Article

Not peer-reviewed version

Detection of Patterns of Victimization and Risk of Gender Violence Through Machine Learning Algorithms

Edna Rocio Bernal-Monroy , Erika Dajanna Castañeda-Monroy , Rafael Ricardo Renteria-Ramos , Sixto Enrique Campaña-Bastidas * , Jessica Barrera , Tania Maribel Palacios-Yampuezan , Olga Lucia González Gustin , Carlos Fernando Tobar-Torres , Zeneida Rocio Ceballos-Villada

Posted Date: 4 October 2024

doi: 10.20944/preprints202410.0314.v1

Keywords: Data science; Machine learning; Pacific; San Andrés de Tumaco; Gender violence; Violence against women

 doi:10.20944/preprints202410.0314.v1

Article

Detection of Patterns of Victimization and Risk of Gender Violence Through Machine Learning Algorithms

Edna Bernal [1], Erika Castañeda [2], Rafael Rentería [1], Sixto Campaña [1,*], Jessica Barrera [3], Tania Palacios [1], Olga González [1], Carlos Tobar [1] and Zeneida Ceballos [1]

[1] Universidad Nacional Abierta y a Distancial (UNAD); edna.bernal@unad.edu.co (E.B.); rafael.renteria@unad.edu.co (R.R.); tania.palacios@unad.edu.co (T.P.); lucia.gonzalez@unad.edu.co (O.G.); carlos.tobar@unad.edu.co (C.T.); zeneida.ceballos@unad.edu.co (Z.C.)
[2] Universidad Pedagógica y Tecnológica de Colombia (UPTC); erika.castaneda01@uptc.edu.co
[3] Centro de Investigación y Desarrollo Tecnológico en Ciencias aplicadas (CIDTCA); jnbarrera191@gmail.com
* Correspondence: sixto.campana@unad.edu.co

Abstract: This paper explores the application of machine learning techniques and statistical analysis to identify patterns of victimization and the risk of gender-based violence in San Andrés de Tumaco, Nariño, Colombia. Models were developed to classify women according to their vulnerability and risk of suffering various forms of violence, which were integrated into a decision-making tool for local authorities. The algorithms employed include K-means for clustering, artificial neural networks, random forest, decision trees, and multiclass classification algorithms combined with fuzzy classification techniques to handle incomplete data. Implemented in Python and R, the models were statistically validated to ensure their reliability. Analyses based on health data revealed key patterns of victimization and risks associated with gender-based violence in the region. This study presents a data science model that uses a social determinants approach to assess the characteristics and patterns of violence against women in the Pacific region of Nariño. The research was conducted within the framework of the Orquídeas Program of the Colombian Ministry of Science, Technology, and Innovation.

Keywords: data science; machine learning; Pacific; San Andrés de Tumaco; gender violence; violence against women

1. Introduction

Violence takes many forms. Historically, the impact and type of violence have differed depending on gender, with violence against women presenting the most alarming statistics [1,2]. Worldwide, according to the United Nations Entity for Gender Equality and the Empowerment of Women (ONU Women), one in three women has experienced violence in her lifetime, with intimate partner violence being the predominant form. Despite efforts to reduce this type of violence worldwide, in 2022, more than five women or girls were killed by a family member every hour [1,3].

Violence against women is even more severe [2,4,5] in countries with internal armed conflicts, such as Colombia. In the Pacific region of Nariño, the township of San Andrés de Tumaco is one of the areas most affected by this phenomenon, where violence persists and impacts the health and well-being of all its inhabitants. According to [6], this district has 27.52% unsatisfied basic needs, 53.7% multidimensional poverty, and a high rate of victimization. This bulletin also reports that 52% of the victims in San Andrés de Tumaco are women, the highest figure in the country.

The study [7] presents research on gender gaps that reveals that, after age 40, a significant percentage of women face limitations in their daily performance due to time constraints that affect

 doi:10.20944/preprints202410.0314.v1

rest and the realization of physical and mental self-care activities, concluding that these impacts are acquired and preventable [1,7,8].

Gender-based violence significantly affects women and is a public health and safety problem [1,8]. According to the National Demographic and Health Survey (NDHS), between 2010 and 2015, there was an increase in the percentage of women aged 15-49 years who did not seek help in cases of violence, possibly reflecting a lack of access to support resources or a reluctance to seek help [9]. The report "Cifras Violeta VI" of the Gender Observatory of Nariño presents data on sexual, family, and intimate partner violence between 2015 and 2019. It shows an annual increase in each aspect [10], pointing to the problem of violence that has been experienced in this territory for several years.

Violence against women in San Andrés de Tumaco, Nariño, is complex and requires precise characterization to implement effective intervention strategies [8,11]. The social determinants of health, based on gender-based violence, are fundamental to recognizing the actors involved and the various forms of violence faced by women beyond the commonly reported physical violence, including the identification of psychological sequelae and violence [12,13].

Establishing patterns in official national health reports is a pivotal tool for characterizing the needs and the violence faced by women in the district. In this sense, classification techniques and identification of characteristics through Machine Learning (ML) are relevant, given their versatility. In the world, the study of these algorithms for identifying and analyzing patterns of gender violence has gained momentum, as shown below.

This article is organized as follows: Section 2 details the review of works related to our proposal; Section 3 presents the methodology used for pattern detection; Section 4 shows the results of analyzing health reports with ML algorithms; Section 5 presents the discussion of the results; and, finally, Conclusions are offered.

2. Related Works

Data analysis algorithms for pattern detection using ML techniques have offered innovative alternatives for studying gender-based violence. For example, in [14], a deep neural network model is implemented to identify gender violence in Twitter messages in Mexico, achieving an effectiveness of 80%. Similarly, [15] proposes using ML and big data techniques to predict the incidence of gender violence in Spain based on data collected over a decade. The study demonstrates the possibility of predicting the number of reports of violence using a multi-objective evolutionary search approach for the selection of variables and random forest algorithms.

In [16], text mining, data linkage, and deep learning techniques are employed in police and health records to predict future family and domestic violence crimes, easing the complex data analysis and improving the accuracy of incident prediction.

Also, studies such as [17] discuss how latent variable and clustering methods have been used differently in intersectional approaches. In [18], a methodological approach based on a multivariate model is applied in a spatio-temporal context to analyze areas of high incidence of violence against women.

The above is related to what [19] mentions, explaining how applying different ML techniques for studying and identifying patterns of gender violence represents, worldwide, an efficient alternative. However, these authors also indicate that, in Colombia, no research has been undertaken to implement these technologies in the study of gender violence issues. The importance of harnessing new technologies to identify the population's needs and to generate targeted intervention strategies is highlighted.

In this context, this research developed a tool for screening patterns of victimization and risk of gender violence using ML algorithms.

3. Methodology

This section discusses implementing a decision-making tool in San Andrés de Tumaco, Colombia, to analyze its effects on health using ML algorithms. It includes the following stages:

1. Identification of sources, data collection and preparation

 doi:10.20944/preprints202410.0314.v1

2. Exploration of ML techniques for pattern detection
3. Evaluation of the techniques in pattern detection

3.1. Identification of Sources, Data Collection and Preparation

Health data related to suicide attempts, poisonings, and gender-based violence were selected for this study, considering the first two as possible consequences of the latter. Some studies have shown that gender violence is closely related to an increase in suicide attempts and self-injurious behaviors [20]. Intoxication can be a direct or indirect consequence of these outbreaks of violence, either as a form of self-injury or due to substance abuse as a coping mechanism [21,22].

In Colombia, the Public Health Surveillance System (SIVILIGA) reports this type of data in events 356, 365, and 875, respectively. The SIVIGILILA provides a detailed framework for the notification and follow-up of these events, organized into two principal sections for filling out the information: Basic data and complementary data. The basic data perform a sociodemographic analysis of the event and classify the reported cases according to their characteristics; these data are organized in 82 fields available in the notification form of each report, whereas the complementary data explicitly describe the event.

Of the latter, event 356 uses 52 fields to collect information on suicide attempts, including details on circumstances, methods, risk and protective factors, and sociodemographic and clinical characteristics. Event 365 uses 49 fields to report chemical poisonings, recording the circumstances of exposure, symptoms, toxic agents, and occurrence area. Event 875 has 41 fields to monitor cases of gender and domestic violence, characterizing the victim and the aggressor and studying each case [23]. Collecting the information of these events and their databases was given thanks to the support of the district mayor's office of San Andrés de Tumaco; 737 reports were gathered between 2017 and 2023 for event 356, 708 for event 365, and 3680 for event 875.

The selection of variables for analyzing these databases was focused on the preliminary exploration of the patterns of vulnerability and risk of violence against women in the three events based on the data shared by these databases. In this sense, we identified similarities in the reports describing a "woman in sitiation of violation" rather than the analysis of the event that occurred. The complementary data study pursued identifying the patterns of occurrence of the different kinds of violence and the connection they may have with each other. Annex 1 presents the basic and complementary data variables for each SIVIGILA event collected.

3.2. Exploration of ML Techniques for Pattern Detection

The exploratory analysis for detecting patterns of risk of occurrence of gender violence in San Andrés de Tumaco relies on the characteristics established in the SIVIGILA event databases, so the use of algorithms such as decision trees, random forests, Artificial Neural Networks (ANNs), and clustering are the techniques that usually show better results for analyzing patterns in small datasets [14,24].

The analysis for each of the datasets included the search for similarities and correlations between suicide attempts and intoxications with gender and intrafamily violence to explain the recurrence of these events. In these cases, a balancing was performed on the datasets by subsampling to avoid overfitting the algorithms when analyzing the patterns in the basic data since the number of reports of suicide attempts and intoxications is lower than the cases of gender and intrafamily violence.

Decision trees are a suitable tool for these cases, as they are popular in pattern detection and decision-making due to their interpretability and ease of use. Research such as [25] or [26] have demonstrated their effectiveness in identifying risk factors associated with gender-based violence.

In this study, we employed decision trees to understand the relationships between different variables and the events of interest, providing a solid basis for predictions and initial exploratory analysis. Identifying the relationship among the variables that clearly and in detail characterize gender violence allows us to recognize and explain those cases, suicide attempts, and intoxications, which possibly occur as a consequence of an episode of gender and intrafamily violence. For this purpose, the most relevant variables were selected, the cut-off point was maximized, and the

 doi:10.20944/preprints202410.0314.v1

impurities caused by data that offer little explainable information were minimized thanks to the evaluation of impurity measures such as the Gini index and entropy [27].

Decision trees were applied to patients referred to psychiatry and social work as a possible predictor of relevant characteristics of women with suicide attempts, presumably due to acts of abuse in different forms. Likewise, to disaggregate the phenomenon of intoxication in women in San Andrés de Tumaco, this technique was implemented according to the classification of the alert situation in the reports; it searches for common patterns with the reported cases of gender violence.

In addition, a random forest with the same analysis parameters of the relationship between the events was explored to mitigate the limitation of overfitting. Random forests average multiple decision trees and present the best results for dataset analysis, thanks to their outstanding handling of data variability. This technique has been successfully used in studies of gender-based violence to improve the accuracy of predictions [2,24,25,28].

Random forests are applied to available health events as internally multiple subgroups of the original dataset are created by sampling with replacement; consequently, for each subgroup, a decision tree is elaborated, and, at each node, a subgroup of features of size "mtry" is randomly selected. The best-split point is only determined from the selected features. The maximum depth of the trees is limited with the parameter "max_depth", and cross-validation is used to choose the most efficient number of trees, recognizing the decision capacity of the algorithm.

Other powerful ML algorithms that can capture complex nonlinear relationships in the data are ANNs, as well as the techniques mentioned, which have also been applied in studies of gender violence pattern detection with promising results [14].

The ANNs' model for the available SIVIGILA health events was implemented by searching for the number of hidden layers and neurons that would fit the amount of input and output data of the network and the classification in each event; likewise, a review was performed to determine the activation function that would best contribute to the algorithm for the type of data available, accounting for optimization parameters such as number of neurons (input and output) and layers (hidden), and the activation algorithm for a better adjustment of weights. The evaluation methods in Section 3 tested the degree of assertiveness when implementing these algorithms.

Moreover, implementing clustering methods permitted the identification of groups of potential victim characteristics and possible collectives that, despite suffering gender-based violence, may present some other pattern(s) of vulnerability(ies) associated with it [17]. Clustering methods, such as K-means and K-medoids, identified groups within data that share characteristics. K-means is suitable for continuous numerical data, while K-medoids can handle both numerical and categorical data and is more robust to outliers. The latter is applied to a dataset in an unsupervised form, in contrast to the previous algorithms. In the context of gender-based violence, these methods can reveal hidden patterns and subgroups of victims with similar behaviors and experiences. It can be valuable for recognizing groups that tend to normalize violence and that other methods do not detect.

For the recognition of patterns of victimization and risk of gender violence, a robust and diverse approach was applied based on the exploration of decision trees, random forests, and ANNs. It perceives the influences of variables and the connection between health events in an optimized model that detects the minimum viable number of variables, explaining these phenomena through a prototype with at least 70% assertiveness. Therefore, it considers the limited number of observations per type of aggression so that the phenomenon of violence against women in San Andrés de Tumaco is explainable from the analysis of health data.

This study prioritized the analysis with Partitioning Around Medoids (PAM) clustering technique to display the risks and patterns of victimization and focus on the group of women most prone to gender-based violence. In this way, territorial entities can decide on targeted intervention strategies to analyze actual data on their population. The methodology maximized the capacity of the algorithms to address the complexity and variability inherent to the data in this type of study, providing a more complete and accurate perspective of the phenomenon.

 doi:10.20944/preprints202410.0314.v1

3.3. Evaluation of the Techniques in Pattern Detection

It was based on performance metrics such as precision, sensitivity, and specificity, as well as analysis through confusion matrices.

Besides, the interpretation of the relevance of the variables and the results with techniques such as Multiple Correspondence Analysis (MCA) enabled the identification of the most influential factors in the predictions [29–31]. Within the evaluation methods, simulations were also applied to model the distributions and frequencies of past events, taking into account only the variables obtained as minimum viable. These simulations were performed by adjusting probabilistic distributions to historical data and generating new samples that respected the statistical properties of the original data. They worked exclusively with the reports of event 875 to establish violence projections between 2025 and 2028, calculating the rates of gender-based violence in women in San Andrés de Tumaco from 2017 to 2023. Violence rates were applied to three decimal places. The simulations for projecting the figures were conducted in a decision-making tool developed with Power BI. It integrated the presentation of statistics from previous reports and their forecasts.

Furthermore, with the application of random forests, the best-performing algorithm was identified based on the cross-validation using the metrics of different possibilities.

4. Results

This section details the results after applying the different ML methods. For designing the event analysis model, a statistical analysis was performed to understand data behavior, the relationships between them, and their connection with the phenomenon of violence.

A first exploratory analysis complemented this study by prior cleaning and labeling the data on suicide attempts, intoxications, and gender and intrafamily violence. Few isolated cases with a predominance of violence were identified, standing out for their severe consequences for the health and life of the victims. This hindered the recognition of profiles of potential victims, as it highlighted the general context and covered the patterns of cases with lower alertness. Pointing out these cases facilitated a more contextual approach, considering the experience and type of violence suffered and not only the consequences of the event. Thus, a partial discrimination of the predominant cases was made to define the most common cases and understand the phenomenon and the situation of the victim and their environment.

Similarly, based on the exploration and recognition of the databases, the categories of the variables of the SIVIGILA 356, 365, and 875 events were compiled in the *Dictionary of variables* in Annex 1. It identified the characteristics of the information consolidated in this system based on the study by the officials of the health entities to report cases [32–34]. The dictionary recognizes the variables corresponding to the basic and complementary data in the reports of the events. It is worth noting that for the variables whose description is filled in as “unknown”, no specific definition was found in the SIVIGILA sources reviewed.

The confusion matrix shows that the decision tree has limited accuracy in predicting referrals to social work of women who attempt suicide. Of the 92 cases evaluated, 26 were correctly predicted as non-referred and 12 as referred. However, the model failed in 54 cases: 20 women were incorrectly predicted as referred (false positives) and 34 as non-referred (false negatives), indicating difficulties in correctly identifying cases that need a referral to social work. Of the databases provided by the mayor’s office of San Andrés de Tumaco, those with a significant number of unreported or empty records were identified, which influenced the exploratory process of selecting the variables that contribute most to the profiles of the victims.

The data in Table 1 were used for the statistical and exploratory analysis. The variables of events 356 and 365 were related to the detection of patterns in the algorithms, given the context from the researchers; however, when victimization patterns on the potential victim profiles were identified, the definitive variables were the area where the event occurred, the life cycle, the type of health service of the abused woman or girl, and whether the report indicated possible partner problems. Likewise, the nature of the violence and whether the woman or girl lives with the aggressor were

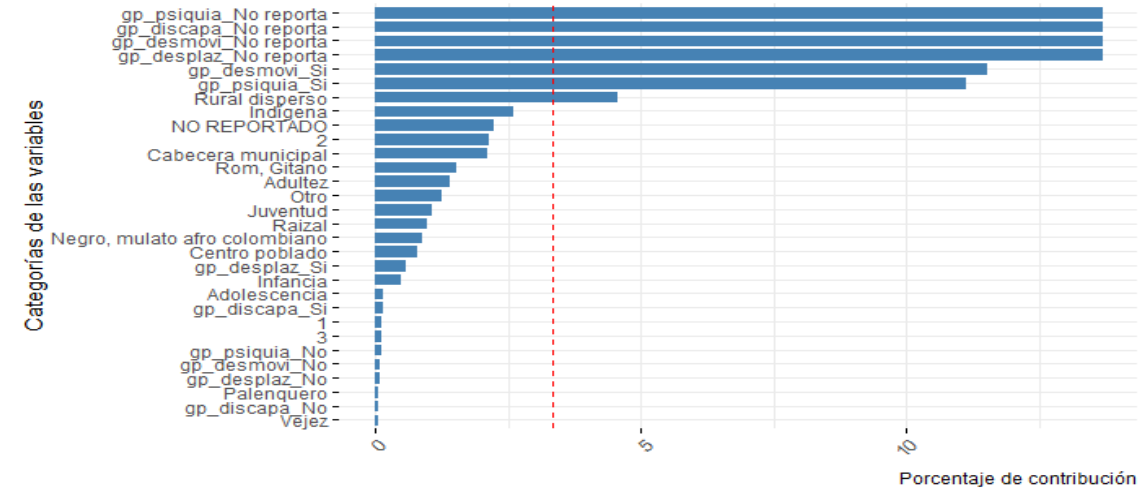
 doi:10.20944/preprints202410.0314.v1

analyzed, as well as whether she is in a situation of alert and was referred to a mental health or social work service.

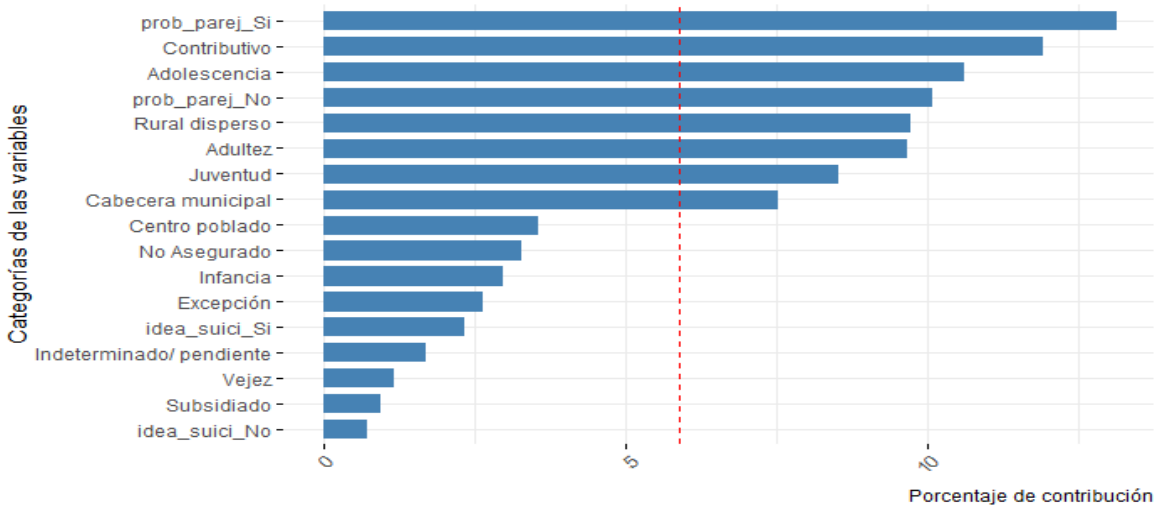
Table 1. Variables used in exploratory statistical analysis.

Variables	Event
area_,estrato_,ciclo_vida,per_etn_,gp_desplaz,gp_desmovi,gp_discapa,gp_psiquia,tip_ss_,inten_prev,pac_hos_,prob_parej,prob_econo, prob_legal, prob_labor, prob_famil	356
area_,estrato_,ciclo_vida,tip_ss_,pac_hos_,sit_ale,tip_exp, con_fin_,sit_ale,est_civ	365
area_, ciclo_vida, tip_ss_,naturaleza, conv_agre	875

In this same sense, variables related to socioeconomic level and population groups were identified. However, despite being a crucial sociodemographic factor for characterizing populations, this study did not report conclusive data on the phenomenon due to the lack of information and adequate documentation during case registration. Figure 1 illustrates how the categories of the variables analyzed throughout the research helped define the profiles of victims in events 356, 365, and 875. Figure 1a demonstrates the impact of unreported cases on variables related to population groups and socioeconomic strata. Figures 1b, 2 and 3 show the contribution of definitive variables to the victim profiles of each studied event.



(a)



(b)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2024 doi:10.20944/preprints202410.0314.v1

Figure 1. Contribution of variable category to victim profile of event 356. Attempted suicide in San Andrés de Tumaco. (a). Considering "not-reported" cases. (b). Discriminating the non-reported cases.

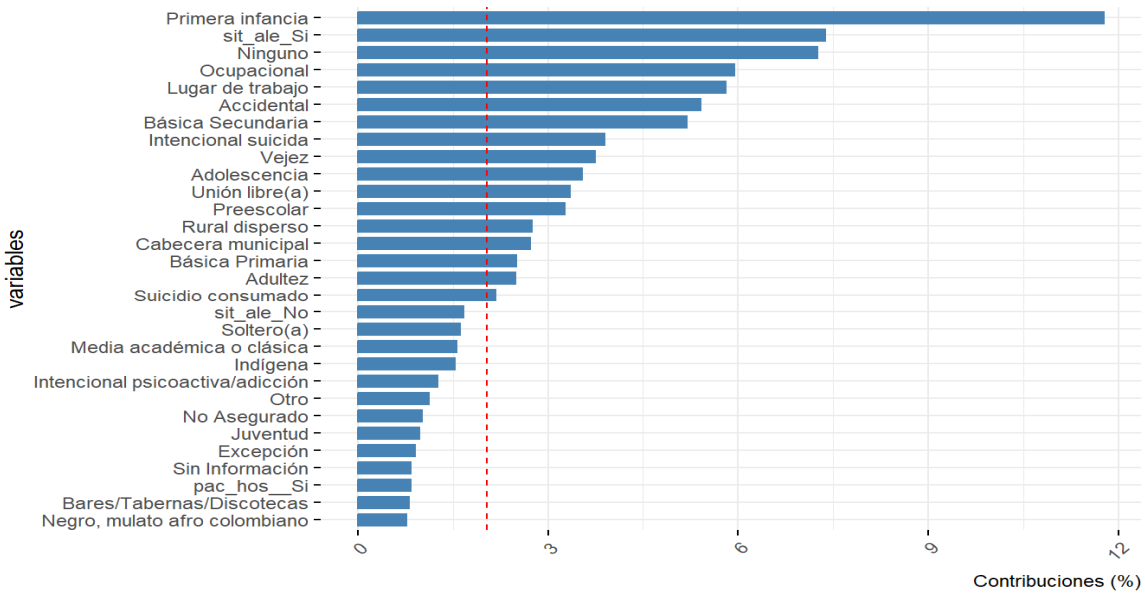


Figure 2. Contribution of the categories of definitive variables to the victim profile for event 365 - Intoxications in San Andrés de Tumaco.

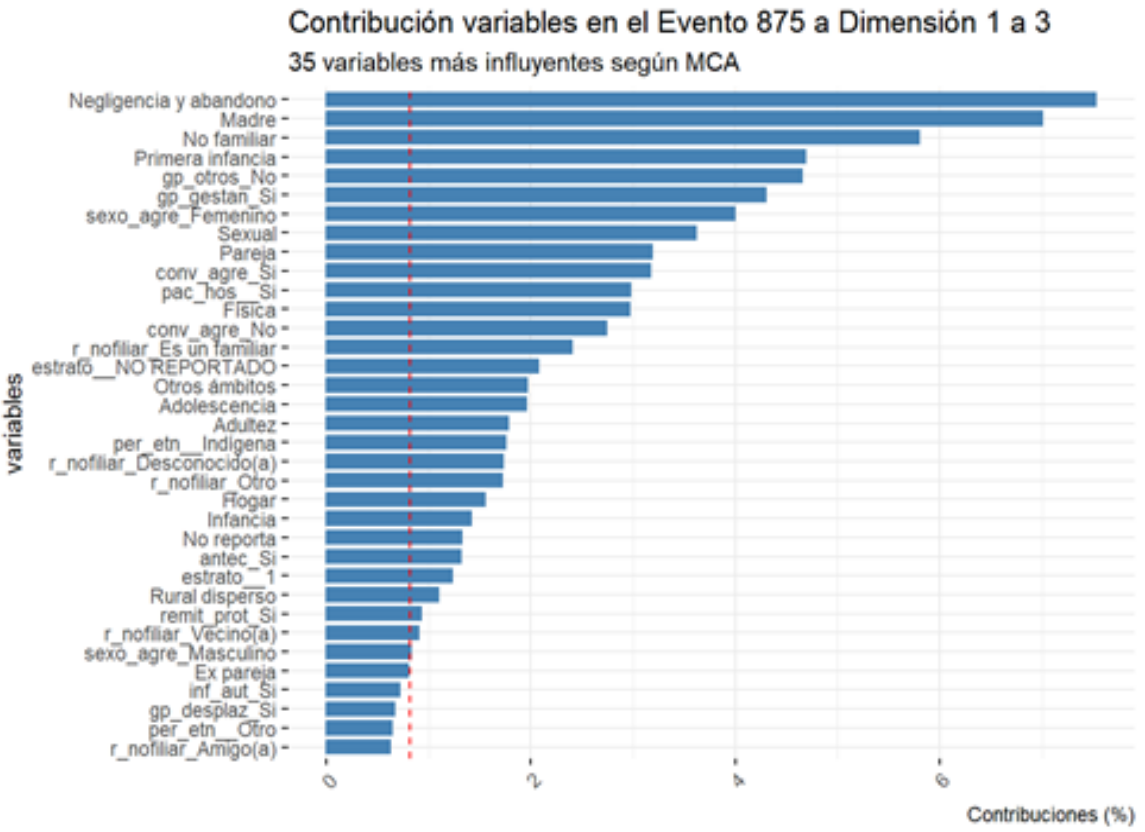


Figure 3. Contribution of the categories of definitive variables to the victim profile of event 875. Gender-based and domestic violence in San Andrés de Tumaco.

Similarly, MCA was conducted on the identified variables, and clustering was applied to group common characteristics among the women affected in each event. The MCA of the event data

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2024 doi:10.20944/preprints202410.0314.v1

revealed that factors such as the life cycle, area of occurrence, and health status of the victim are shared elements in cases of gender-based violence. These factors permitted the recognition of patterns of occurrence of violence in urban and dispersed rural areas. In addition, it was observed that suicide attempts are related to the life cycles of youth and adulthood, whereas gender and domestic violence are principally associated with childhood, especially in cases of neglect and abandonment, suggesting a possible relationship between violence experienced by mothers and neglect in the care of their daughters.

The cluster classification in Figure 4 revealed the connection between cases and the degree of vulnerability and normalization of violence. In particular, in the case of minors, violence is not normalized, thus reflected in a higher reporting rate, in contrast to what is observed in adult women.

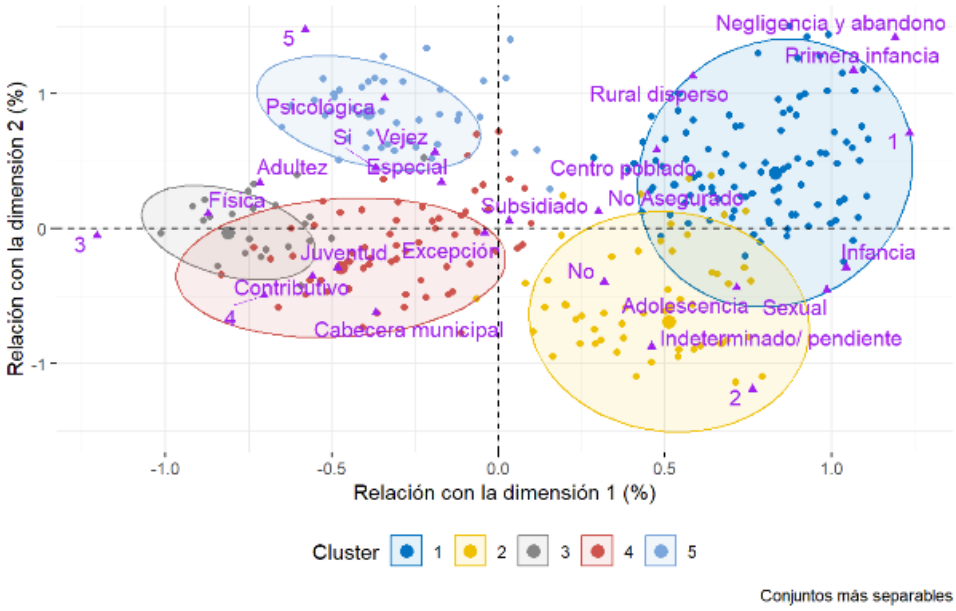


Figure 4. Scatter plot of the grouping provided by PAM clustering for event 875. Gender and domestic violence in San Andrés de Tumaco.

Furthermore, it was found that victims who live with their aggressor suffer principally from physical and psychological violence, neglect, neglect and abandonment, aspects that belong to the life cycles of early childhood, youth, and adulthood. In contrast, victims who do not live with their aggressor were predominantly found in the life cycles of childhood and adolescence and suffered from sexual violence.

4.1. Implementation of Decision Trees

Based on the selection of the variables that most contributed to characterizing the victim profile and the application of decision tree algorithms, the following results are obtained. As shown in the decision tree in Figure 5, the classification of women referred to social work services was obtained from training with variables such as alcohol_abuse, area, partner_prob, economic_prob, education, ss_type, and life_cycle, and aimed to predict the referral to social work for women with suicide attempts by analyzing several characteristics.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2024 doi:10.20944/preprints202410.0314.v1

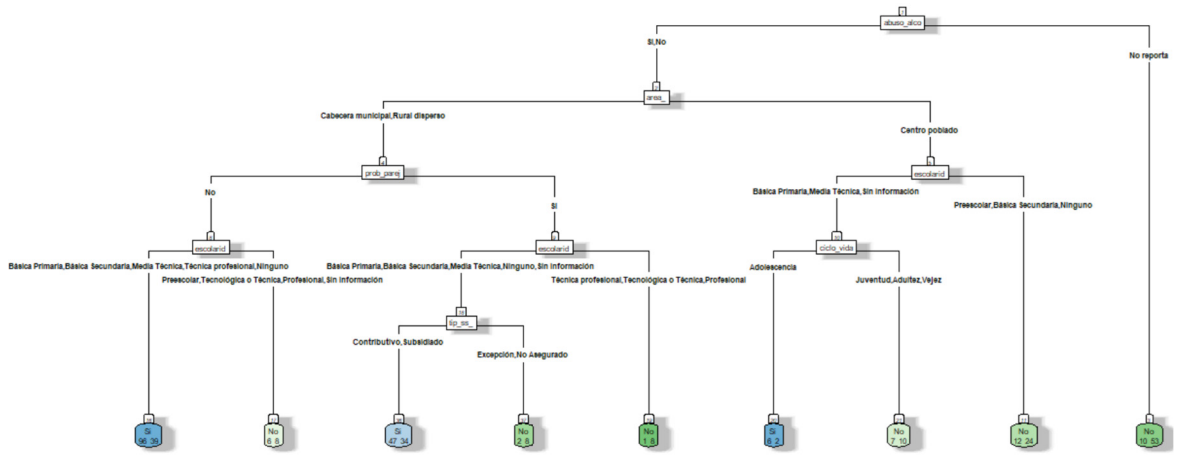


Figure 5. Decision tree focused on predicting women with suicide attempts referred to social work.

Alcohol abuse was reported by 57.10% of women, and 61.04% were from urban or sparsely populated rural areas. When there were no relationship problems, 68.46% of the cases were referred to social work, whereas 50% were referred to social work when relationship problems existed. Women with lower levels of education had a referral rate of 71.11%, compared to 42.86% of those with higher levels of education. Additionally, 15.87% of cases did not report alcohol abuse, and among these, 84.13% were non-referred to social work. These data underline that factors such as alcohol abuse, area of residence, partner issues, and educational level influence referral to social work.

The confusion matrix in Figure 6 demonstrates the ability of the decision tree to predict referral to social work for women with suicide attempts. The values reveal that of the 92 cases evaluated, the model correctly predicted 26 as non-referred and 12 as referred, with 20 false positives and 34 false negatives.

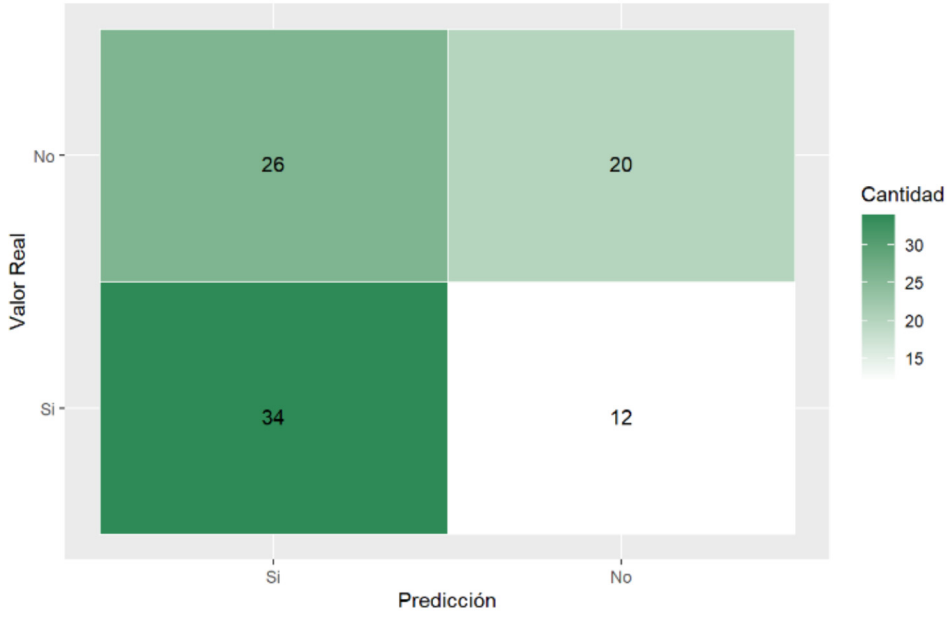


Figure 6. Confusion matrix of the trained decision tree for predicting social work referrals of women with suicide attempts.

The classification metrics in Table 2 indicate that the model's accuracy is 58.69%, sensitivity 56.66%, precision 73.91%, and F1 score 64.15%, suggesting that, although its accuracy is reasonable, its ability to identify all cases needing referral (sensitivity) adequately is limited.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2024 doi:10.20944/preprints202410.0314.v1

Table 2. Classification metrics for referral to social work according to the decision tree for women with suicide attempts.

F1.Score	Precision	Sensitivity	Accuracy
64,15%	73,91%	56,66%	58,69%

The results of the classification of women referred to psychiatric services through the decision tree trained with the variables life_cycle, area, marital_status, ethnicity, education, health_insurance_type, suicidal_ideation, and partner_prob, as shown in Figure 7, aim to predict the referral to psychiatry of women who had suicide attempts, using several descriptive variables. The data indicate that 68.42% of urban women and 82.61% of single or divorced women are referred to psychiatry, compared to 53.33% of those in cohabiting relationships. Regarding ethnicity, 64% of women from Indigenous communities and 41.46% of Afro-Colombian women are referred. Furthermore, 39.13% of women with basic education are referred. In the contributory health system, 60% are referred, while 39.06% are referred in the subsidized regime. When suicidal ideation is present, there is a referral rate of 53.33%, a figure that decreases to 34.69% when such ideations are absent. Adolescent women present a referral rate of 52.38%, whereas the result for women in other stages of life is 21.43%.

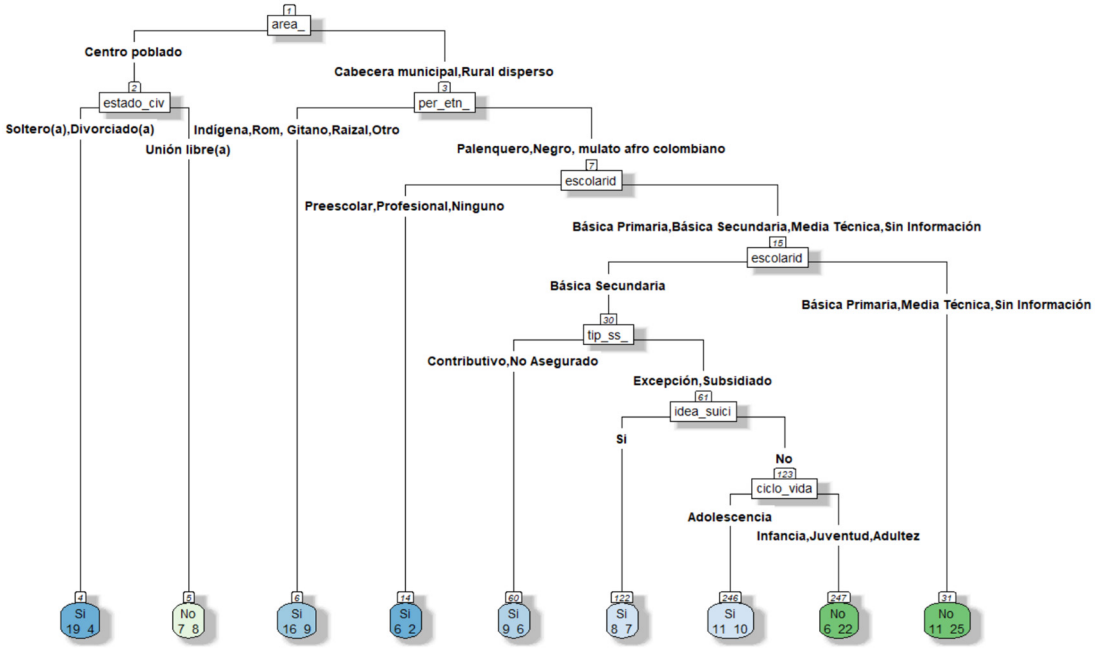


Figure 7. Decision tree focused on predicting women with suicidal attempts referred to psychiatry.

The confusion matrix in Figure 8 and the decision tree classification metrics in Table 3 predicting referrals to psychiatry show an accuracy of 65.21%, with a sensitivity of 62.96% and a precision of 73.91%, resulting in an F1 score of 68%. The matrix revealed that 10 cases of referral (true positives) and six cases of non-referred (true negatives) were properly predicted, although there were also 17 false negatives and 13 false positives. These results indicate a moderate ability of the model to predict psychiatric referrals, with a good balance between precision and sensitivity, despite the significant number of false negatives presented.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2024 doi:10.20944/preprints202410.0314.v1

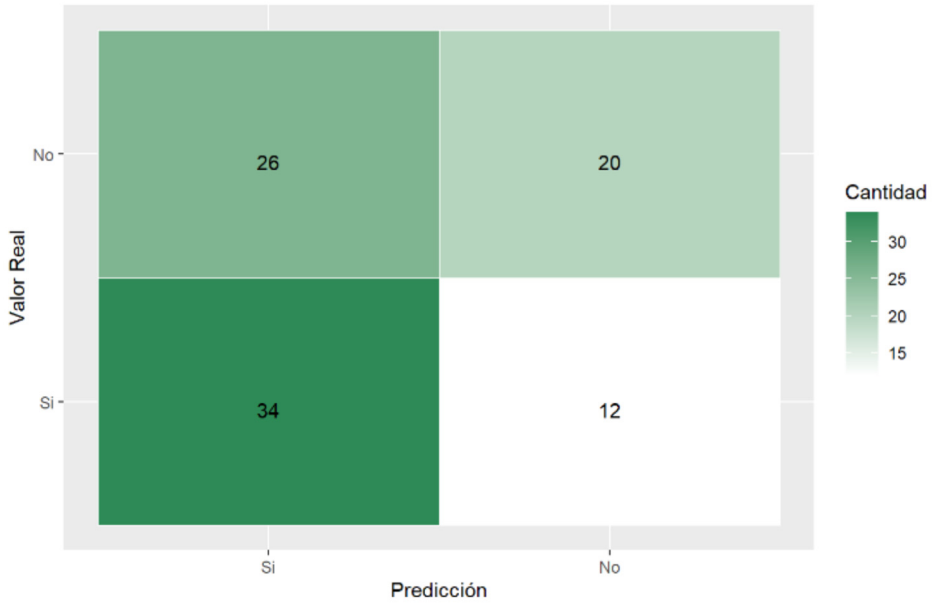


Figure 8. Confusion matrix of the trained decision tree for predicting psychiatric referral of women with suicide attempts.

Table 3. Classification metrics for psychiatric referral using a decision tree in women with suicide attempts.

F1.Score	Precision	Sensitivity	Accuracy
68%	73,91%	62,96%	65,21%

For the analysis of event 365, poisonings, the tree in Figure 9 was elaborated, capturing the characteristics of poisonings of women in the "alert situations". The root node indicates 81.51% of cases classified as "no" and 18.49% as "yes". The variable 'life cycle' divides the sample into: early childhood and childhood, adolescence, youth, adulthood, etc. The majority of accidental poisonings are classified as "yes". The type of health regime indicates that most cases occur in the subsidized or uninsured regime, where there is a higher proportion of cases classified as "yes". In contrast, the branch infancy, adolescence, youth, adulthood, old age shows that most cases are classified as "no". In summary, most acute poisonings that occur in early childhood are accidental, and factors such as type of exposure and health regime influence them.

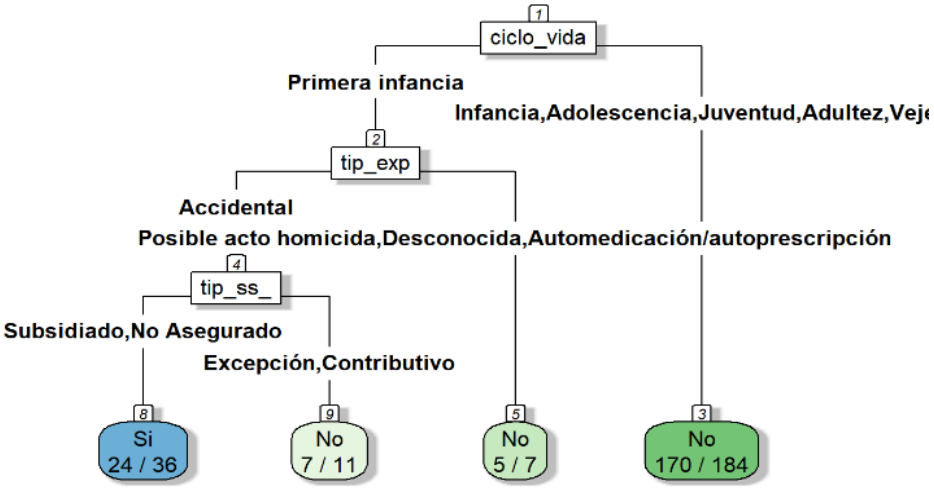


Figure 9. Decision tree focusing on predicting women with referred poisonings in alert situation.

 doi:10.20944/preprints202410.0314.v1

The confusion matrix in Figure 10 and the decision tree model metrics in Table 4 show that the accuracy is 79.66%, indicating proper performance in predicting alert situations. However, the sensitivity (46.15%) reveals difficulties when correctly identifying all alert situations; the accuracy of 54.54% points to a high false alarm rate. The F1 score of 50% reflects a moderate balance between the function to detect alert situations and avoid false alarms, suggesting possible improvements in model sensitivity and accuracy.

Table 4. Classification metrics for psychiatric referral using a decision tree in women with poisoning.

F1.Score	Precision	Sensitivity	Accuracy
50%	54,54%	46,15%	79,66%

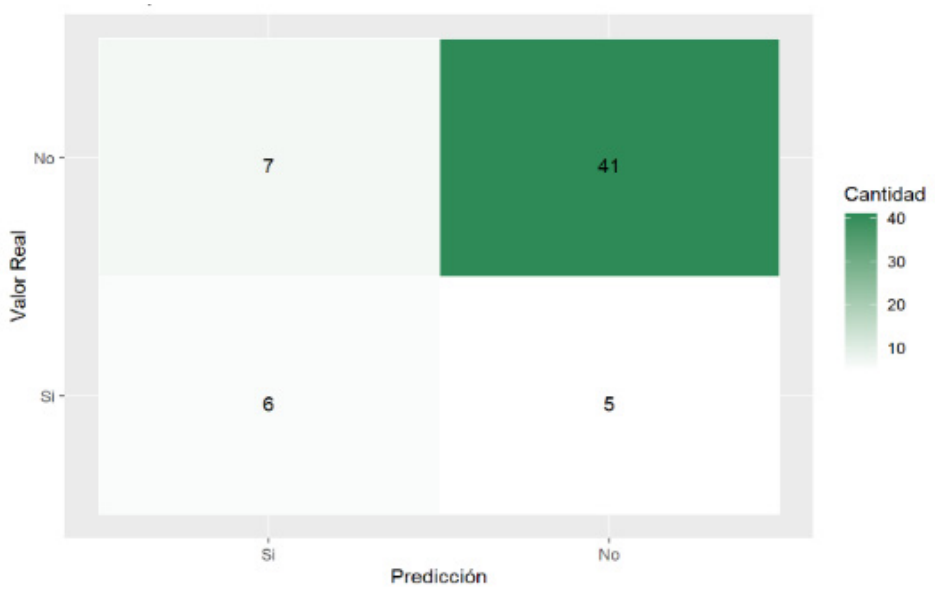


Figure 10. Confusion matrix of the trained decision tree for predicting alert situations in poisonings.

For the analysis of event 875, gender violence, the tree in Figure 11 was elaborated with the characteristics that classify the nature of the situations of physical, psychological, neglect and abandonment, and sexual alert, using variables such as life cycle, ethnicity, type of social service, and occurrence area. Its configuration was based on categories such as life_cycle, per_ethn_, type_ss_, area, and nature, determining that the majority of physical mistreatment occurs in Afro-Colombian black or Mulatto youths and adults, and sexual abuse is more frequent in adolescents and older persons of the same group. Indigenous people and other groups in dispersed rural areas have a higher incidence of neglect and abandonment, and Indigenous people in infancy and early childhood suffer more psychological mistreatment.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2024 doi:10.20944/preprints202410.0314.v1

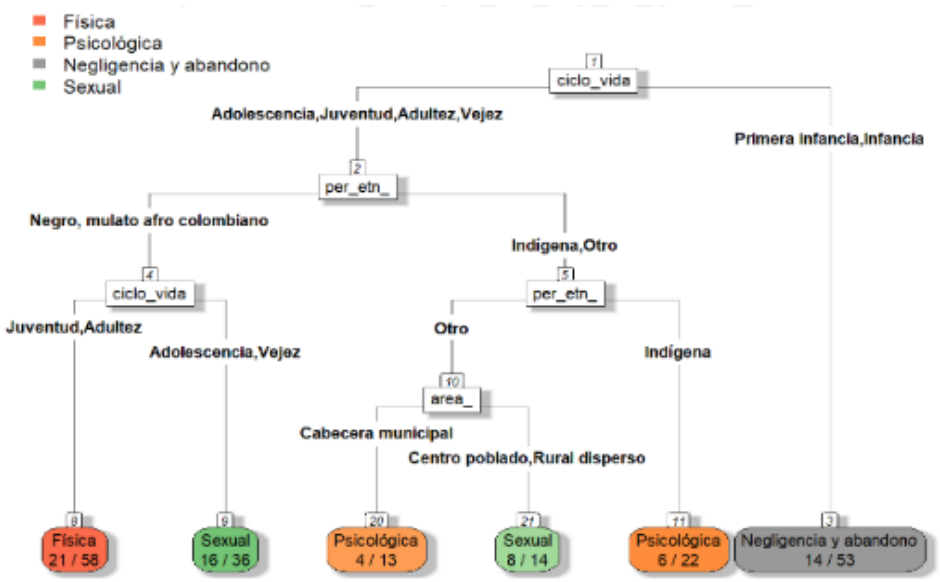


Figure 11. Decision tree focused on predicting the nature/type of violence suffered by women in gender-based and intrafamily violence.

The confusion matrix of the decision tree model in Figure 12 and the classification metrics in Table 5 reflect its correct performance in predicting cases of neglect and abandonment, with an accuracy of 83.33% and a sensitivity of 83.33%. However, its ability to correctly predict physical, psychological, and sexual violence is limited, especially in the case of sexual violence, with a sensitivity of 25% and an accuracy of 33.33%.

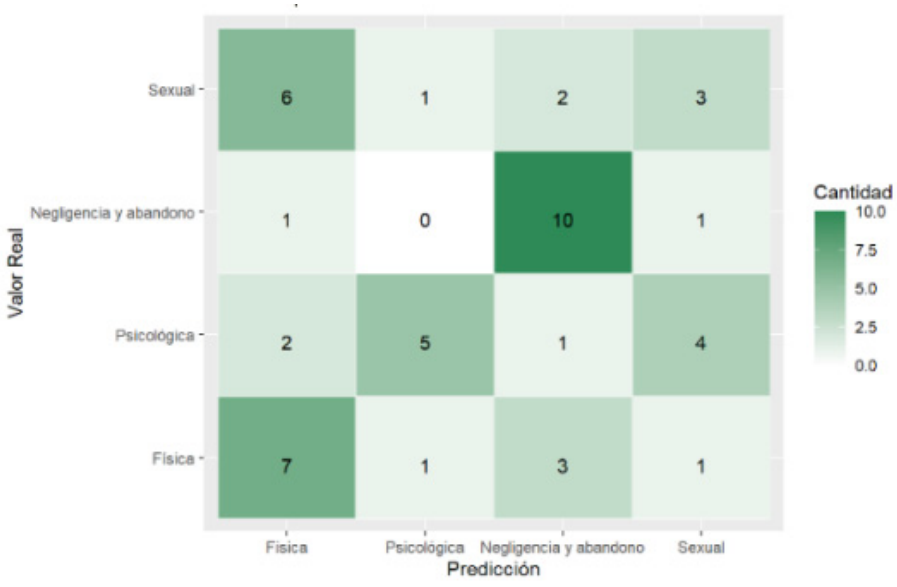


Figure 12. Confusion matrix of the decision tree trained to predict the type of violence suffered by women.

Table 5. Metrics for classifying violence type using a decision tree.

F1.Score	Precision	Sensitivity	Accuracy	Class
50%	43,75%	58,33%	66,66%	Physical
52,63%	71,42%	41,66%	68,05%	Psychological

 doi:10.20944/preprints202410.0314.v1

71,42%	62,50%	83,33%	83,33%	Neglect and abandonment
28,57%	33,33%	25%	54,16%	Sexual

Even though the model is relatively effective for some categories, it needs considerable improvement to identify all types of violence.

4.2. Implementation of Random Forests

A random forest model was trained using data from event 356 to improve accuracy in identifying relevant characteristics for the referral of women to social work. By recognizing the importance of each variable in the event, it was found that the most significant were education level, socioeconomic status, area of occurrence, life cycle, type of health service, place of intoxication, alcohol abuse, ethnic affiliation, marital status, relationship problems, economic issues, gestational week, and suicidal thought.

Given the problem of unreported data in the variable stratum, this was excluded from the training. The use of the other typologies was scaled in a new algorithm, focusing the analysis on the characteristics of the individual. A cross-validation was performed (Figure 13), and the best behavior of the random forests was identified only with the variables that did not generate bias or confusion: schooling, area, life cycle, type of social security, alcohol abuse, ethnicity, marital status, relationship problems, economic problems, and suicidal thoughts.

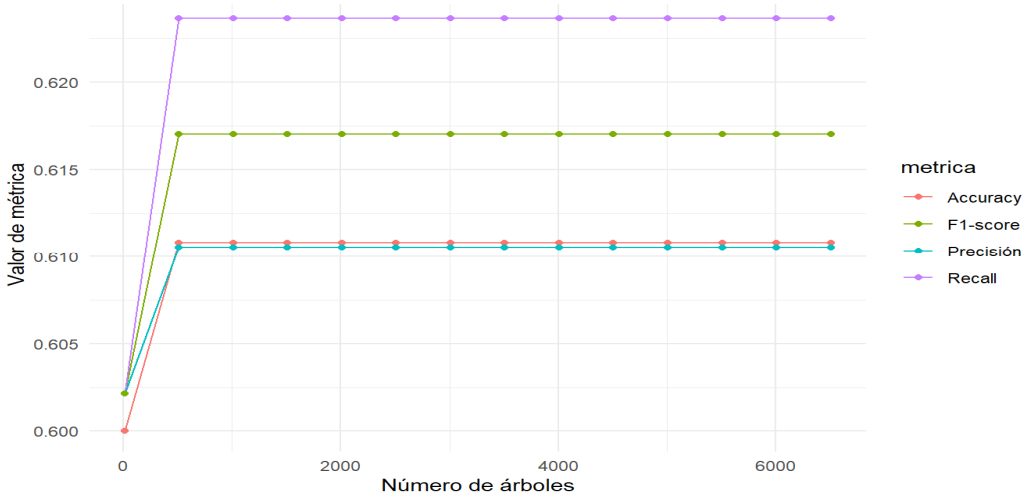


Figure 13. Cross-validation applied to random forests in the prediction of referrals of battered women to social work services.

In Figure 14 and Table 6, the random forest model trained with the principal variables shows a moderate performance, with metrics of around 60%. It has an accuracy of 61.1%, a precision of 61.1%, a recall of 62.4%, and an F1 score of 61.7%, indicating a moderate ability for effective identification and referral to social work. The selection of these variables has improved the model's accuracy and reflects adequate identification of the relevant features.

 doi:10.20944/preprints202410.0314.v1

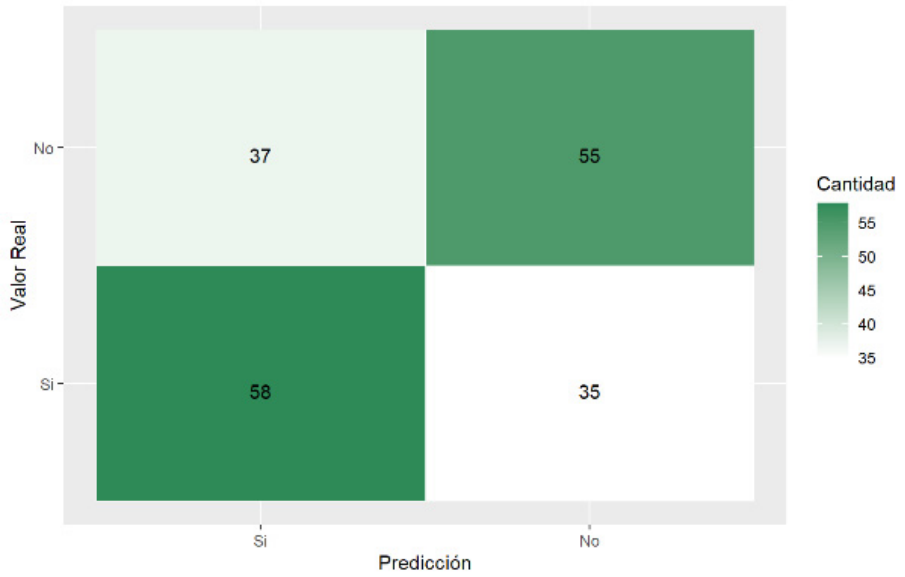


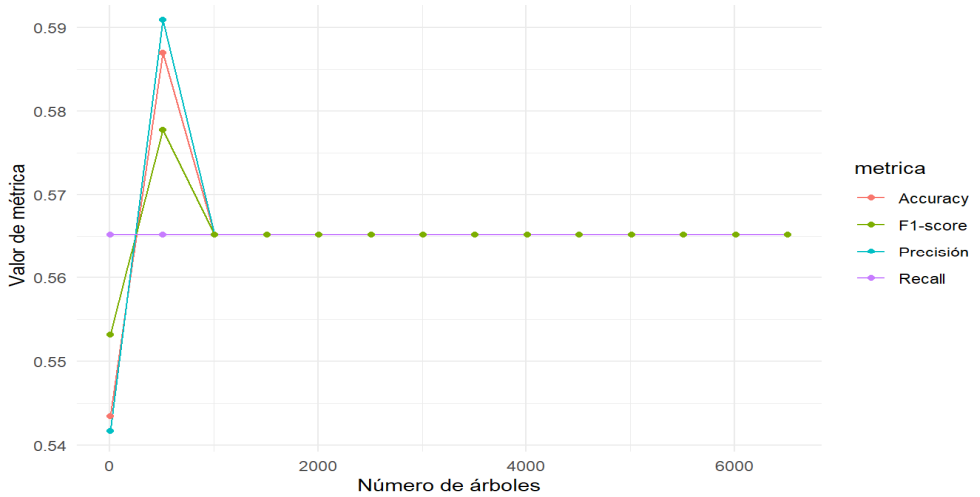
Figure 14. Confusion matrix random forest for predicting referrals to social work of women with suicide attempts.

In Figure 14 and Table 6, the Random Forest model trained with the key variables shows moderate performance, with metrics around 60%. It has an accuracy of 61.1%, a precision of 61.1%, a recall of 62.4%, and an F1-score of 61.7%, indicating a moderate capacity for effective identification and referral to social work. The selection of these variables has improved the model's precision and reflects an effective identification of relevant characteristics.

Table 6. Classification metrics for referral to social work in suicide attempts according to random forest.

F1.Score	Sensitivity	Precision	Accuracy
61,70%	62,36%	61,05%	61,08%

Also, to characterize the profile of women with suicide attempts referred to psychiatry, a random forest model was trained with complementary data variables, identifying the most important: schooling, stratum, area, life cycle, type of social security, place of attempt, alcohol abuse, ethnicity, marital status, marital problems, week of gestation, and suicidal ideation. Additional tests with these variables showed that life cycle, area, schooling, type of social security, suicidal ideation, ethnicity, marital status, and marital problems were the most relevant and did not generate bias. Cross-validation with different numbers of trees in the model confirmed these results, as shown in Figure 15.



 doi:10.20944/preprints202410.0314.v1

Figure 15. Cross-validation applied to random forests in predicting referrals of battered women to psychiatric services.

For predicting psychiatric referral, the random forest model to predict obtained an accuracy of 58.7%, a precision of 59.1%, a recall of 56.5%, and an F1 score of 57.8% (Table 7).

Table 7. Random forest classification metrics for psychiatry referral in suicide attempts.

F1.Score	Sensitivity	Precision	Accuracy
57,77%	56,52%	59,09%	58,69%

The confusion matrix in Figure 16 shows nine true negative cases, 14 false positives, 13 false negatives, and 10 true positives, indicating moderate algorithm performance.

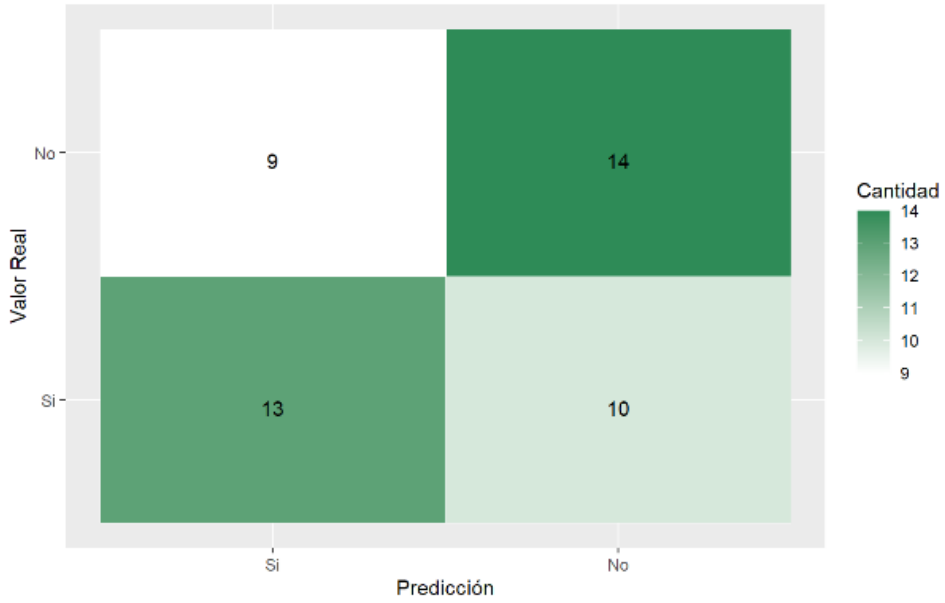


Figure 16. Random forest confusion matrix trained for predicting psychiatric referrals in women with suicide attempts.

With the implementation of the random forest model to improve the decision-making capacity in alert cases and review additional variables of event 365, the observations were balanced for improved prediction. It was observed that the life cycle significantly influences poisoning cases, and marital status emerges as a possible principal factor, indicating the relevance of the support network in situations of violence that can be life-threatening.

Figure 17 includes the confusion matrix of the random forest model's ability to predict alert situations in poisoning. In this case, it provides one true positive (TP), one true negative (TN), 10 false positives (FP), and 10 false negatives (FN).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2024 doi:10.20944/preprints202410.0314.v1

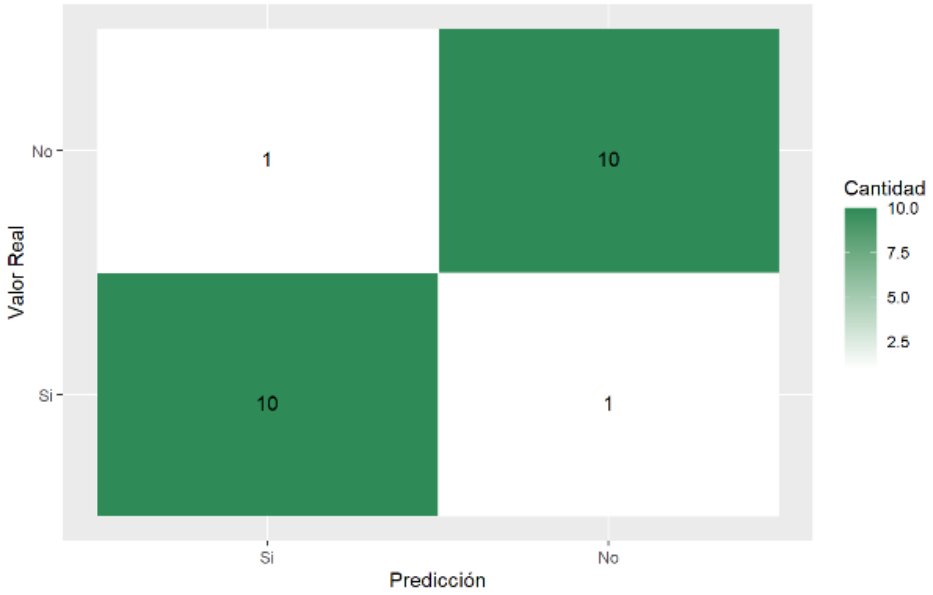


Figure 17. Confusion matrix random forest trained to predict alertness in poisoning.

Despite these values, the performance metrics (Table 8) are high: accuracy is 0.909; precision, 0.909; recall, 0.909, and F1 score is also 0.909, suggesting that the model exhibits high overall accuracy and efficiency even when the confusion matrix indicates that the model faces difficulties in correctly distinguishing between the "yes" and "no" classes in specific situations. These metrics could be due to an unbalanced dataset or the model benefiting from invisible factors in the confusion matrix.

Table 8. Classification metrics for the alert situation in random forest poisonings.

F1.Score	Sensitivity	Precision	Accuracy
90,90%	90,90%	90,90%	90,90%

It may be related to observing more frequent alerts reported as early childhood accidents. Exposures that intentionally seek to harm life were also identified. However, insufficient observations or characteristic variables do not generate an effective context-sensitive value classifier.

To improve the decision-making capacity on the types of violence, cross-validation of different random forest models was performed with event 875 (Figure 18). The relationship between the aggressor and the victim and the socioeconomic aspect influenced the type of violence.

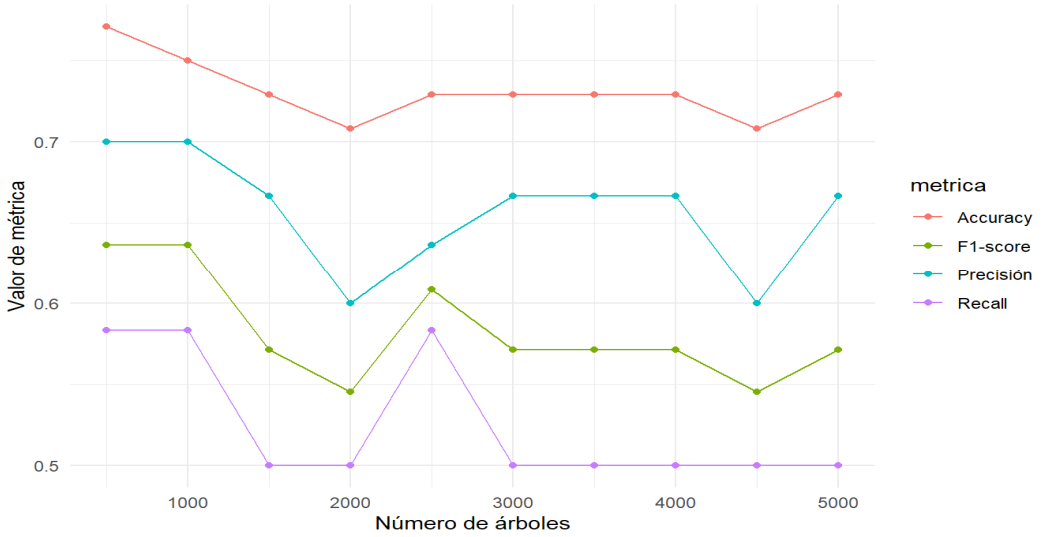


Figure 18. Cross-validation applied to random forests in predicting the nature/type of violence suffered by women in gender and domestic violence.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2024 doi:10.20944/preprints202410.0314.v1

After training the best random forest model, the most influential variables were identified: r_fam_vic, cycle_life, sex_agre, per_ethn_, r_nofiliar, pac_hos_, conv_agre, ambito_lug, area_, tip_ss_, and estrato_. However, not all of these variables are unique to the victim. Therefore, a model was trained with the variables: cycle_life, per_ethn_, area_, conv_agre, and tip_ss_ to improve the characterization of the potential victims.

The confusion matrix in Figure 19 and the classification metrics in Table 9 indicate that the random forest model trained under cycle_life, per_ethn_, area_, conv_agre, and tip_ss_ has a sound predictive capacity to classify violence. The accuracy, sensitivity, precision, and F1 score metrics indicate acceptable performance, standing out in predicting physical and psychological violence. However, the model presents accuracy and sensitivity problems for the categories neglect and abandonment and sexual, suggesting improvements in future model iterations.

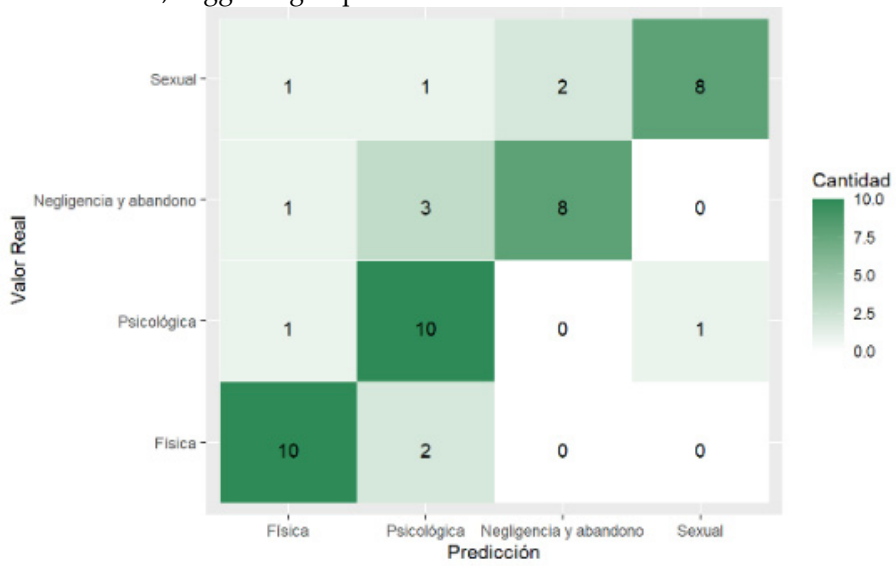


Figure 19. Confusion matrix. Random forest trained to predict the nature/type of violence suffered by women in gender and domestic violence.

Table 9. Metrics for classifying the type of violence using a decision tree.

F1.Score	Precision	Sensitivity	Accuracy	Class
50%	43,75%	58,33%	66,66%	Physical
52,63%	71,42%	41,66%	68,05%	Psychological
71,42%	62,50%	83,33%	83,33%	Neglect and abandonment
28,57%	33,33%	25%	54,16%	Sexual

4.3. Artificial Neural Networks (ANN)

The neural network designed to characterize psychiatric referrals in cases of attempted suicide included four hidden layers with 500, 100, six, and two neurons, and a convergence threshold of 0.0, with a maximum step size of 1e+14. The training was repeated three times, using the "rprop+" algorithm and the logistic activation function. Although the network showed better metrics after several adjustments to its hyperparameters, it did not achieve high predictive capacity. This model took five times longer to train due to the small number of observations and the need to balance the data.

The confusion matrix in Figure 20 for the trained neural network revealed poor performance in predicting referrals to social work in reports of suicide attempts, as Table 10 shows. The high mean prediction error rate, 62.86%, and entropy problems indicated that the model is ineffective in correctly distinguishing between cases that are and are not referred to social work.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 October 2024 doi:10.20944/preprints202410.0314.v1

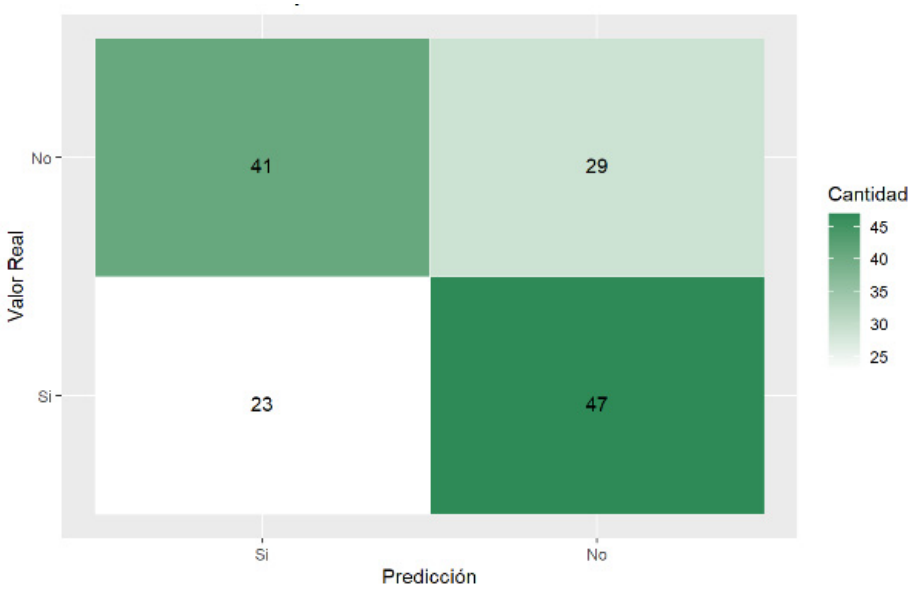


Figure 20. ANN confusion matrix trained for predicting social work referral.

Table 10. ANN classification metrics for social work referral.

F1.Score	Sensitivity	Precision	Accuracy
34,32%	32,85%	35,93%	37,14%

For women who attempted suicide and were referred to psychiatric services, there is a four-layer hidden neural network with the same configuration as the social work referral network. Although this model demonstrated improved metrics, it did not achieved predictive capacity, and the same entropy and prediction error problems as the previous network were generated, becoming noticeable in the confusion matrix in Figure 21 and the metrics in Table 11. Also, the training time was five times longer due to the need to balance the collected data.

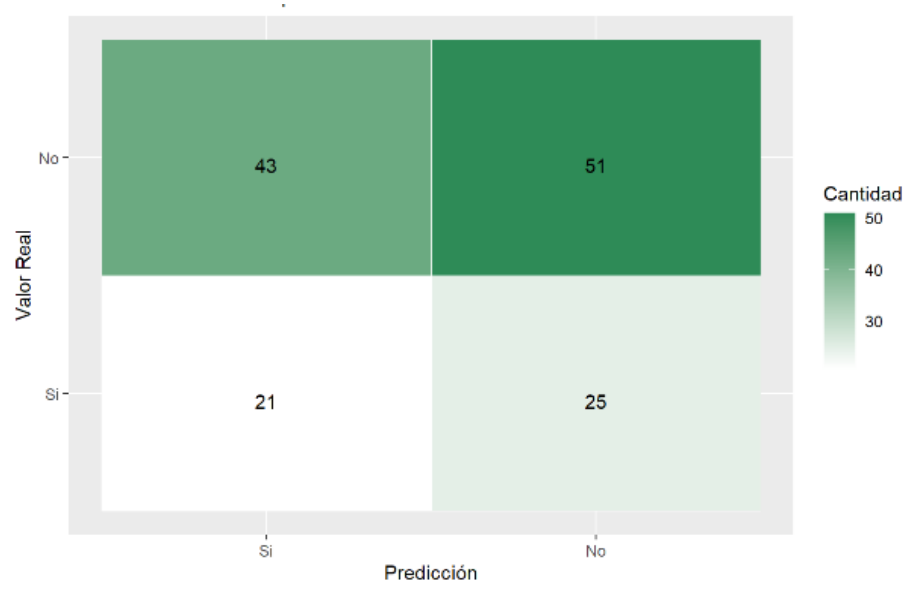


Figure 21. ANN confusion matrix trained for predicting referral to a psychiatric service.

Table 11. ANN classification metrics for psychiatric referral.

F1.Score	Sensitivity	Precision	Accuracy
38,18%	45,65%	32,81%	51,43%

 doi:10.20944/preprints202410.0314.v1

The confusion matrix in Figure 21 for the neural network trained to predict referral to psychiatry showed limited performance, with an accuracy of 51.43%. The neural network achieved 43 true negatives and 25 true positives but also had 51 false positives and 21 false negatives. With an accuracy of 32.81%, a recall of 45.65%, and an F1 score of 38.18%, it is clear that the model suffers from poor predictive capacity. In addition, the mean entropy was not calculable and the mean prediction error was 48.57%, indicating that the characterization of the referral phenomenon to areas such as psychiatry or social work presents a high level of complexity.

For the analysis of gender and domestic violence with ANNs, the predictive capacity of the model was evaluated without analyzing the variables involved in decision-making. The neural network with a hidden layer structure (500, 100, 10, 4) and specific parameters were used to optimize the predictive performance.

The confusion matrix in Figure 22 revealed that the neural network model contained a high mean prediction error (1.970297) and an undefined mean entropy (NaN), indicating poor performance. Considering the information in Table 12, the classification yielded better results in cases of physical and sexual violence, with an accuracy of 75.69% and 74.56%, and a sensitivity of 80.63% and 69.61%, respectively. However, its performance was poor concerning psychological violence, as its accuracy stands at only 3.77%, its sensitivity at 12.50%, and it did not identify cases of neglect and abandonment, reflecting the lack of correct predictions for this category.

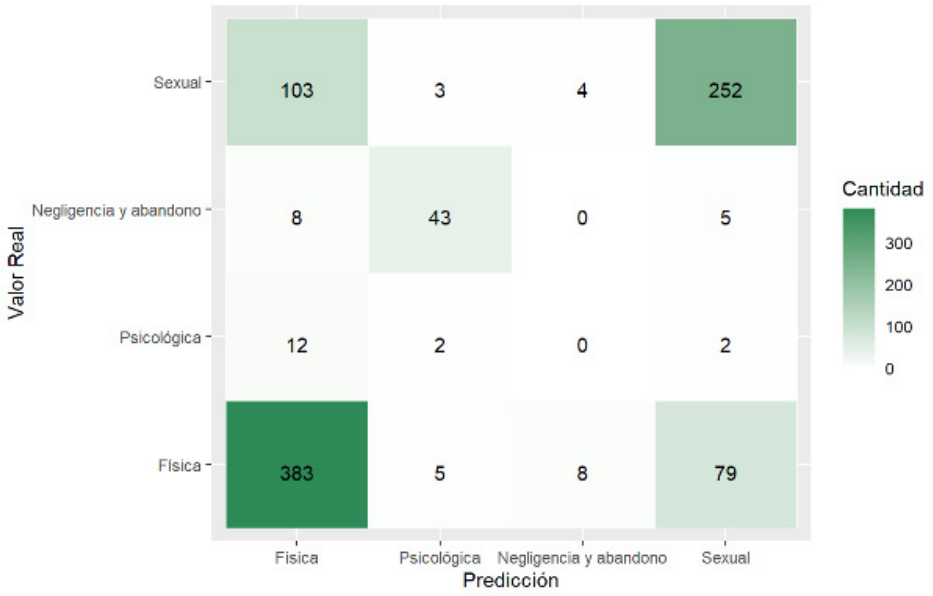


Figure 22. ANN confusion matrix trained to predict the nature/type of violence suffered by women in gender and domestic violence.

Table 12. Metrics for classifying violence type using an ANN.

F1.Score	Precision	Sensitivity	Accuracy	Class
78,08%	75,69%	80,63%	76,14%	Physical
5,79%	3,77%	12,50%	53,39%	Psychological
NaN%	0%	0%	49,29%	Neglect and abandonment
72,00%	74,55%	69,61%	76,94%	Sexual

Finally, the projection design based on the rates of gender and domestic violence provided by the statistical analysis of the SIVIGILA 875 event is presented concerning the results of implementing clustering and random forest algorithms. Table 13 shows the violence rates in the 875 event reports.

Based on the population reports from the National Administrative Department of Statistics (DANE), from 2024 to 2028, hypothetical profiles of potential victims were created through a detailed statistical analysis of the dataset needed to project violence in San Andrés de Tumaco. A distribution

 doi:10.20944/preprints202410.0314.v1

was established based on the observation rates of event 875, applying the mean to obtain a portrait of the trend of violence cases. The observation rate, calculated with the number of women per 1000 inhabitants, allowed a wider representativeness of the dataset and a richer interpretation of the frequencies, in a sample of 1000 women as the analysis unit (Table 13).

Table 13. Rates of gender-based violence in San Andrés de Tumaco from 2018 to 2023.

Rate of women who experienced violence per 1000 women	Rate of violence cases against women per 1000 women	Women who experienced gender-based and domestic violence	Cases of gender-based and domestic violence against women	Total female population	Año
2,73	2,84	355	366	128.752	2018
2,87	3.08	374	401	129.923	2019
2,23	2,45	296	324	132.176	2020
2.71	2,93	363	392	133.515	2021
2,78	3,02	374	406	134.277	2022
3,45	3,81	467	515	135.117	2023

These victim profiles were used to make specific predictions, applying clustering techniques according to the initial exploratory analysis.

The results showed that the random forest and clustering algorithms were effective in the projections. Combining these methods made it possible to accurately classify the different types of violence and identify relevant patterns in the data. The projections obtained with these models provided a comprehensive and detailed view of gender violence in San Andrés de Tumaco, facilitating the identification of critical areas for future interventions.

The projections and the visualization of all the database figures analyzed throughout this research were recorded in the decision-making tool shown in Figures 23 and 24.

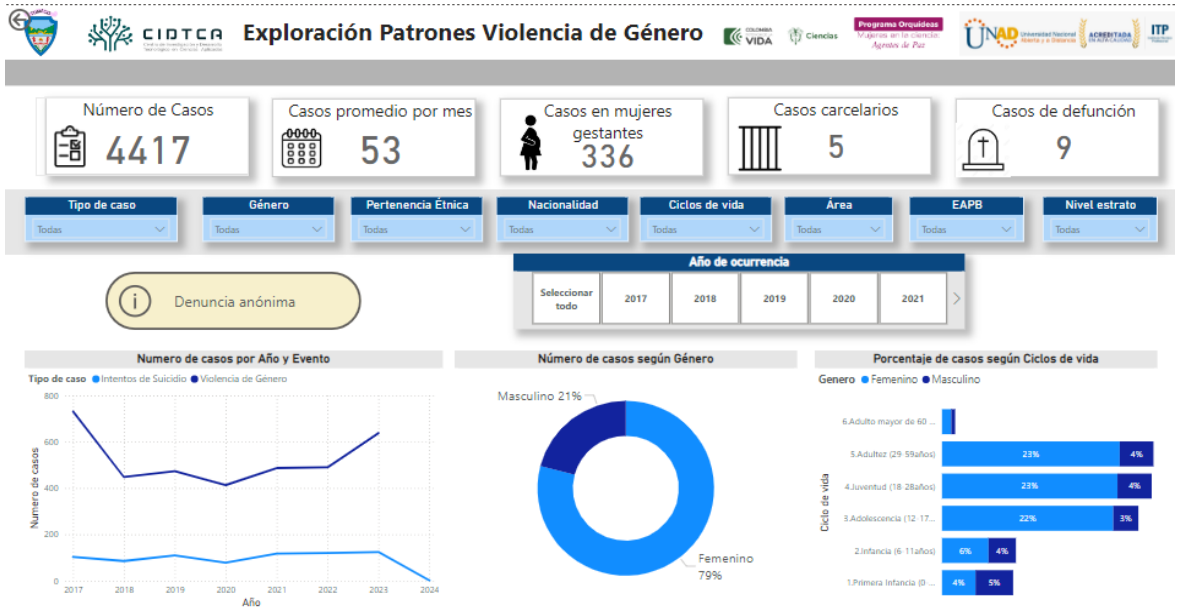


Figure 23. Interface of the section presenting gender violence figures in San Andrés de Tumaco and the anonymous reporting button from the decision-making tool developed in Power BI.

 doi:10.20944/preprints202410.0314.v1

Figure 24. Interface of the section presenting projections of the gender-based violence rate in San Andrés de Tumaco for 2024 and 2028 from the decision-making tool developed in Power BI.

5. Discussion

The analysis of the SIVIGILA system has unveiled notable deficiencies in the completion of the reports, thus affecting the quality and accuracy of the data collected. The lack of automation in entering information by health institution officials contributes to common errors such as duplication of information in incorrect fields and outdated records in the reporting forms. In addition, although the mayor's office showed interest in detailing the results by neighborhoods and ethnic groups in the district, it was found that the location in the databases does not comply with uniform standards. This inconsistency made it difficult to filter and explore the information, preventing an accurate and detailed local and sectoral analysis, albeit some vulnerabilities were characterized by area of occurrence. These observations underscore the immediate need to train the personnel responsible for filling out the event notification forms to mitigate erroneous data entry and improve the quality of the information reported in the SIVIGILA. Despite the application of artificial neural networks, the data collected failed to identify representative patterns, limiting the ability of these techniques to provide accurate predictions.

In contrast, the random forest algorithm proved to be more effective if combined with subsampling techniques, as a notable improvement in the accuracy and robustness of the results was observed. In addition, the use of clustering techniques revealed patterns of normalization of gender violence, especially about the vulnerability of the people involved in the reported cases. The evaluation of the effectiveness of the models was conducted using performance metrics such as accuracy, sensitivity and specificity, and the analysis of confusion matrices. The analysis of the importance of the variables, complemented with techniques such as MCA, led to the identification of the most influential factors in the predictions, improving the interpretation of the results and providing a more complete and detailed view of the data.

This study illustrates that the combined use of ML and clustering techniques can provide a detailed and accurate understanding of patterns in datasets, especially in gender violence. The validation techniques allowed a margin close to 70% in the prediction capacity, underlining the methodology's effectiveness.

The analysis reveals several critical areas that require intervention to address gender violence in San Andrés de Tumaco. The findings suggest the need to implement specific strategies to improve the situation and support affected women.

First, educational interventions are essential. The proposal includes developing programs from childhood that promote gender equality, mutual respect, and peaceful conflict resolution. These programs should be integrated into the school curriculum and complemented by workshops and

 doi:10.20944/preprints202410.0314.v1

campaigns in the community. Teacher training and community sensitization are essential to create meaningful and lasting cultural change.

Regarding access to mental health services, it is crucial to improve the availability and accessibility of therapy and psychological support for women victims of gender violence. Health centers and community-based organizations should offer these services, ensuring cultural sensitivity and accessibility. Adequate care will enable women to recover emotionally and healthily and rebuild their lives.

Financial support is also vital. Programs that include job training, access to microcredit, employment programs, and financial assistance are recommended to help women achieve economic independence. This support can reduce their dependence on abusers and facilitate their ability to escape violent situations.

Strengthening community support networks is another crucial aspect. Clear protocols should be established for responding to gender-based violence, training community leaders, and creating safe spaces for victims. Active community participation in preventing and responding to violence contributes to have a supportive and protective environment for affected women.

In addition, alcohol abuse awareness and prevention are essential to address a leading risk factor in gender-based violence. Educational campaigns, regulations on the sale of alcohol, and treatment programs for people with abuse problems are proposed. Addressing excessive alcohol consumption can reduce the risk of violence and promote a safer environment.

Finally, the promotion of legal rights and resources is essential. Women require information about their legal rights and the resources available for protection. Training in legal rights, establishing legal advice centers, and promoting these services, free or low-cost, will ensure that women can access justice and protect themselves from violence.

These proposals are designed to address gender-based violence from multiple angles to create a comprehensive and sustainable approach that benefits women in San Andrés de Tumaco.

6. Conclusions

The implementation of ML algorithms for the identification of violence patterns is presented as a promising alternative in the study of the needs of vulnerable women. These algorithms permit a deeper understanding of the context where violence emerges and provide a robust tool for planning and consolidating intervention strategies. Government entities must prioritize collecting orderly and accurate information to maximize the impact of these technologies. A well-structured and reliable database supports the implementation of ML algorithms and the continuous adaptation of decision-making tools, improving the effectiveness of strategies designed to combat violence and assist affected women.

Recognition of women's needs and the non-normalization of violence by the community are crucial factors for the effectiveness of decision-making tools, as has been raised in this research. Women must understand the consequences of violence and the alternatives available both in their support network and in government entities. This understanding facilitates information collection effectively, improving the quality of statistical analysis and the formulation of intervention strategies. In addition, involving men and minors in recognizing the problem is critical to a generational transformation, creating a solid foundation for combating violence and fostering lasting cultural change.

The identification of vulnerability patterns in women in San Andrés de Tumaco revealed the prevalence of various types of violence throughout the different life cycles of victims, demonstrating that women face risks continuously in their lives. A particularly alarming finding is the high prevalence of sexual violence among minors, indicating that this population is quite vulnerable. In addition, the analysis shows that, in youth, changes in the social context and romantic relationships increase the incidence of physical and psychological violence. These results underline the need for specialized approaches adapted to each stage of life to comprehensively and effectively address violence against women, from prevention to intervention and support. In this regard, the implementation of intervention strategies focused on the different life cycles of women in San Andrés

 doi:10.20944/preprints202410.0314.v1

de Tumaco is essential to mitigate violence. If specific approaches are developed for each stage of life, the violence women face in different circumstances can be addressed more effectively. This approach not only seeks to reduce the incidence of violence but also to foster a culture of prevention and support that breaks the cycle of abuse and prevents its transmission from generation to generation. These strategies have to be adapted to the specific needs of each age group, thus ensuring a positive and lasting impact on the community.

Appendix A

Table A1. Dictionary of variables of the basic data of SIVIGILA events 356, 365, and 875.

N°	Variable Name	Meaning in the Event	N°	Variable Name	Meaning in the Event
1	cod_eve	Código del evento	43	gp_psiquia	Centros psiquiátricos
2	fec_not	Fecha de la notificación	44	gp_vic_vio	Víctimas de violencia armada
3	semana	Semana de la notificación	45	gp_otros	Otros grupos poblacionales
4	año	Año de notificación	46	fuente	Fuente notificación
5	cod_pre	Códigos prestadores de salud	47	cod_pais_r	Código país de residencia
6	cod_sub	Código UPGD que recibe el caso	48	cod_dpto_r	Código departamento de residencia
7	pri_nom_	Primer nombre	49	cod_mun_r	Código municipio de residencia
8	seg_nom_	Segundo nombre	50	fec_con_	Fecha de consulta (dd/mm/aaaa)
9	pri_ape_	Primer apellido	51	ini_sin_	Fecha de inicio de síntomas (dd/mm/aaaa)
10	seg_ape_	Segundo apellido	52	tip_cas_	Clasificación inicial de caso
11	tip_ide_	Tipo de identificación	53	pac_hos_	Paciente Hospitalizado
12	num_ide_	Número de identificación	54	fec_hos_	Fecha de hospitalización (dd/mm/aaaa)
13	edad_	Edad	55	con_fin_	Condición final
14	uni_med_	Unidad de medida de la edad	56	fec_def_	Fecha de defunción (dd/mm/aaaa)
15	nacionali_	Código nacionalidad	57	ajuste_	Desconocido
16	nombre_ nacionalidad	Nombre nacionalidad	58	telefono_	Teléfono
17	sexo_	Sexo	59	fecha_nto_	Desconocido
18	cod_pais_o	Código del país de ocurrencia	60	cer_def_	Número certificado de defunción
19	cod_dpto_o	Departamento de ocurrencia	61	cbmte_	Causa básica de muerte
20	cod_mun_o	Municipio de ocurrencia	62	uni_modif	Desconocido
21	area_	Área de ocurrencia	63	nuni_modif	Desconocido
22	localidad_	Localidad de ocurrencia	64	fec_arc_xl	Desconocido
23	cen_pobla_	Cabecera municipal/centro poblado/rural disperso	65	nom_dil_f_	Nombre del profesional que diligenció la ficha
24	vereda_	Vereda/zona	66	tel_dil_f_	Teléfono del profesional que diligenció la ficha
25	bar_ver_	Código barrio de ocurrencia	67	fec_aju_	Fecha de ajuste (dd/mm/aaaa)
26	dir_res_	Dirección de residencia	68	nit_upgd	NIT de UPGD
27	ocupacion_	Ocupación	69	fm_fuerza	Desconocido
28	tip_ss_	Tipo de régimen en salud	70	fm_unidad	Desconocido
29	cod_ase_	Administradora de Planes de beneficios.	71	fm_grado	Desconocido
30	per_etn_	Pertenencia étnica	72	version	Desconocido

 doi:10.20944/preprints202410.0314.v1

31	nom_grupo_	Nombre grupo étnico	73	nom_eve	Nombre del evento
32	estrato_	Estrato	74	nom_upgd	Nombre de la UPGD
33	gp_discapa	Persona en condición de discapacidad	75	npais_proce	Nombre país de procedencia
34	gp_desplaz	Desplazados	76	ndep_proce	Nombre departamento de procedencia
35	gp_migrant	Migrantes	77	nmun_proce	Nombre municipio de procedencia
36	gp_carcela	Personas privadas de la libertad	78	npais_resi	Nombre país de residencia
37	gp_gestan	Gestantes	79	ndep_resi	Nombre departamento de residencia
38	sem_ges_	Semanas de gestación	80	nmun_resi	Nombre municipio de residencia
39	gp_indigen	Habitantes de la calle	81	ndep_notif	Nombre departamento de notificación
40	gp_pobicbf	Población infantil a cargo del ICBF	82	nmun_notif	Nombre municipio de notificación
41	gp_mad_com	Madres comunitarias	83	nreg	Desconocido
42	gp_desmovi	Desmovilizados			

Table A2. Dictionary of variables of the complementary data of SIVIGILA events 356, 365, and 875.

Event 875		Event 365		Event 356	
Meaning in the Event	**Variable Name**	**Meaning in the Event**	**Variable Name**	**Meaning in the Event**	**Variable Name**
Violencia no sexual	naturaleza	Grupo de sustancias	grupo_sust	fecha ocurrencia (dd/mm/aaaa)	fec_ocurr
Violencia sexual	nat_viosex	Código y nombre del producto	cod_sust	Desconocido	día_ocurrencia
Actividad	actividad	Desconocido	clasificac	Intentos previos	inten_prev
Orientación sexual	orient_sex	Desconocido	categoria	Número de intentos	intentos
Identidad de género	ident_gene	Desconocido	nom_pro	Estado Civil.	estado_civ
Persona consumidora de SPA	consum_spa	Tipo de exposición	tip_exp	Escolaridad	escolarid
Persona con jefatura de hogar	persona_con_ jefatura_ de_hogar	Lugar donde se produjo la intoxicación	lugar_expo	Conflictos con pareja o expareja	prob_parej
Antecedente de violencia	antec	Fecha de exposición (dd/mm/aaaa)	fec_exp	Enfermedad crónica dolorosa o discapacitante	enfermedad_ cronica
Alcohol víctima	presencia_de_ alcohol_u_otra_ sustancia_en_ la_víctima	Hora (0 a 24)	hor_exp	Problemas económicos	prob_econo
Sexo	sexo_agre	Vía de exposición	via_exp	Muerte de un familiar	muerte_fam
Parentesco con la víctima	r_fam_vic		fec_aspers	Escolar / educativa	esco_educ
Convive con el agresor (a)	conv_agre	Escolaridad.	escolarida	Problemas jurídicos	prob_legal
Agresor no familiar	r_nofiliar	Afiliado a A.R.L.	afi_arp	Suicidio de un familiar o amigo	suici_fm_a

 doi:10.20944/preprints202410.0314.v1

¿Hecho violento ocurrido en el marco del conflicto armado?	zona_conf	Código y nombre de la A.R.L	cod_arp	altrato físico / psicológico / sexual	maltr_fps
Mecanismo utilizado para la agresión	mecanismo_ utilizado_ para_la_agresión	Estado civil.	est_civ	Problemas laborales	prob_labor
Quemadura Cara	que_cara	¿El caso hace parte de un brote?: SI, NO	parte_brot	Problemas familiares	prob_famil
Quemadura Cuello	que_cuello	Número de casos en este brote	num_cas_br	Consumo de SPA(Sustancias psicoactivas)	prob_consu
Quemadura Mano	que_mano	Fecha investigación epidemiológica brote (dd/mm/aaaa)	fec_inv_br	Antecedentes Familiares de conducta suicida	hist_famil
Quemadura Píes	que_pie	Situación de alerta	sit_ale	Ideación suicida persistente	idea_suici
Quemadura Pliegues	que_pliegu	Se tomaron muestras de toxicología	muest_toxi	Plan organizado de suicidio	plan_suici
Quemadura Genitales	que_genita	Tipo de muestras solicitada	tipo_muest	Antecedentes trastorno psiquiátrico	antec_tran
Quemadura Tronco	que_tronco	Nombre de la prueba toxicológica	prueba	Trastorno depresivo	tran_depre
Quemadura Miembro superior	que_miesup	Desconocido	fec_expres	Trastornos de personalidad	trast_ personalidad
Quemadura Miembro inferior	que_mieinf	Diligencie Valor resultado /unidades	result_pru	Trastorno Bipolar	trast_ bipolaridad
Grado Quemadura	cla_gra	Desconocido	hor_inv_br	Esquizofrenia	esquizofre
Extensión	ext_que	Grupo de sustancias	grupo_sust	Antecedentes de violencia o abuso	antec_v_a
Fecha del hecho (dd/mm/aaaa)	fec_hecho	Código y nombre del producto	cod_sust	Abuso de alcohol	abuso_alco
Escenario	escenario	Desconocido	clasificac	Ahorcamiento o asfixia	ahorcamien
Ámbito de la violencia según lugar de ocurrencia	ambito_lug	Desconocido	categoria	Elemento Cortopunzante	arma_corto
Profilaxis VIH	sp_its	Desconocido	nom_pro	Arma de Fuego	arma_fuego
Profilaxis Hep B	prof_hep_b	Tipo de exposición	tip_exp	Inmolación	inmolacion
Otras Profilaxis	prof_otras	Lugar donde se produjo la intoxicación	lugar_expo	Lanzamiento al vacio	lanz_vacio
Anticoncepción de emergencia	ac_anticon	Fecha de exposición (dd/mm/aaaa)	fec_exp	Lanzamiento a vehículo	lanz_vehic

References

1. L. Sardinha, M. Maheu-Giroux, H. Stöckl, S. R. Meyer, and C. García-Moreno, “Global, regional, and national prevalence estimates of physical or sexual, or both, intimate partner violence against women in 2018,” The Lancet, vol. 399, no. 10327, pp. 803–813, Feb. 2022, doi: 10.1016/S0140-6736(21)02664-7.
2. A. M. Thurston, H. Stöckl, and M. Ranganathan, “Natural hazards, disasters and violence against women and girls: a global mixed-methods systematic review,” BMJ Glob Health, vol. 6, no. 4, p. e004377, Apr. 2021, doi: 10.1136/BMJGH-2020-004377.

3. "Hechos y cifras: Poner fin a la violencia contra las mujeres | ONU Mujeres." Accessed: Jul. 04, 2024. [Online]. Available: https://www.unwomen.org/es/what-we-do/ending-violence-against-women/facts-and-figures#83918
4. E. Zamora-Moncayo, R. A. Burgess, L. Fonseca, M. González-Gort, and R. Kakuma, "Gender, mental health and resilience in armed conflict: listening to life stories of internally displaced women in Colombia," BMJ Glob Health, vol. 6, no. 10, p. e005770, Oct. 2021, doi: 10.1136/BMJGH-2021-005770.
5. S. Svallfors, "Hidden Casualties: The Links between Armed Conflict and Intimate Partner Violence in Colombia," Politics & Gender, vol. 19, no. 1, pp. 133–165, Mar. 2023, doi: 10.1017/S1743923X2100043X.
6. "Boletín 8 - Datos para la paz - Corte octubre 2023," 2023. Accessed: Dec. 25, 2023. [Online]. Available: https://datospaz.unidadvictimas.gov.co/archivos/datosPaz/boletin_datos_paz_octubre_fronteras.pdf
7. ONU Mujeres, Universidad de Nariño, and Observatorio de género de Nariño, "MUJERES Y HOMBRES: BRECHAS DE GÉNERO EN NARIÑO," 2020.
8. E. García Restrepo, D. Cardona, and A. F. Tirado Otálvaro, "La violencia contra las mujeres en Colombia, un desafío para la salud pública en cuanto a su prevención, atención y eliminación," CES Derecho, vol. 12, no. 1, pp. 167–175, Aug. 2021, doi: 10.21615/cesder.12.1.9.
9. Ministerio de Salud y Protección Social, "Resumen Ejecutivo Encuesta Nacion de Demografía y Salud," 2015.
10. Observatorio de género de Nariño, "Informe Cifras Violeta, Edición VI – Violencia contra las mujeres en Nariño 2015 -2019 – Observatorio de Género de Nariño," 2021. Accessed: Jul. 11, 2024. [Online]. Available: https://observatoriogenero.udenar.edu.co/cifras_violeta_vi/
11. J. R. Sanín, "Violence against Women in Politics: Latin America in an Era of Backlash," https://doi.org/10.1086/704954, vol. 45, no. 2, pp. 302–310, Jan. 2020, doi: 10.1086/704954.
12. A. MariaGiammarioli, E. Longo, R. Bucciardini, A. MariaGiammarioli, E. Longo, and R. Bucciardini, "Gender-Based Violence is a Never to be Forgotten Social Determinant of Health: A Narrative Literature Review," Women's Health Problems - A Global Perspective [Working Title], Jun. 2023, doi: 10.5772/INTECHOPEN.110651.
13. M. Dawson and M. Carrigan, "Identifying femicide locally and globally: Understanding the utility and accessibility of sex/gender-related motives and indicators," https://doi.org/10.1177/0011392120946359, vol. 69, no. 5, pp. 682–704, Aug. 2020, doi: 10.1177/0011392120946359.
14. C. M. Castorena, I. M. Abundez, R. Alejo, E. E. Granda-Gutiérrez, E. Rendón, and O. Villegas, "Deep Neural Network for Gender-Based Violence Detection on Twitter Messages," Mathematics 2021, Vol. 9, Page 807, vol. 9, no. 8, p. 807, Apr. 2021, doi: 10.3390/MATH9080807.
15. I. Rodríguez-Rodríguez, J. V. Rodríguez, D. J. Pardo-Quiles, P. Heras-González, and I. Chatzigiannakis, "Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques," Applied Sciences 2020, Vol. 10, Page 8244, vol. 10, no. 22, p. 8244, Nov. 2020, doi: 10.3390/APP10228244.
16. S. Ulrika Velupillai et al., "Utilizing Text Mining, Data Linkage and Deep Learning in Police and Health Records to Predict Future Offenses in Family and Domestic Violence," Front Digit Health, vol. 3, p. 602683, Feb. 2021, doi: 10.3389/FDGTH.2021.602683.
17. G. R. Bauer, M. Mahendran, C. Walwyn, and M. Shokoohi, "Latent variable and clustering methods in intersectionality research: systematic review of methods applications," Soc Psychiatry Psychiatr Epidemiol, vol. 57, no. 2, pp. 221–237, Feb. 2022, doi: 10.1007/S00127-021-02195-6/TABLES/3.
18. G. Vicente, T. Goicoa, and M. D. Ugarte, "Bayesian inference in multivariate spatio-temporal areal models using INLA: analysis of gender-based violence in small areas," Stochastic Environmental Research and Risk Assessment, vol. 34, no. 10, pp. 1421–1440, Oct. 2020, doi: 10.1007/S00477-020-01808-X/TABLES/6.
19. C.-C. Pinto-Muñoz, J.-A. Zuñiga-Samboni, H.-A. Ordoñez-Erazo, C.-C. Pinto-Muñoz, J.-A. Zuñiga-Samboni, and H.-A. Ordoñez-Erazo, "Machine Learning Applied to Gender Violence: A Systematic Mapping Study," Revista Facultad de Ingeniería, vol. 32, no. 64, p. e15944, Jun. 2023, doi: 10.19053/01211129.V32.N64.2023.15944.
20. K. M. Devries et al., "Intimate partner violence and incident depressive symptoms and suicide attempts: a systematic review of longitudinal studies," PLoS Med, vol. 10, no. 5, 2013, doi: 10.1371/JOURNAL.PMED.1001439.
21. E. Lynn, A. Doyle, M. Keane, K. Bennett, and G. Cousins, "Drug Poisoning Deaths Among Women: A Scoping Review," https://doi.org/10.15288/jsad.2020.81.543, vol. 81, no. 5, pp. 543–555, Oct. 2020, doi: 10.15288/JSAD.2020.81.543.
22. P. Bandara et al., "Domestic violence and self-poisoning in Sri Lanka," Psychol Med, vol. 52, no. 6, pp. 1183–1191, Apr. 2022, doi: 10.1017/S0033291720002986.
23. A. Del Pilar and M. Ramírez, "Modelo para la caracterización y clasificación de los tipos de violencia intrafamiliar desde los registros del sistema de salud".
24. L. Breiman, "Random forests," Mach Learn, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324/METRICS.

25. G. Oh, J. Song, H. Park, and C. Na, "Evaluation of Random forest in Crime Prediction: Comparing Three-Layered Random forest and Logistic Regression," Deviant Behav, Sep. 2022, doi: 10.1080/01639625.2021.1953360.
26. A. Guerrero, J. G. Cárdenas, V. Romero, and V. H. Ayma, "Comparison of Classifiers Models for Prediction of Intimate Partner Violence," Advances in Intelligent Systems and Computing, vol. 1289, pp. 469–488, 2021, doi: 10.1007/978-3-030-63089-8_30.
27. T. S. Biró and Z. Néda, "Gintropy: Gini Index Based Generalization of Entropy," Entropy 2020, Vol. 22, Page 879, vol. 22, no. 8, p. 879, Aug. 2020, doi: 10.3390/E22080879.
28. M. M. Hossain et al., "Prediction on Domestic Violence in Bangladesh during the COVID-19 Outbreak Using Machine Learning Methods," Applied System Innovation 2021, Vol. 4, Page 77, vol. 4, no. 4, p. 77, Oct. 2021, doi: 10.3390/ASI4040077.
29. D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, "MULTI-LABEL CLASSIFIER PERFORMANCE EVALUATION WITH CONFUSION MATRIX," pp. 1–14, 2020, doi: 10.5121/csit.2020.100801.
30. A. Theissler, M. Thomas, M. Burch, and F. Gerschner, "ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices," Knowl Based Syst, vol. 247, p. 108651, Jul. 2022, doi: 10.1016/J.KNOSYS.2022.108651.
31. E. Ileberi, Y. Sun, and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," IEEE Access, vol. 9, pp. 165286–165294, 2021, doi: 10.1109/ACCESS.2021.3134330.
32. Subsistema de información SIVIGILA, "Ficha de notificación individual de Intento de suicidio. Cod INS 356."
33. Subsistema de información SIVIGILA, "Ficha de notificación de intoxicaciones por sustancias químicas Código INS: 365."
34. Subsistema de información SIVIGILA, "Ficha de notificación de vigilancia en salud pública de las violencias de género código INS: 875".