends to repeat over time. The absence of self-inflicted injury is strongly associated, with an odds ratio of 0.12 reducing the odds by 88 percent, since cases that present self-harm tend to be classified differently.

Sexual violence carries the most socially charged signature. The age of the victim is the strongest predictor. With an odds ratio of 0.24 per unit of age, younger victims face drastically higher odds of being notified for sexual violence. Mental disability emerges as a strong vulnerability marker, with an odds ratio of 3.70, signaling that women with mental disabilities are notably more represented among notified cases. The aggressor profile centers on men, with an odds ratio of 7.37 for male perpetrators, and on aggressors who are either unknown or intrafamilial, stepfathers in particular.

4.6. Calibration: the cost of class balancing

Figure 3 contrasts Brier scores with and without class balancing. For physical violence, where the classes are nearly balanced, the impact is negligible. For the two imbalanced outcomes, class balancing substantially degrades probability calibration. In the logistic regression, the Brier score increases by 0.040 for psychological violence and by 0.032 for sexual violence. The trade-off is real and quantified: balancing improves recall but distorts predicted probabilities. For decision-support use, the unbalanced model is preferable, whereas for ranking-only use the balanced version may be acceptable.

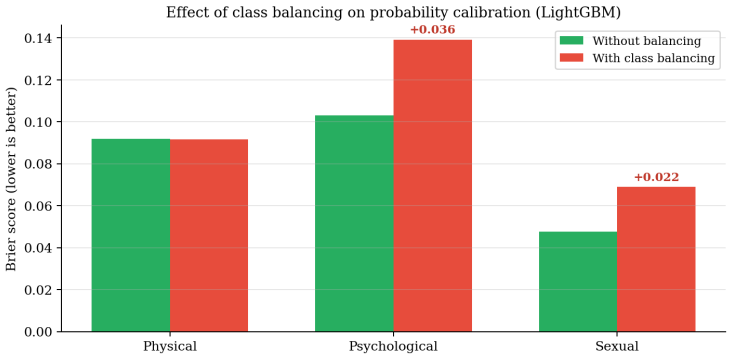


Figure 3. Effect of class balancing on probability calibration, measured by the Brier score using LightGBM. The balanced configuration substantially degrades calibration in imbalanced outcomes.

4.7. The cost of false negatives

The most socially impactful result emerges when the model's errors are translated into human terms (Figure 4). On the 2024 test set, the LightGBM model fails to detect 17.9 percent of physical violence cases (73,734 women), 22.8 percent of sexual cases (33,531 women), and 44.9 percent of psychological cases (approximately 84,988 women). Psychological violence is the hardest type for the model, because the phenomenon itself has a more diffuse signature on the notification form. There is no specific instrument of aggression and there are fewer direct markers. The model mirrors the systemic underreporting already present in the health information system.

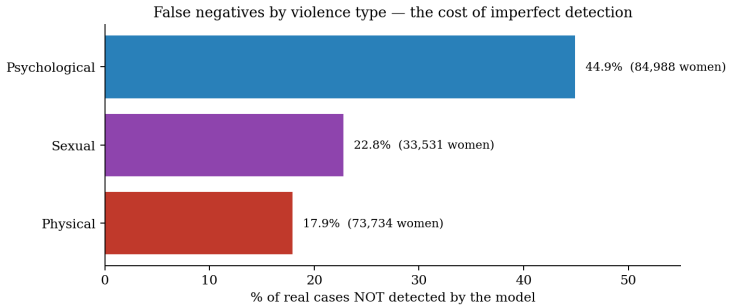


Figure 4. False negatives by violence type. Psychological violence shows nearly twice the rate of missed cases compared to the other types, which points to where the notification system itself needs to be strengthened.

5. Discussion

The empirical finding of three distinct predictive signatures, confirmed both at the bivariate level and at the modeling level, supports the hypothesis that different types of violence demand type-specific public policy responses. Strategies that work for physical violence will not address the diffuse pattern of psychological violence, nor the relational and age-driven profile of sexual violence.

The geographic concentration of sexual violence in Northern states aligns with a documented epidemiology of vulnerability in Amazonian Brazil. The Marajo Archipelago case illustrates the mechanism. The difficulty of access to riverine communities, the limited fiscalization, and the high incidence of intrafamilial violence combine to produce both high real prevalence and concentration in specific territories [Bernardo and Rodrigues 2024]. The model identifies stepfathers (OR 1.57) and unknown aggressors (OR 2.07) as significant markers of sexual violence, and this matches

the Marajo field data, where stepfathers and ex-stepfathers account for 22 percent of cases and fathers for another 13 percent. In other words, the predictive signature recovers a social pattern documented qualitatively in the literature.

Psychological violence is the most challenging type for the model to detect. We argue that this is not a failure of the model, but a faithful mirror of a notification system that struggles to capture diffuse, non-physical forms of violence. The 44.9 percent false-negative rate represents roughly 85,000 women whose experiences may be missed by any automated tool built on these data. Several technical mitigations could reduce this rate without discarding calibration, and we outline them as future work. Cost-sensitive learning that penalizes false negatives directly, rather than naive class balancing, can raise recall while limiting the calibration damage we documented. A calibration-aware decision threshold can then meet a target recall while monitoring the Brier score, and alternative boosting losses such as the focal loss concentrate learning on the hard, ambiguous cases that characterize diffuse minority signatures. None of these can fully overcome under-notification at the source, so improving prediction here ultimately requires improving the notification instrument itself, not only the modeling algorithm.

6. Ethical Considerations and Limitations

This study is framed as an effort in epidemiological understanding. It identifies factors associated with each type of violence in order to inform public policy. It is not, and should not be used as, an individual triage instrument, since such a use would carry risks of stigmatization, a false sense of security, and reinforcement of existing biases.

SINAN records notified violence, not experienced violence. Women with less access to health services, such as those in rural areas, peripheries, or the Amazonian interior, are underrepresented in the data. The model learns notification patterns, not victimization patterns. The pandemic period also adds heterogeneity, since notification dynamics shifted markedly between 2020 and 2022.

Most disability-related fields in SINAN are systematically underfilled and did not emerge as predictors. The exception is mental disability, which appears as a strong predictor for sexual violence, with an odds ratio of 3.70. This is a meaningful finding: it signals that women with mental disabilities face a particular vulnerability that the notification system manages to capture, even when other disability fields remain blank. All findings reported here are associational, since the study design is observational and cross-sectional and the variables flagged as predictors are not necessarily causes. Regarding the use of AI tools, ChatGPT and Claude were used to assist with linguistic revision and structural drafting. All technical decisions, results, and interpretations are the full responsibility of the authors.

7. Conclusion

We have presented a comparative study of predictive signatures for three types of violence against women in the Brazilian SINAN-VIOL data, contrasting a regularized logistic regression with two classifier committees. Three substantive conclusions emerge. First, each type of violence has its own distinct predictive signature, which supports type-specific policy responses. Second, the phenomenon is essentially additive, which means that the linear model is sufficient and preferable from the point of view of interpretation. Third, the false-negative patterns of the model mirror, rather than correct, the underreporting patterns of the source data.

The concentration of sexual violence in Northern Brazil, the relational signature pointing to intrafamilial perpetration, and the diffuse predictive structure of psychological violence each have direct implications for how public-health and protection systems should evolve. Future work includes subgroup-equity analysis, calibration-aware threshold optimization, and integration with complementary data sources such as police records and public-security databases, in order to triangulate notification biases.

Acknowledgments

We acknowledge the Brazilian Ministry of Health for the public availability of SINAN data, and the Graduate Program in Computer Science and Computational Mathematics (PPG-CCMC) of the University of São Paulo (USP) for its support.

References

Bernardo, A. M. C. S. and Rodrigues, A. L. L. d. S. M. (2024). Violência sexual contra criança e adolescente na amazônia: um panorama no arquipélago do marajó. Fonte Segura, Fórum Brasileiro de Segurança Pública.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Coelho, M. d. M. F., Oriá, M. O. B., Tamboril, A. C. R., Moreira, W. C., Oliveira, B. A. d., Martins, M. C., et al. (2025). Temporal evolution and projections of violence against women in Brazil: Scenarios from 2013 to 2033. *Journal of Interpersonal Violence*, page 08862605251355969.

González-Prieto, Á., Brú, A., Nuño, J. C., and González-Álvarez, J. L. (2021). Machine learning for risk assessment in gender-based crime. *arXiv preprint arXiv:2106.11847*.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

Kozhevnikova, S., Yukhnenko, D., Scola, G., and Fazel, S. (2025). Machine learning for violence prediction: a systematic review and critical appraisal. *arXiv preprint arXiv:2511.23118*.

Lima, G. C. C., Passos, C. M. d., Pinheiro, A. L. S., Ribeiro, Í. J. S., and Maia, E. G. (2025). Temporal trend and epidemiological profile of notifications of violence against women in Brazil: 2014–2023. *Epidemiologia e Serviços de Saúde*, 34:e20240475.

Souto, H., Mello, R., and Furtado, A. (2019). An acoustic scene classification approach involving domestic violence using machine learning. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, pages 705–716. SBC.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.