

Article

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

Edna Rocio Bernal-Monroy ¹, Erika Dajanna Castañeda-Monroy ², Rafael Ricardo Rentería-Ramos ¹ ,
Sixto Enrique Campaña-Bastidas ^{1,*}, Jessica Barrera ³ , Tania Maribel Palacios-Yampuezan ¹,
Olga Lucía González Gustin ¹, Carlos Fernando Tobar-Torres ¹ and Zeneida Rocio Ceballos-Villada ¹ 

¹ Department of Engineering, Technical Professional Institute, Faculty of Basic Sciences and Engineering, Universidad Nacional Abierta y a Distancia (UNAD), Calle 14 Sur # 14-23 Barrio Restrepo, Bogotá 111511, Colombia; edna.bernal@unad.edu.co (E.R.B.-M.); rafael.renteria@unad.edu.co (R.R.R.-R.); tania.palacios@unad.edu.co (T.M.P.-Y.); lucia.gonzalez@unad.edu.co (O.L.G.G.); carlos.tobar@unad.edu.co (C.F.T.-T.); zeneida.ceballos@unad.edu.co (Z.R.C.-V.)

² Department of Engineering, Faculty of Electronic Engineering, Universidad Pedagógica y Tecnológica de Colombia (UPTC), Calle 4 A Sur # 15-134, Sogamoso 152211, Colombia; erika.castaneda01@uptc.edu.co

³ Engineering Department, Centro de Investigación y Desarrollo Tecnológico en Ciencias Aplicadas (CIDTCA), Cl. 11 #37-05, Pasto 520001, Colombia; jnbarrera191@gmail.com

* Correspondence: sixto.campana@unad.edu.co

Abstract: This paper explores the application of machine learning techniques and statistical analysis to identify the 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 forests, decision trees, and multiclass classification algorithms combined with fuzzy classification techniques to handle the incomplete data. Implemented in Python and R, the models were statistically validated to ensure their reliability. Analysis based on health data revealed the key victimization patterns and risks associated with gender-based violence in the region. This study presents a data science model that uses a social determinant approach to assess the characteristics and patterns of violence against women in the Pacific region of Nariño. This research was conducted within the framework of the Orquídeas Program of the Colombian Ministry of Science, Technology, and Innovation.

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



Academic Editor: Antony Bryant

Received: 2 October 2024

Revised: 17 January 2025

Accepted: 21 January 2025

Published: 17 February 2025

Citation: Bernal-Monroy, E.R.; Castañeda-Monroy, E.D.; Rentería-Ramos, R.R.; Campaña-Bastidas, S.E.; Barrera, J.; Palacios-Yampuezan, T.M.; González Gustin, O.L.; Tobar-Torres, C.F.; Ceballos-Villada, Z.R. Detection of Victimization Patterns and Risk of Gender Violence Through Machine Learning Algorithms. *Informatics* **2025**, *12*, 21. <https://doi.org/10.3390/informatics12010021>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

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]. According to the United Nations Entity for Gender Equality and the Empowerment of Women (ONU Women), one in three women worldwide 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 [3].

Violence against women is even more severe [4,5] in countries with internal armed conflict such as Colombia. In the Pacific region of Nariño, the township of San Andrés de Tumaco is one of the most affected areas 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.

Another study [7] presents research on the gender gap that reveals that after age 40, a significant percentage of women face limitations in their daily performance due to time constraints that affect rest and the realization of physical and mental self-care activities, concluding that these impacts are acquired and preventable [8].

Gender-based violence significantly affects women and is a public health and safety problem [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 the 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 [11]. The social determinants of health based on gender violence are fundamental to recognizing the actors involved and the various forms of violence faced by women beyond commonly reported physical violence, including the identification of psychological sequelae and violence [12,13].

Establishing patterns in the official national health reports is a pivotal tool for characterizing the needs and the violence faced by women in this district. In this sense, classification techniques and the identification of characteristics through machine learning (ML) are relevant, given their versatility. Throughout 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 a review of the 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 a 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 instance, in [14], a deep neural network model is implemented to identify gender violence in Twitter messages in Mexico, achieving an effectiveness of 80%. Similarly, ref. [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 using police and health records to predict future family and domestic violence crimes, easing 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 with a high incidence of violence against women.

The above is related to what the authors of [19] mention, explaining how applying different ML techniques for studying and identifying patterns of gender violence represents an efficient alternative worldwide. 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, we developed a tool for screening patterns of victimization and the risk of gender violence using ML algorithms.

3. Methodology

This section discusses the implementation in San Andrés de Tumaco, Colombia, of a decision-making tool, aimed at analyzing its effects on health using ML algorithms. This was developed through the following stages:

1. The identification of sources, data collection, and preparation;
2. The exploration of ML techniques for pattern detection;
3. The evaluation of these techniques in pattern detection.

3.1. Identification of Sources, Data Collection and Preparation

In Colombia, the Public Health Surveillance System (SIVIGILA) collects and reports updated and detailed information on public health events and violence across all regions of the country. This system includes specialized databases that enable the precise tracking of various phenomena, such as suicide attempts, poisonings, and gender-based violence.

In the context of this research, events 356, 365, and 875 were selected, corresponding to "Suicide Attempts", "Poisonings", and "Gender-Based and Domestic Violence", respectively. These events are particularly relevant for analyzing gender-based violence from the perspective of the social determinants of health. Several studies have demonstrated the close relationship between gender-based violence and the increase in suicide attempts and self-harming behaviors [20]. Moreover, poisoning may be a direct or indirect consequence of these forms of violence, either as an act of self-harm, or as a result of substance abuse as a coping mechanism [21,22].

The dataset provided by SIVIGILA was characterized by organizing the information into two main sections: basic data and complementary data. Having basic data allowed for the sociodemographic analysis of the events and categorized the cases based on their specific characteristics, including 82 fields in notification form. On the other hand, complementary data offered a more detailed description of the type of event that occurred, taking into account various specific characteristics related to each case.

The complementary data from the datasets analyzed in this work for events 356, 365, and 875 contain specific variables that allow for the analysis of the context and situation of recorded violence. For event 356, related to suicide attempts, 465 records of women were collected, along with 52 variables that detail the circumstances of the attempt, the methods used, risk and protective factors, as well as the sociodemographic and clinical characteristics of the affected individuals. These data include age, municipality of residence, substance use, and history of depressive disorders, which facilitate the identification of risk patterns and the exploration of the social determinants associated with suicidal behavior. For event 365, related to poisonings, 297 records were available, and 49 variables document exposure to chemicals, the amount ingested, the type of toxic agent, and related circumstances, such as prior substance use and the level of medical attention received. These data are essential for modeling the relationships between risk factors and the consequences of poisoning, enabling the identification of patterns and predicting behaviors linked to social vulnerability and violence. Finally, for event 875, which addresses gender-based and

domestic violence, 3030 records were available, and 41 variables were used to characterize both victims and perpetrators, as well as to detail the circumstances of the event, such as the type of violence (physical, psychological, sexual, or economic) and the history of prior violence. This dataset is crucial for identifying patterns of violence and vulnerability, as well as assessing the effectiveness of social and health responses. Between 2017 and 2023, the district mayor's office of San Andrés de Tumaco collected a total of 737 reports for event 356, 708 for event 365, and 3680 for event 875, enabling in-depth analysis of the dynamics and social factors influencing these phenomena. Appendix A presents variables of the basic and complementary data for each of the collected SIVIGILA events.

3.2. Exploration of ML Techniques for Pattern Detection

Exploratory analysis for detecting gender violence occurrence patterns in San Andrés de Tumaco relies on the characteristics established in the SIVIGILA event database. Algorithms, such as decision trees, random forests, artificial neural networks (ANNs), and clustering usually show better results for analyzing patterns in small datasets [23,24].

Analysis of each of the datasets included the search for similarities and correlations between the suicide attempts and the intoxications with gender and intrafamily violence to explain the recurrence of these events. In these cases, 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 suicide attempt reports and intoxications is lower than the number of 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 that given in [25] or [26] has demonstrated their effectiveness in identifying the 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 relationships among the variables that clearly characterize gender violence in detail allows us to recognize and explain those cases, suicide attempts, and intoxications, which possibly occur as a consequence of an episode of gender or intrafamily violence. For this purpose, the most relevant variables were selected, the cut-off point was maximized, and the impurities caused by the 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 the patients referred to psychiatry and social work professionals as a possible predictor of the relevant characteristics of women who had attempted suicide, 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, 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 [23–25,28].

Random forests were applied to the available health events, as multiple internal subgroups of the original dataset were created by sampling with replacement; consequently, for each subgroup, a decision tree was elaborated, and at each node a subgroup of features of size “mtry” was randomly selected. The best-split point was only determined from the selected features. The maximum depth of the trees was limited with the parameter

“max_depth”, and cross-validation was 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 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 of 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 the number of neurons (input and output) and layers (hidden), and the activation algorithm for the better adjustment of weights. The evaluation methods given in Section 3 tested the degree of assertiveness when implementing these algorithms.

Moreover, implementing clustering methods allowed for 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 the 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 when working with outliers. The latter was applied to the 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. This can be valuable for recognizing groups that tend to normalize violence and that other methods do not detect.

For the recognition of victimization patterns and gender violence risk, a robust and diverse approach was applied based on the exploration of decision trees, random forests, and ANNs. This perceives the influences of variables and the connection between the 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. It is worth noting that the models were implemented in Python and R and subjected to statistical validation to ensure their reliability.

This study prioritized analysis with the Partitioning Around Medoid (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.

3.3. Evaluation of the Techniques in Pattern Detection

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

Moreover, 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 previous events, taking into account only the minimum number of variables obtained as viable. These simulations were performed by adjusting probabilistic distributions of the 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 against women in San Andrés de Tumaco from 2017 to 2023. The violence rates are given with three decimal places. Simulations for projecting the figures were conducted using a decision-making tool developed with Power BI (Microsoft Power BI 2016). This integrated the presentation of statistics from the previous reports and their forecasts. The implementation of the trained models for each event studied was initially carried out locally in R (4.4.1), and later migrated to Python (3.13) for integration with the decision-making tool developed in PowerBI.

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

4. Results

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

The first exploratory analysis complemented this study by initially cleaning and labeling the data on suicide attempts, intoxications, and gender and intrafamily violence. A few isolated cases with the predominance of violence were identified, standing out for their severe consequences on the health and life of the victims. This hindered the recognition of the profiles of potential victims, as it highlighted the general context and covered the patterns of cases with less alertness. Pointing out these cases facilitated a more contextual approach, considering the experience and type of violence suffered, not only the consequences of the event. Thus, 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 variables of the SIVIGILA 356, 365, and 875 events were compiled in the dictionary of variables in Appendix A. This identified the characteristics of the information consolidated in this system based on the study of notification forms used by the officials of 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 was 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 workers of the women who attempted 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 the cases that needed a referral to social workers. 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.

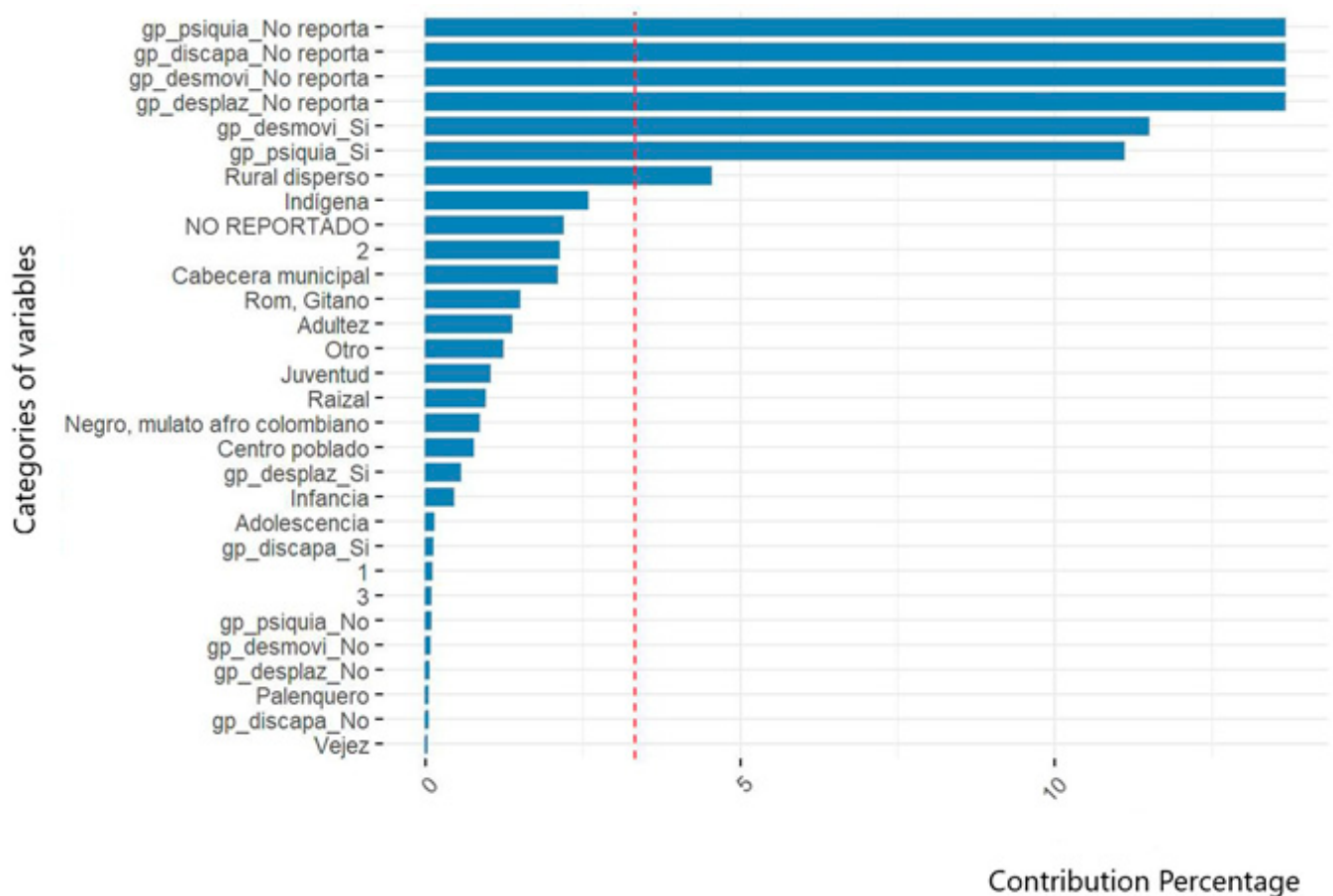
The data in Table 1 were used for statistical and exploratory analyses. The variables of events 356 and 365 were related to the detection of patterns in the algorithms, given the context from the researchers; however, when victimization patterns on the potential victim profiles were identified, the definitive variables were the area where the event occurred, 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 violent event and whether

the woman or girl lived with the aggressor were analyzed, as well as whether she was alerted or was referred to a mental health or social work service.

Table 1. Variables used in exploratory statistical analysis. (For more information see Appendix A).

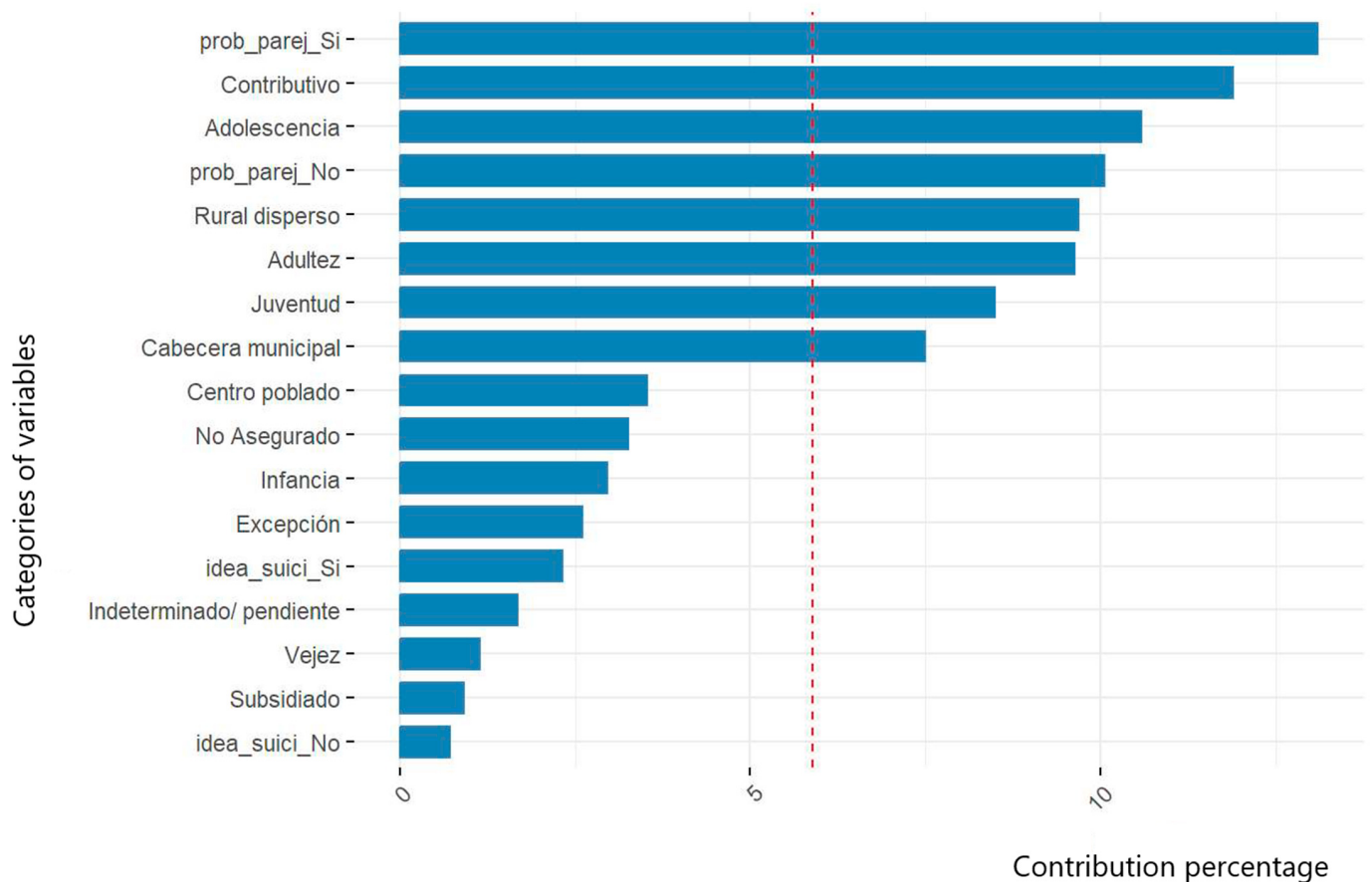
Event	Variables
356	area_, stratum_, life_cycle, per_ethnic_, gp_displacement, gp_dislocation, gp_disability, gp_psychia, ss_type, prev_intensive, hos_pac_, couple_problem, economic_problem, legal_problem, labor_problem, family_problem
365	area_, stratum_, life_cycle, ss_type, pac_hos_, sit_ale, typ_exp, con_fin_, sit_ale, est_civ
875	area_, life_cycle, tip_ss_, nature, conv_agre

In this regard, the 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 this research helped define the profiles of victims in events 356, 365, and 875. Figure 1a demonstrates the impact of unreported cases on the 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)

Figure 1. Cont.



(b)

Figure 1. Contribution of variable category to victim profiles of event 356. Suicide attempt in San Andrés de Tumaco. (a) Considering “not-reported” cases. (b) Discriminating non-reported cases. The red line corresponds to the significance threshold to underline the contribution of the variables to the aggregate.

Similarly, MCA was conducted on the identified variables, and clustering was applied to the common group characteristics among the women affected in each event. The MCA of the event data revealed that factors such as life cycle, the area of occurrence, and the health status of the victim are shared elements in cases of gender-based violence. These factors allowed for the recognition of violence occurrence patterns in urban and dispersed rural areas. In addition, it was observed that suicide attempts are related to the life cycles of women in 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 the violence experienced by mothers and neglect in the care of their daughters.

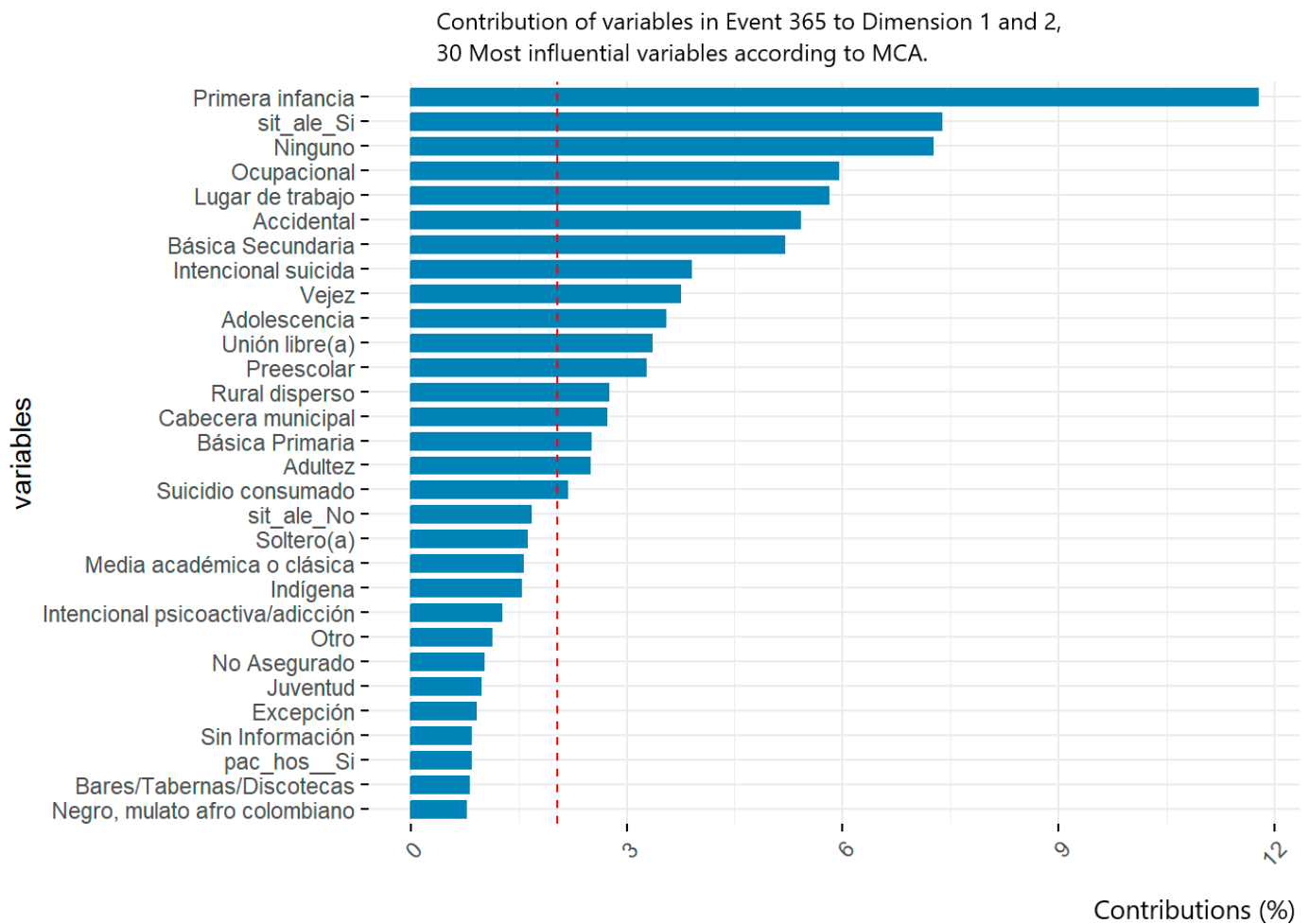


Figure 2. Contribution of definitive variable categories to the victim profile for event 365—intoxication in San Andrés de Tumaco. The red line corresponds to the significance threshold to underline the contribution of the variables to the aggregate.

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

Furthermore, it was found that victims who live with their aggressors suffer principally from physical and psychological violence, neglect, and abandonment, aspects that belong to the life cycles of women in 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.

Contribution of the 35 most influential variables in Event 875 to dimensions 1 to 3 According to MCA

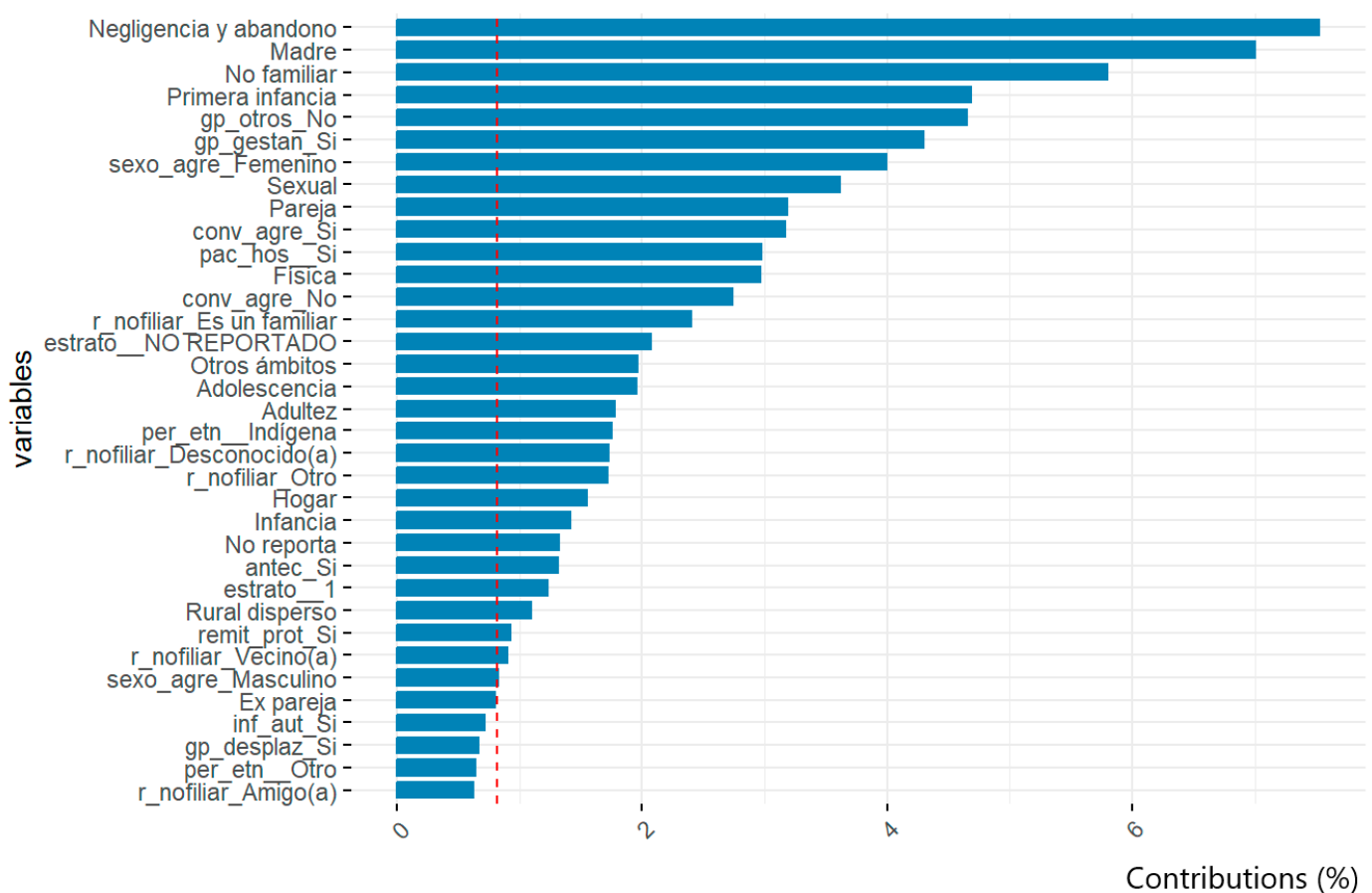


Figure 3. Contribution of definitive variable categories to victim profiles of event 875. Gender-based and domestic violence in San Andrés de Tumaco. The red line corresponds to the significance threshold to underline the contribution of the variables to the aggregate.

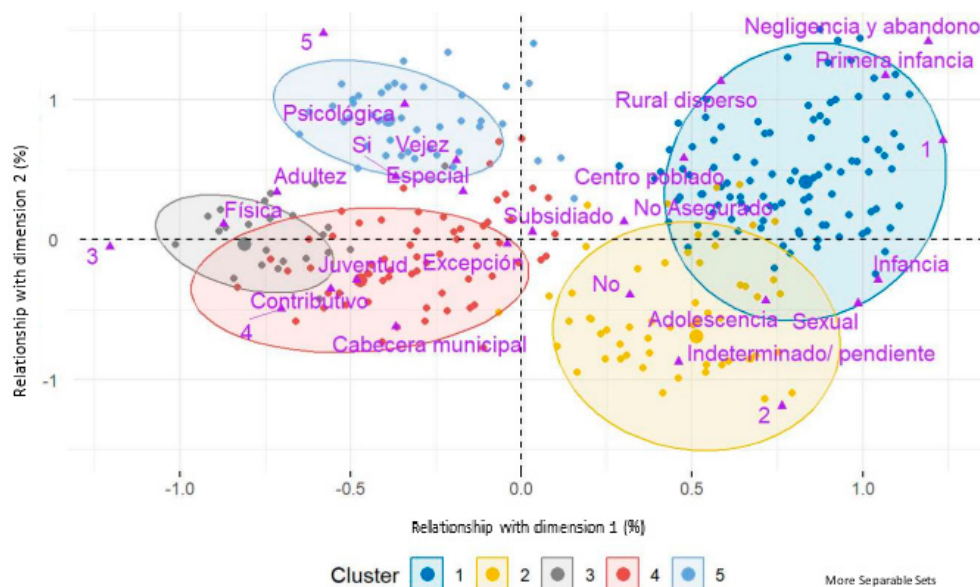


Figure 4. Scatter plot of grouping provided by PAM clustering for event 875. Gender and domestic violence in San Andrés de Tumaco.

4.1. Implementation of Decision Trees

Based on the selection of variables that most contributed to characterizing the victim profile and the application of decision tree algorithms, the following results were 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 referral to social work for women who had attempted suicide by analyzing several characteristics.

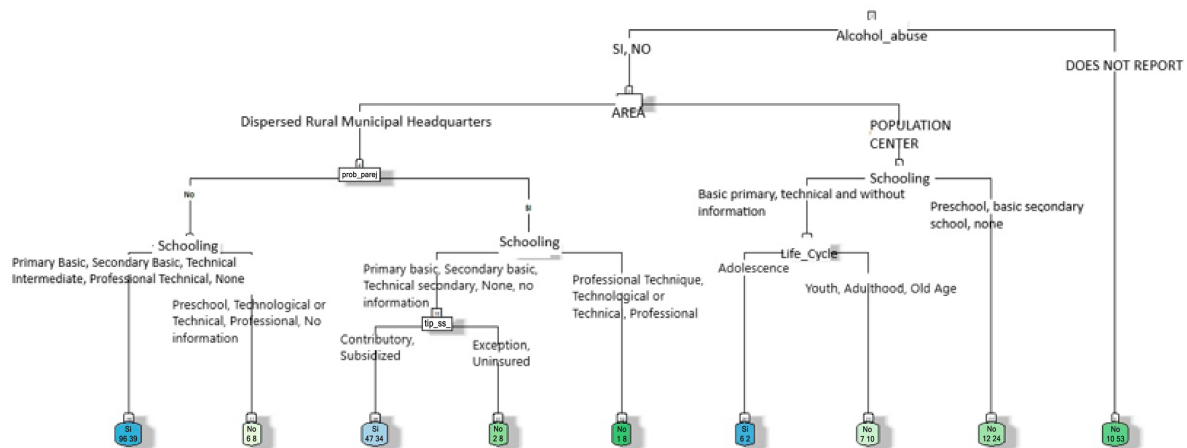


Figure 5. Decision tree for predicting social work referrals of women who had attempted suicide.

Alcohol abuse was reported by 57.10% of the 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 workers, whereas 50% were referred to social workers when relationship problems existed. The 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 not referred to social workers. These data underline the idea that factors such as alcohol abuse, the area of residence, partner issues, and educational level influence referrals to social workers.

The confusion matrix shown in Figure 6 demonstrates the ability of decision trees to predict referral to social workers for women who had attempted suicide. 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.

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

Table 2. Classification metrics for social work referral according to decision tree for women who had attempted suicide.

Accuracy	Recall	Precision	F1 Score
58.69%	56.66%	73.91%	64.15%

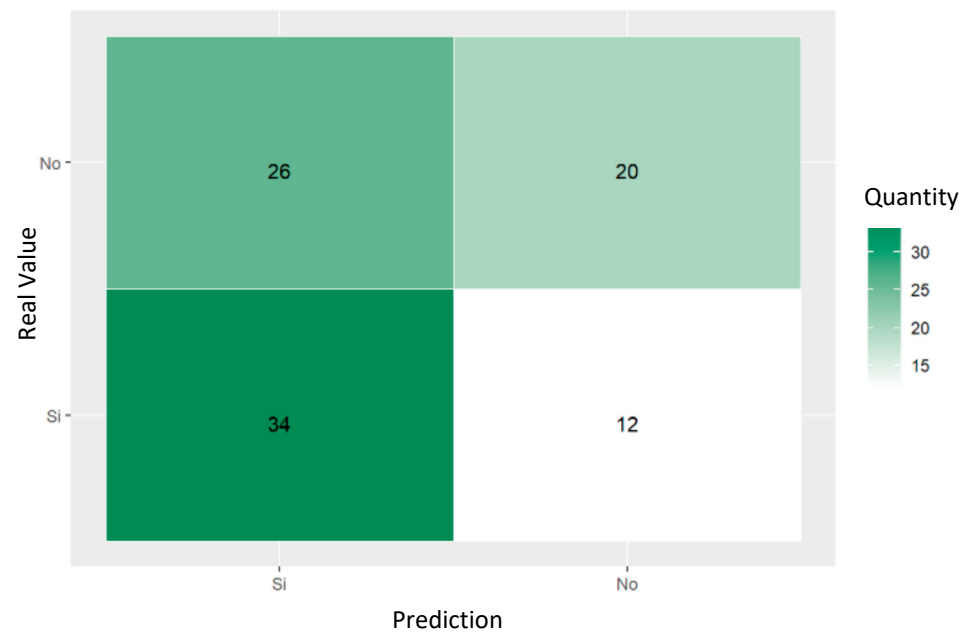


Figure 6. Confusion matrix of trained decision tree for predicting social work referrals of women who had attempted suicide.

The classification results of the women referred to psychiatric services obtained 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 services of the women who had attempted suicide 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 services 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 under 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%.

Women who were referred to psychiatry for suicide attempts. The higher the density of the group. That is, if there is greater probability, the darker the color and in the blue ones belong to that if it is more likely to be referred to psychiatry services, the darker then the branch of the tree, the more likely it is to be referred to psychiatry services, and the lighter the less likely, the darker the more likely it is almost certainly to be referred to psychiatric services for attempted suicide.

The confusion matrix shown in Figure 8 and the decision tree classification metrics shown in Table 3 predicting referrals to psychiatry services 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 ten cases of referral (true positives) and six cases of non-referral (true negatives) were properly predicted, although there were also 17 false negatives and 13 false positives. These results indicate the 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.

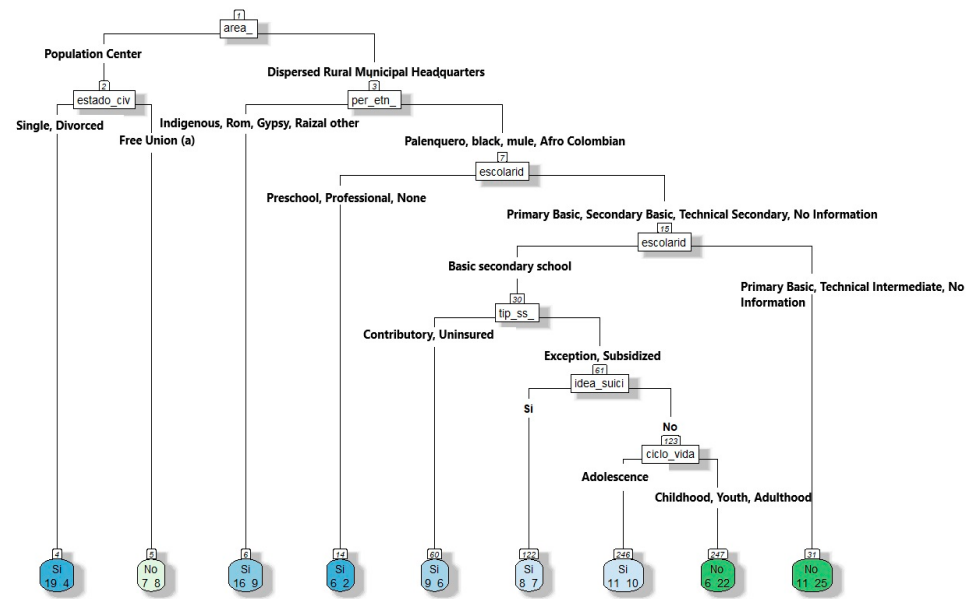


Figure 7. Decision tree for predicting referral of women who had attempted suicide to psychiatry services.

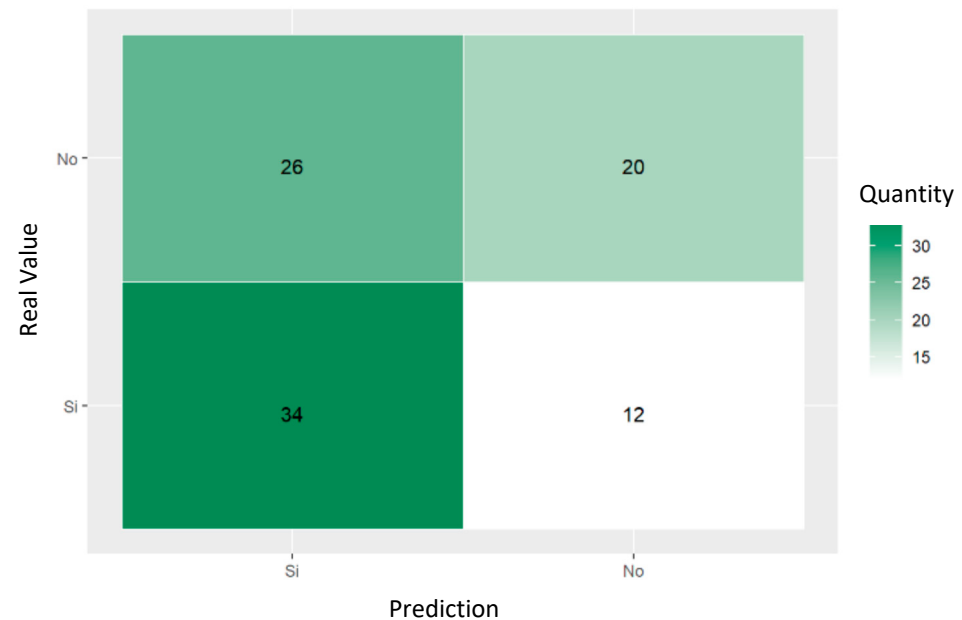


Figure 8. Confusion matrix of trained decision tree for predicting psychiatric referral of women who had attempted suicide.

Table 3. Classification metrics for psychiatric referral using decision tree in women who had attempted suicide.

Accuracy	Sensitivity	Precision	F1 Score
65.21%	62.96%	73.91%	68%

For the analysis of event 365, poisonings, the tree shown in Figure 9 was elaborated upon, capturing the characteristics of the poisonings of women in “alert situations”. The root node indicates 81.51% of the cases classified as “no”, and 18.49% as “yes”. The variable ‘life cycle’ divides the sample into early childhood, childhood, adolescence, youth, adulthood, etc. The majority of accidental poisonings are classified as “yes”. The type of health regime indicates that most cases occur under a subsidized or uninsured regime,

where there is a higher proportion of cases classified as “yes”. In contrast, the branch with infancy, adolescence, youth, adulthood, and 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 the type of exposure and health regime influence them.

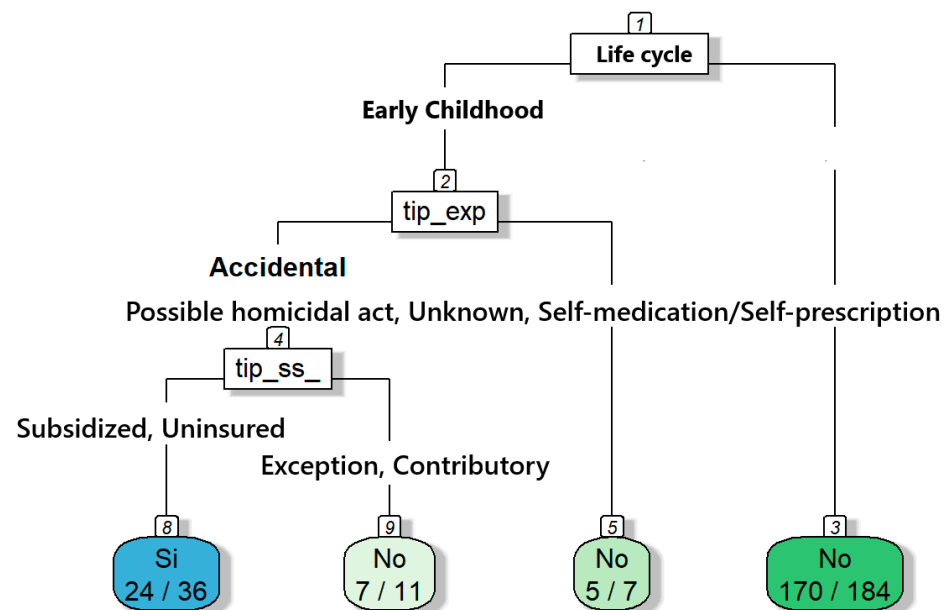


Figure 9. Decision tree focusing on predicting women who have been referred for poisoning in alert situation.

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

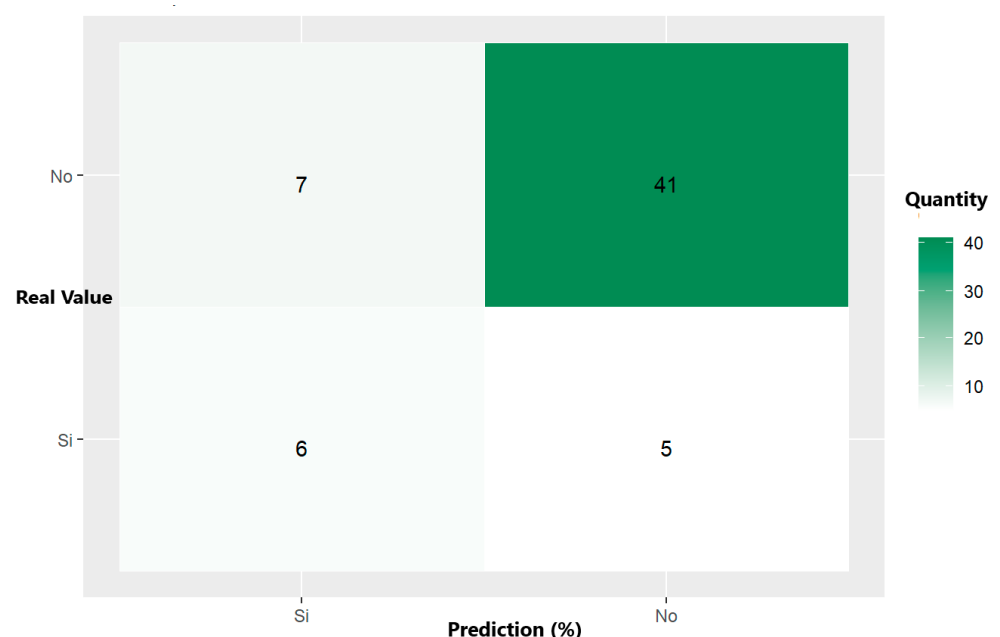
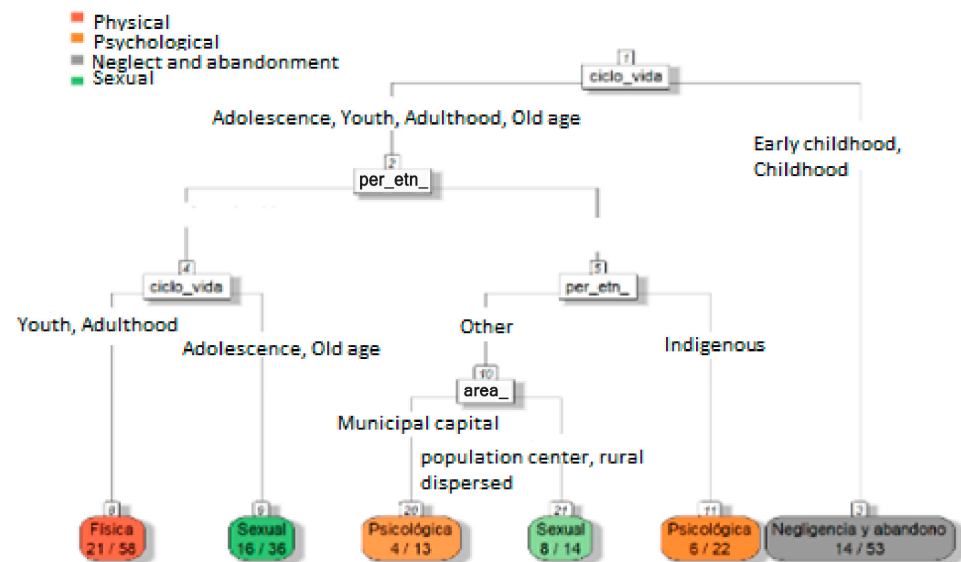


Figure 10. Confusion matrix of trained decision tree for predicting alert situations for poisonings.

Table 4. Classification metrics for psychiatric referral using decision tree in women who are poisoned.

Accuracy	Sensitivity	Precision	F1 Score
79.66%	46.15%	54.54%	50%

For the analysis of event 875, gender violence, the tree shown in Figure 11 was elaborated upon with characteristics that classify the nature of situations of physical, psychological and sexual violence, neglect, and abandonment using variables such as life cycle, ethnicity, the type of social service, and the 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 mixed race youths and adults, and sexual abuse is more frequent in adolescents and older people 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.

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

The confusion matrix of the decision tree model shown in Figure 12 and the classification metrics given in Table 5 reflect its strong 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%.

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

Class	Accuracy	Sensitivity	Precision	F1 Score
Physical	66.66%	58.33%	43.75%	50%
Psychological	68.05%	41.66%	71.42%	52.63%
Neglect and abandonment	83.33%	83.33%	62.50%	71.42%
Sexual	54.16%	25%	33.33%	28.57%

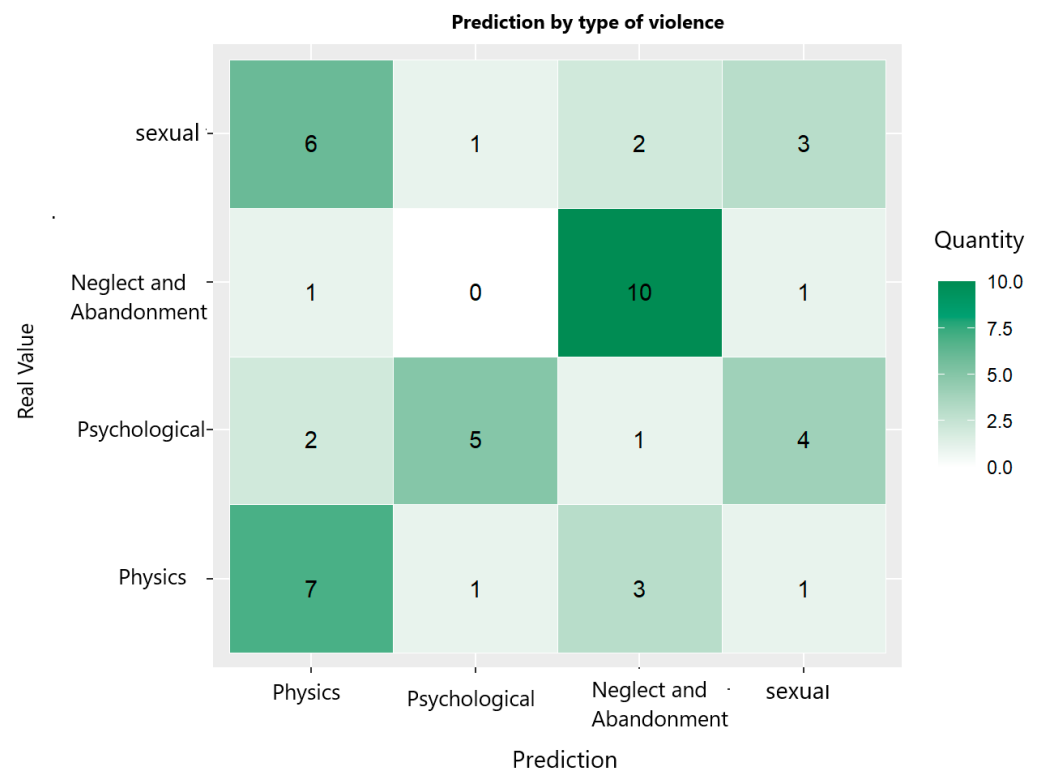


Figure 12. Confusion matrix of decision tree trained to predict type of violence suffered by women.

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

4.2. Implementation of Random Forests

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

Given the problem of unreported data in the variable stratum, this was excluded from training. The use of the other typologies was scaled in a new algorithm, focusing the analysis of characteristics of the individuals. 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, the type of social security, alcohol abuse, ethnicity, marital status, relationship problems, economic problems, and suicidal thoughts.

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 61.08%, a precision of 61.05%, a recall of 62.36%, and an F1 score of 61.70%, indicating a moderate ability for effective identification and referral to social workers. The selection of these variables has improved the model's accuracy and reflects the adequate identification of relevant features.

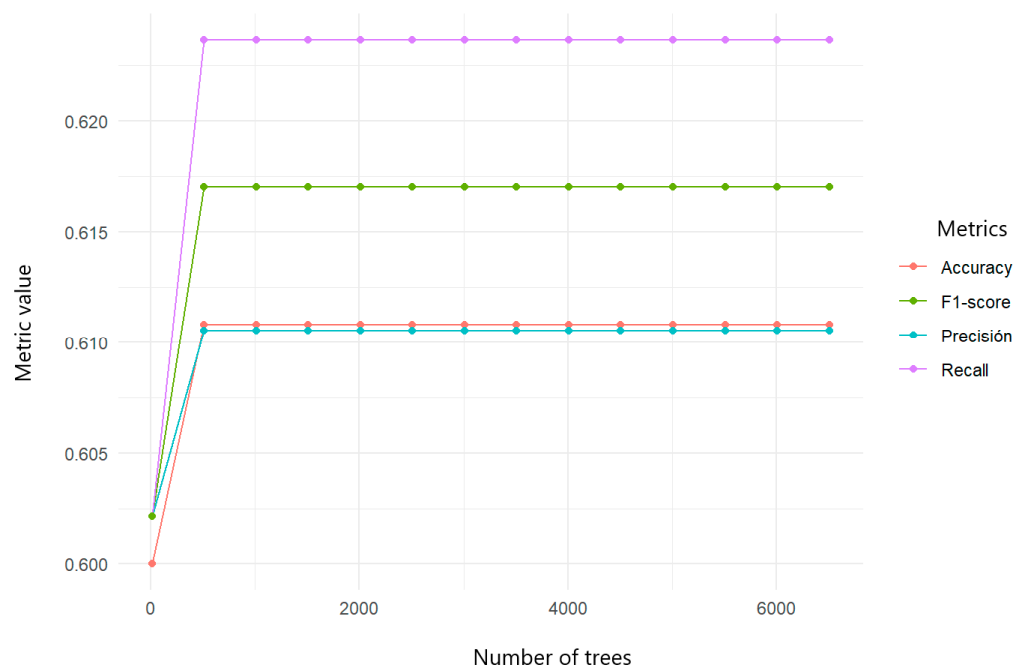


Figure 13. Cross-validation applied to random forests in prediction of referrals of women experiencing abuse to social work services.

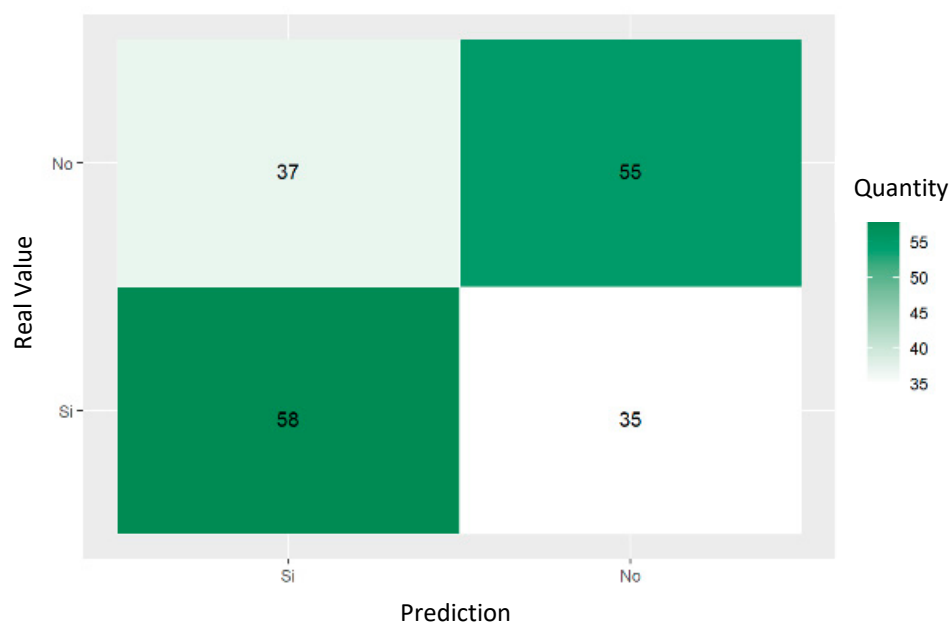


Figure 14. Confusion matrix random forest for predicting referrals to social workers of women who had attempted suicide.

Table 6. Classification metrics for referral to social workers in suicide attempt cases according to random forest.

Accuracy	Precision	Recall	F1 Score
61.08%	61.05%	62.36%	61.70%

Also, to characterize the profiles of the women who had attempted suicide referred to psychiatry services, a random forest model was trained with complementary data variables, identifying the most important: schooling, stratum, area, life cycle, the type of social security, place of attempt, alcohol abuse, ethnicity, marital status, marital problems, the

week of gestation, and suicidal ideation. Additional tests with these variables showed that life cycle, area, schooling, the 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.

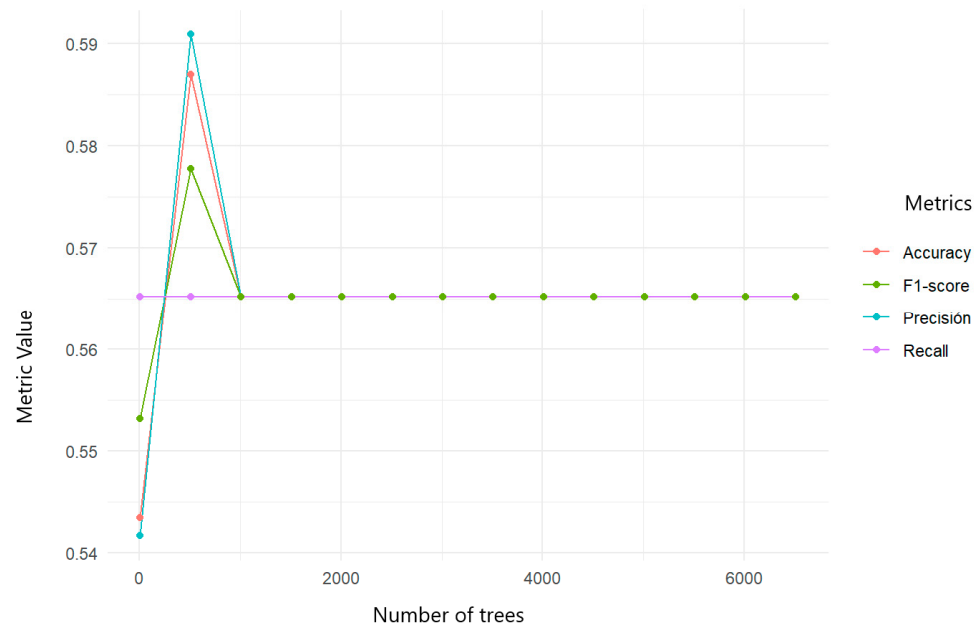


Figure 15. Cross-validation applied to random forests in predicting referrals of women who had been physically assaulted to psychiatric services.

For predicting psychiatric referral, the random forest model obtained an accuracy of 58.69%, a precision of 59.09%, a recall of 56.52%, and an F1 score of 57.77% (Table 7).

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

Accuracy	Precision	Recall	F1 Score
58.69%	59.09%	56.52%	57.77%

The confusion matrix shown in Figure 16 shows 9 true negative cases, 14 false positives, 13 false negatives, and 10 true positives, indicating a moderate algorithm performance.

With the implementation of the random forest model to improve the decision-making capacity in alert cases and review the additional variables of event 365 (see Table A2, Appendix A), 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 a 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), ten false positives (FPs), and ten false negatives (FNs).

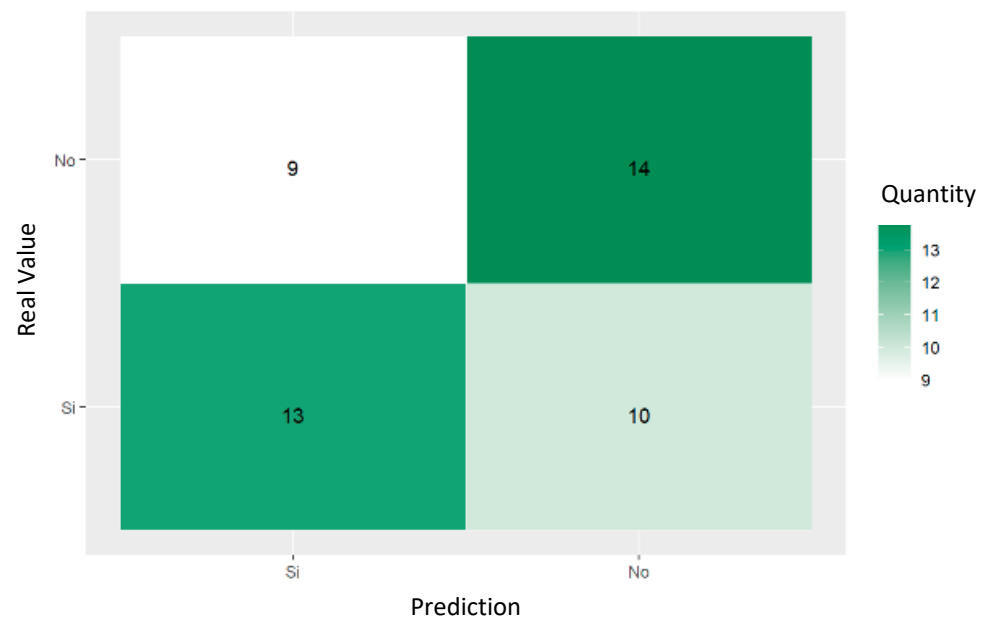


Figure 16. Random forest confusion matrix trained for predicting psychiatric referrals of women who had attempted suicide.

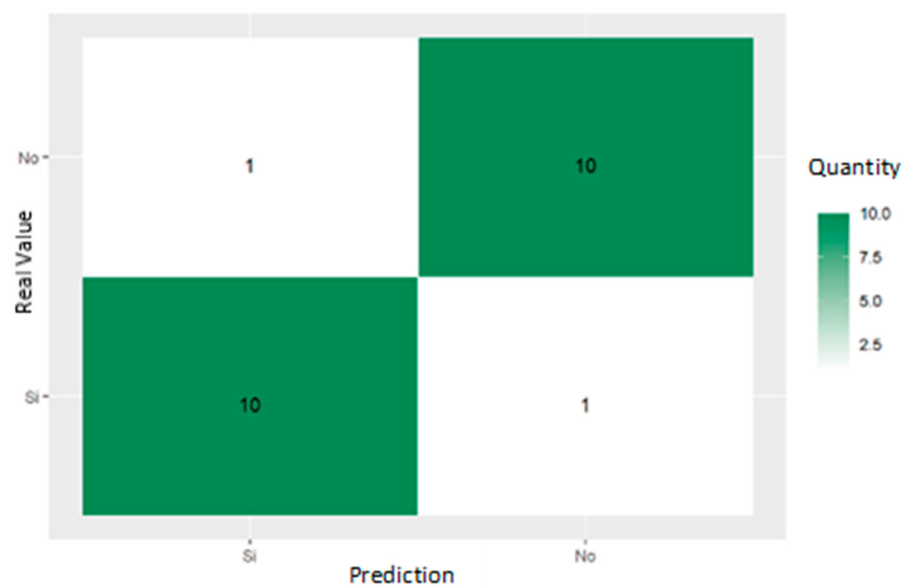


Figure 17. Confusion matrix random forest trained to predict alertness in poisoning.

Despite these values, the performance metrics (Table 8) are high; accuracy is 90.90; precision is 90.90; recall is 90.90, and the F1 score is also 90.90, 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 alert situation in random forest poisonings.

Accuracy	Precision	Recall	F1 Score
90.90%	90.90%	90.90%	90.90%

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

To improve the decision-making capacity for the types of violence, the 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.

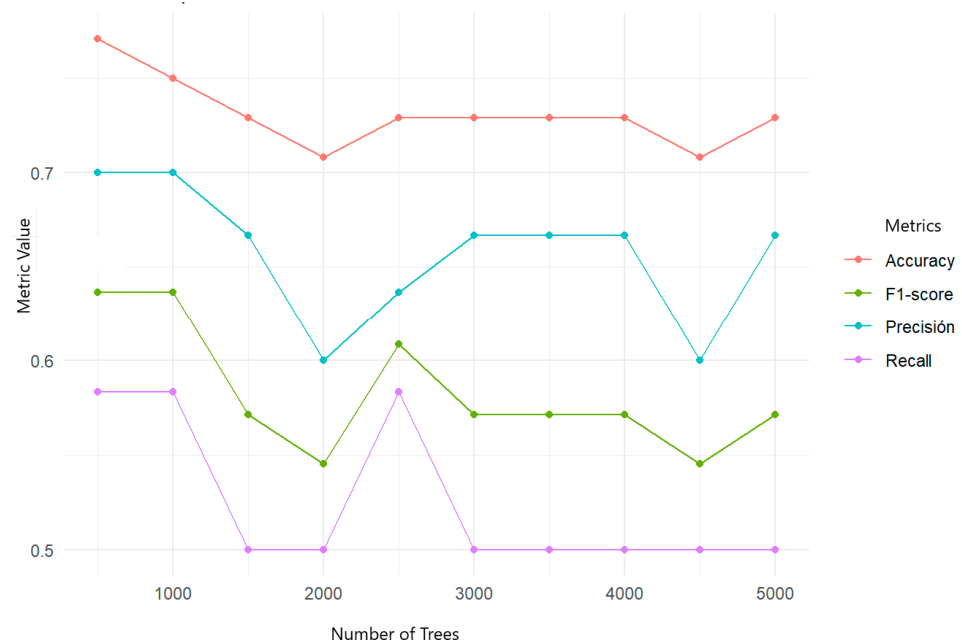


Figure 18. Cross-validation applied to random forests in predicting nature/type of violence suffered by women in relation to gender and domestic violence.

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 other variables, *cycle_life*, *per_ethn_*, *area_*, *conv_agre*, and *tip_ss_* (see Table A1) to improve the characterization of potential victims.

The confusion matrix shown in Figure 19 and the classification metrics given in Table 9 indicate that the random forest model trained with *cycle_life*, *per_ethn_*, *area_*, *conv_agre*, and *tip_ss_* (see Table A2) has a sound predictive capacity to classify violence. The accuracy, sensitivity, precision, and F1 score metrics indicate an 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 violence, suggesting improvements in future model iterations.

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

Class	Accuracy	Recall	Precision	F1 Score
Physical	66.66%	58.33%	43.75%	50%
Psychological	68.05%	41.66%	71.42%	52.63%
Neglect and abandonment	83.33%	83.33%	62.50%	71.42%
Sexual	54.16%	25%	33.33%	28.57%

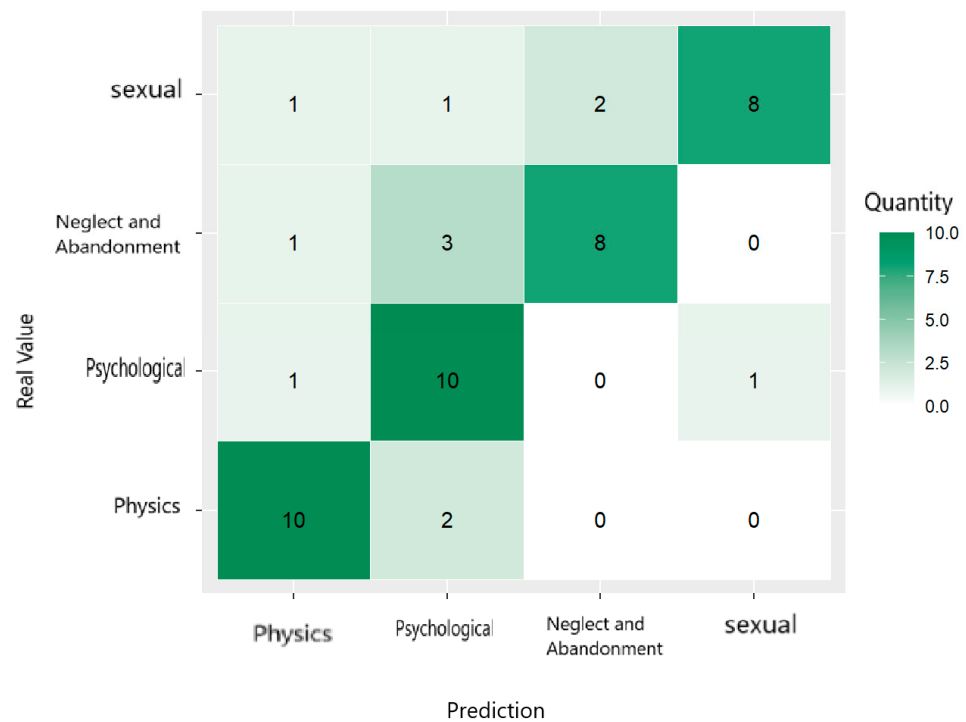


Figure 19. Confusion matrix. Random forest trained to predict nature/type of violence suffered by women in relation to gender and domestic violence.

4.3. Artificial Neural Network (ANN)

The neural network designed to characterize psychiatric referrals in the cases of attempted suicide included four hidden layers with five hundred, one hundred, six, and two neurons, and a convergence threshold of 0.0, with a maximum step size of 1×10^{14} . 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 a 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 shown in Figure 20 for the trained neural network revealed a poor performance in predicting referrals to social workers in the reports of suicide attempts, as Table 10 shows. The high mean prediction error rate, 62.86%, and entropy problems indicate that the model is ineffective in correctly distinguishing between the cases that are and are not referred to social workers.

Table 10. ANN classification metrics for social worker referral.

Accuracy	Precision	Sensitivity	F1 Score
37.14%	35.93%	32.85%	34.32%

For the 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 achieve predictive capacity, and the same entropy and prediction error problems as those of the previous network were generated, becoming noticeable in the confusion matrix shown in Figure 21 and the metrics given in Table 11. Also, the training time was five times longer due to the need to balance the collected data.

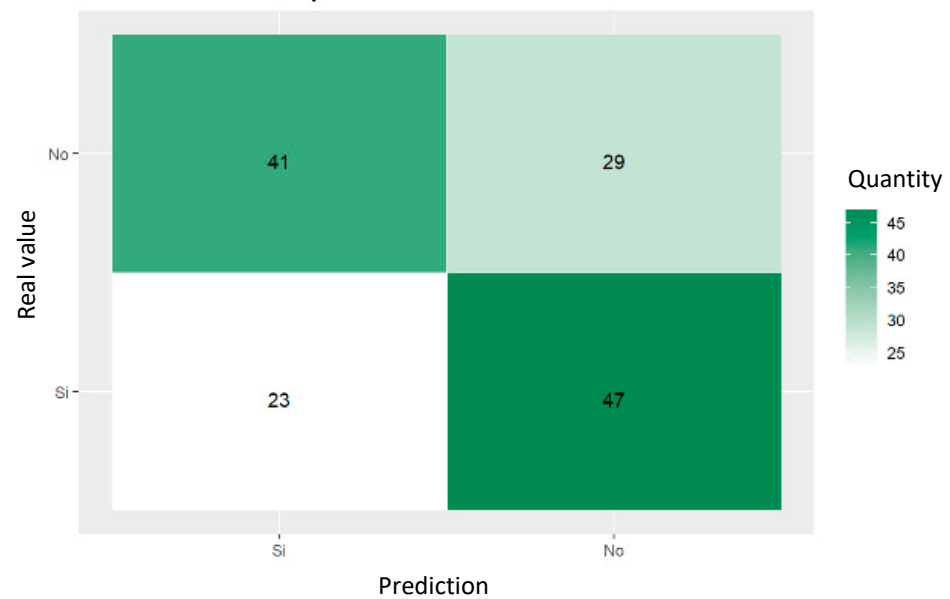


Figure 20. ANN confusion matrix trained for predicting social worker referral.

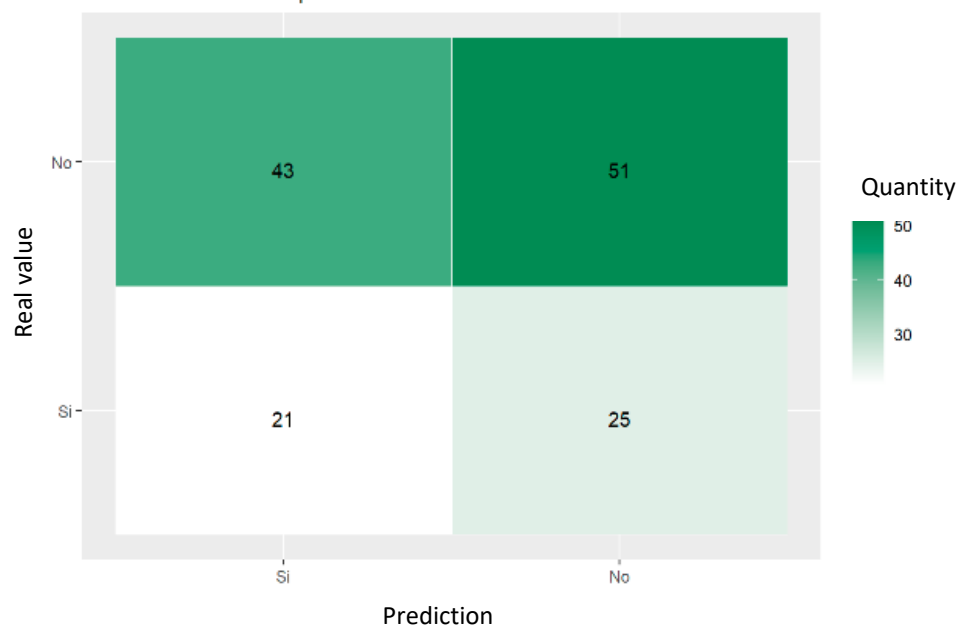


Figure 21. ANN confusion matrix trained for predicting referral to psychiatric service.

Table 11. ANN classification metrics for psychiatric referral.

Accuracy	Precision	Recall	F1 Score
51.43%	32.81%	45.65%	38.18%

The confusion matrix shown in Figure 21 for the neural network trained to predict referral to psychiatry services showed a poor 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 precision of 32.81%, a recall of 45.65%, and an F1 score of 38.18%, it is clear that the model suffers from a 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 services 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, or 4) and specific parameters was used to optimize the predictive performance.

The confusion matrix shown in Figure 22 reveals that the neural network model contained a high mean prediction error (1.970297) and an undefined mean entropy (NaN), indicating a poor performance. Considering the information in Table 12, classification yielded improved results in the cases of physical and sexual violence, with precision of 75.69% and 74.55% and sensitivity of 80.63% and 69.61%, respectively. However, its performance was scarce concerning psychological violence, as its precision stands at only 3.77%, its sensitivity is 12.50%, and it did not identify the cases of neglect and abandonment, reflecting a lack of correct predictions for this category.

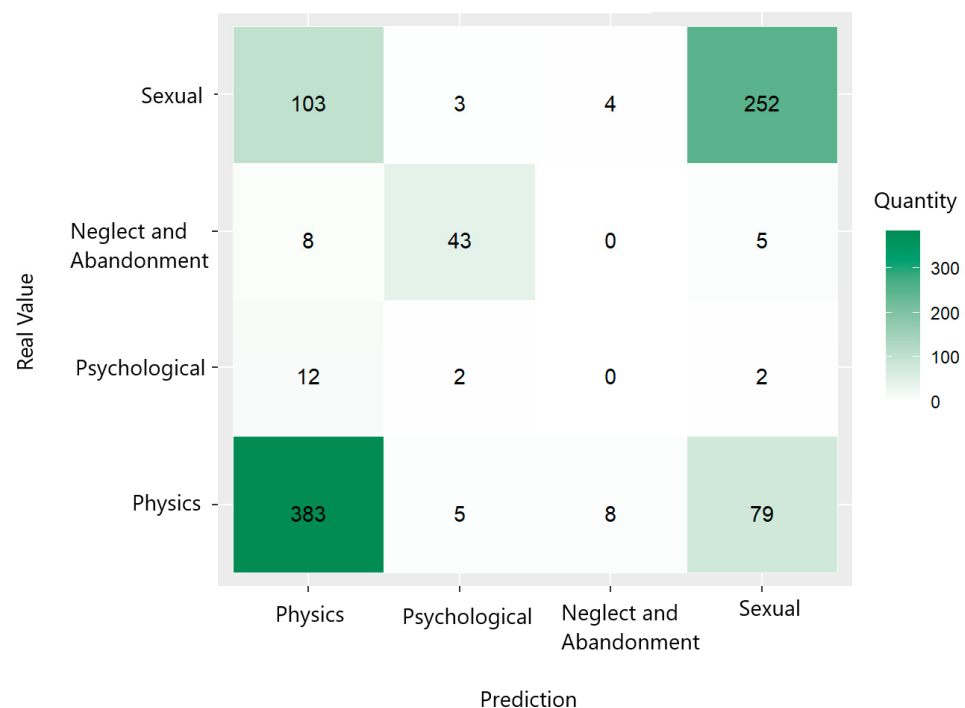


Figure 22. ANN confusion matrix trained to predict the nature/type of violence suffered by women in relation to gender and domestic violence.

Table 12. Metrics for classifying violence type using ANN.

Class	Accuracy	Sensitivity	Precision	F1 Score
Physical	76.14%	80.63%	75.69%	78.08%
Psychological	53.39%	12.50%	3.77%	5.79%
Neglect and abandonment	49.29%	0%	0%	NaN%
Sexual	76.94%	69.61%	74.55%	72.00%

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 detailed statistical analysis of the dataset needed to predict violence in San Andrés de Tumaco. A distribution 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 resulted in 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.

Año	Total Female Population	Cases of Gender-Based and Domestic Violence Against Women	Women Who Experienced Gender-Based and Domestic Violence	Rate of Violence Cases Against Women per 1000 Women	Rate of Women Who Experienced Violence per 1000 Women
2018	128 752	366	355	2.84	2.73
2019	129 923	401	374	3.08	2.87
2020	132 176	324	296	2.45	2.23
2021	133 515	392	363	2.93	2.71
2022	134 277	406	374	3.02	2.78
2023	135 117	515	467	3.81	3.45

These victim profiles were used to make specific predictions, applying clustering techniques according to 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. The interface of the section presenting the gender violence figures in San Andrés de Tumaco and the anonymous reporting button from the decision-making tool developed in Power BI. Source: <https://linktr.ee/OrquideasTumaco> (accessed on 2 October 2024).

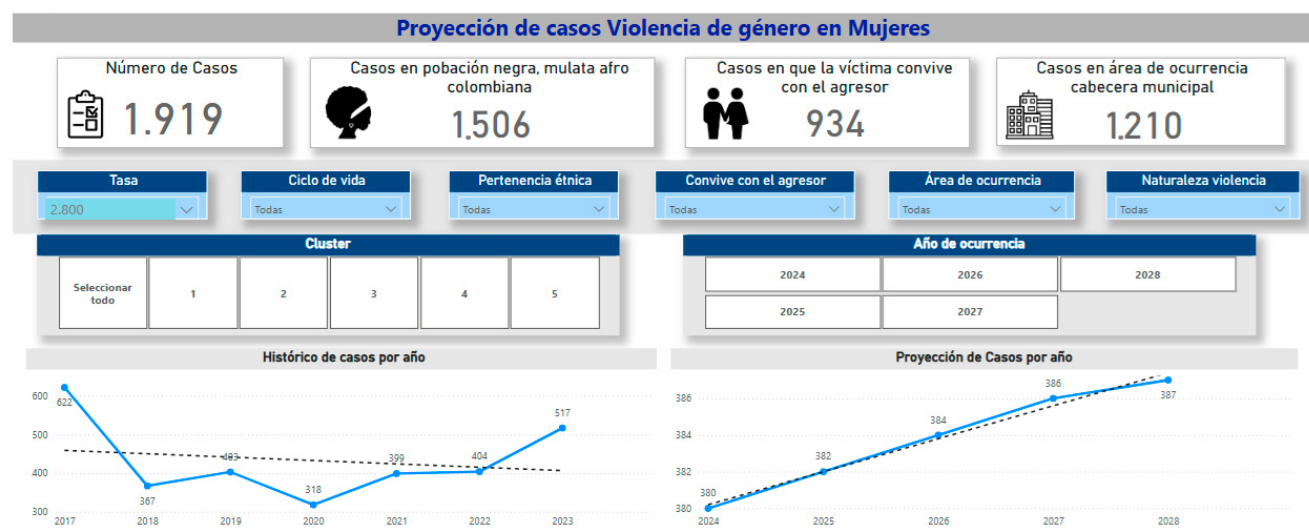


Figure 24. The interface of the section that shows projections of the gender-based violence rate in San Andrés de Tumaco for 2024 and 2028 from the decision-making tool developed in Power BI. Source: <https://linktr.ee/OrquideasTumaco> (accessed on 2 October 2024).

5. Discussion

The analysis of the SIVIGILA system has revealed significant deficiencies in the completion of reports, affecting the quality and accuracy of the collected data. The lack of automation in data entry by health institution personnel contributes to common errors, such as information duplication in incorrect fields and outdated records in reporting forms. Additionally, although the district government of San Andrés de Tumaco expressed interest in detailing the results by neighborhoods and ethnic groups, it was found that the location data in the databases did not adhere to the uniform standards. This inconsistency hindered the filtering and exploration of information, preventing precise and detailed local and sectoral analyses, although some vulnerabilities were characterized by the zone of occurrence.

In this regard, studies such as [35,36] also highlight the importance and urgent need for well-curated datasets that significantly improve the accuracy and reliability of models, especially in critical areas like health, thus preventing biased and inefficient models. These observations emphasize the immediate need to train the personnel responsible for completing event notification forms to mitigate the entry of erroneous data and improve the quality of information reported in the SIVIGILA.

Although the application of artificial neural networks did not identify representative patterns in the collected data, these tests provided valuable insights into the limitations and needs of the datasets. On the other hand, the random forest algorithm proved more effective when combined with under sampling techniques, showing a significant improvement in the accuracy and robustness of the results. Additionally, the use of clustering techniques revealed patterns in the normalization of gender-based violence, particularly highlighting the vulnerability of individuals involved in the reported cases.

These findings align with previous research. For example, Khan et al. [37] proposed using neural networks to predict patterns of domestic violence based on socioeconomic data, where key variables, such as education level, income, and family history, play a crucial role. This approach significantly contributes to the results presented here, providing a methodology that facilitates the identification and prevention of future cases of violence through the analysis of historical patterns.

Similarly, García et al. [38] explored the relationship between the use of classification techniques, such as decision trees and support vector machines, to analyze the impact of

social determinants on youth violence. This study highlights the potential to predict violent behavior, which directly correlates with the focus on gender-based violence analyzed here.

These approaches are essential for understanding the effectiveness of models, as highlighted in the literature, particularly in the study by Pérez et al. [39], who emphasize the importance of confusion matrices in the evaluation of classification models. The results obtained in this research are consistent with those of Pérez et al. regarding the utility of these metrics for interpreting performance, although clustering techniques were also incorporated, providing a more detailed view of the patterns in the data.

Moreover, the analysis of variable importance was performed using techniques like Principal Component Analysis (PCA), which helped identify the most influential factors in predictions. This approach aligns with the findings of Gómez et al. [40], who emphasize the use of PCA to identify the risk factors associated with gender-based violence. In this research, PCA also enhanced the interpretation of the results complemented by clustering, which facilitated the segmentation and more effective analysis of the data.

Finally, the predictive capacity of the model reached a margin close to 70%, highlighting the effectiveness of the methodology applied. This result aligns with the literature, as observed in the work of Zhang et al. [41], which emphasizes the importance of cross-validation techniques to improve the robustness of predictive models. However, this research stands out for its specific focus on gender-based violence, showing a significant advancement in the application of ML models in this context.

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

First, educational interventions are essential. The proposal includes developing programs from early childhood that promote gender equality, mutual respect, and peaceful conflict resolution. These programs should be integrated into the school curriculum and complemented with workshops and community campaigns. Teacher training and community awareness are essential to create a 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-based violence. Health centers and community organizations should offer these services, ensuring cultural sensitivity and accessibility. Proper care will allow for 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 should be 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 to respond to gender-based violence, training community leaders, and creating safe spaces for victims. The active participation of the community in the prevention and response to violence contributes to creating a supportive and protective environment for the affected women.

Additionally, raising awareness and preventing alcohol abuse is essential to address one of the main risk factors for gender-based violence. Educational campaigns, regulations on alcohol sales, 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, promoting legal rights and resources is crucial. Women need information about their legal rights and the resources available for their protection. Legal rights training,

the creation of legal counseling centers, and the promotion of these services, either 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 the women of 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.

The 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 the 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, 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 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.

Author Contributions: Conceptualization, O.L.G.G., C.F.T.-T., T.M.P.-Y. and Z.R.C.-V.; methodology, E.R.B.-M., E.D.C.-M., S.E.C.-B. and R.R.R.-R.; software, J.B.; validation, E.R.B.-M. and E.D.C.-M.; formal analysis, E.R.B.-M. and R.R.R.-R.; investigation, S.E.C.-B. and E.R.B.-M.; resources, R.R.R.-R.; data curation, E.R.B.-M., J.B. and E.D.C.-M.; writing—original draft preparation, E.R.B.-M., E.D.C.-M. and J.B.; writing—review and editing, S.E.C.-B., O.L.G.G., C.F.T.-T., T.M.P.-Y. and Z.R.C.-V.; visualization, J.B.; supervision, R.R.R.-R.; project administration, E.R.B.-M. and S.E.C.-B.; funding acquisition, S.E.C.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Minciencias—Ministry of Science, Technology and Innovation of Colombia, National Fund for Financing Science, Technology and Innovation Francisco José de Caldas through contract no. 112721-276-2023—Orchids program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are not publicly available due to confidentiality reasons and the need to protect the personal data of the participating population. These data are under the domain of the Mayor's Office of San Andrés de Tumaco, which provided them for the research under strict confidentiality agreements. Access to the data is restricted to ensure compliance with ethical and legal principles regarding personal data protection. However, interested parties may request the data from the corresponding author, subject to approval from the relevant authorities and adherence to the established confidentiality agreements.

Conflicts of Interest: The authors declare no conflict of interest

Appendix A

Table A1. Dictionary of Variables of Basic Data of SIVIGILA Events 356, 365, and 875.

Meaning in the Event	Variable Name	No	Meaning in the Event	Variable Name	No
Psychiatric centers	gp_psiquia	43	Event code	cod_eve	1
Victims of armed violence	gp_vic_vio	44	Date of notification	fec_not	2
Other population groups	gp_otros	45	Week of notification	semana	3
Source of notification	fuelle	46	Year of notification	año	4
Code country of residence	cod_pais_r	47	Health care provider code	cod_pre	5
Code department of residence	cod_dpto_r	48	UPGD code receiving the case	cod_sub	6
Code municipality of residence	cod_mun_r	49	First name	pri_nom_	7
Date of consultation (dd/mm/yyyy)	fec_con_	50	Second name	seg_nom_	8
Date of onset of symptoms	ini_sin_	51	First surname	pri_ape_	9
Initial case classification	tip_cas_	52	Second surname	seg_ape_	10
hospitalized patient	pac_hos_	53	Type of identification	tip_ide_	11
Date of hospitalization (dd/mm/yyyy)	fec_hos_	54	Identification number	num_ide_	12
Final condition	con_fin_	55	Age	edad_	13
Date of death (dd/mm/yyyy)	fec_def_	56	Unit of measurement of age	uni_med_	14
Unknown	ajuste_	57	Nationality code	nacionali_	15
Phone	telefono_	58	Nationality	nombre_	16
Unknown	fecha_anto_	59	Sex	nacionalidad	17
Death certificate number	cer_def_	60	Code of country of occurrence	sexo_	18
Basic cause of death	cbmte_	61	Department of occurrence	cod_pais_o	19
Unknown	uni_modif	62	Municipality of occurrence	cod_dpto_o	20
Unknown	nuni_modif	63	Area of occurrence	cod_mun_o	21
Unknown	fec_arc_xl	64	Locality of occurrence	area_	22
Name of the professional who filled out the form	nom_dil_f_	65	County	localidad_	23
Phone of the professional who filled out the form	tel_dil_f_	66	seat/Settlement/Rural dispersed	cen_pobla_	24
Date of adjustment (dd/mm/yyyy)	fec_aju_	67	Rural division/Zone	vereda_	25
NIT de UPGD	nit_upgd	68	Code neighborhood of occurrence	bar_ver_	26
Unknown	fm_fuerza	69	Residence address	dir_res_	27
Unknown	fm_unidad	70	Occupation	ocupacion_	28
Unknown	fm_grado	71	Type of health regime	tip_ss_	29
Unknown	version	72	Benefit plan administrator	cod_ase_	30
Name of the event	nom_eve	73	Ethnicity	per_etn_	31
Name of the UPGD	nom_upgd	74	Name of ethnicity	nom_grupo_	32
Name of country of origin	npais_proce	75	Stratum	estrato_	33
			Disabled person	gp_discapa	33

Table A1. Cont.

Meaning in the Event	Variable Name	No	Meaning in the Event	Variable Name	No
Name of department of origin	ndep_proce	76	Displaced persons	gp_desplaz	34
Name of municipality of origin	nmun_proce	77	Migrants	gp_migrant	35
Name of country of residence	npais_resi	78	Deprived of liberty	gp_carcela	36
Name of department of residence	ndep_resi	79	Pregnant women	gp_gestan	37
Name of municipality of residence	nmun_resi	80	Weeks of gestation	sem_ges_	38
Name of department of notification	ndep_notif	81	Homeless	gp_indigen	39
Name municipality of notification	nmun_notif	82	Child population under ICBF care	gp_pobicbf	40
Unknown	nreg	83	Community mothers	gp_mad_com	41
			Demobilized persons	gp_desmovi	42

UPGD Code: From the following example 11001123345_01 fill out: 11: First two digits identify the Department (DANE codes) 001: Next three digits added to the first two identify the Municipality 1: SIXTH digit if it is 7 identifies a UPGD of the National Police.

Table A2. Dictionary of Variables of Complementary Data of SIVIGILA Events 356, 365, and 875.

Event 356		Event 365		Event 875	
Variable Name	Meaning in the Event	Variable Name	Meaning in the Event	Variable Name	Meaning in the Event
fec_ocurr	Date of occurrence (dd/mm/yyyy)	grupo_sust	Substance group	Nature	Non-sexual violence
día_ocurrencia	Unknown	cod_sust	Product code and name	nat_viosex	Sexual violence
inten_prev	Previous attempts	clasificac	Unknown	actividad	Activity
intentos	Number of attempts	categoria	Unknown	orient_sex	Sexual orientation
estado_civ	Marital status	nom_pro	Unknown	ident_gene	Gender identity
escolarid	Schooling	tip_exp	Type of exposure	consum_spa	Consuming psychoactive substances
prob_parej	Conflicts with partner or ex-partner	lugar_expo	Place where the intoxication occurred	persona_con_jefatura_de_hogar	Head_of_household
enfermedad_cronica	Chronic painful or disabling illness	fec_exp	Date of exposure (dd/mm/yyyy)	antec	History of violence
prob_econo	Economic problems	hor_exp	Time (0 a 24)	presencia_de_alcohol_u_otra_sustancia_en_la_víctima	Alcohol victim
muerte_fam	Death of a family member	via_exp	Route of exposure	sexo_agre	Sex
esco_educ	School/Educational	fec_aspers		r_fam_vic	Relationship to victim
prob_legal	Legal problems	escolarida	Schooling	conv_agre	Cohabitates with the aggressor
suici_fm_a	Suicide of a family member or friend	afi_arp	Affiliated with A.R.L.	r_nofiliar	Non-family aggressor
maltr_fps	Physical/psychological/sexual abuse	cod_arp	Code and name of the A.R.L.	zona_conf	Violent event occurred in the armed conflict?
prob_labor	Labor problems	est_civ	Marital status	mecanismo_utilizado_para_la_agresión	Mechanism used for the aggression

Table A2. Cont.

Event 356		Event 365		Event 875	
Variable Name	Meaning in the Event	Variable Name	Meaning in the Event	Variable Name	Meaning in the Event
prob_famil	Family problems	parte_brot	Is the case part of an outbreak: YES, NO	que_cara	Burn face
prob_consu	Consuming psychoactive substances	num_cas_br	Number of cases in this outbreak	que_cuello	Burn neck
hist_famil	Family history of suicidal behavior	fec_inv_br	Date of epidemiologic investigation outbreak (dd/mm/yyyy)	que_mano	Burn hand
idea_suici	Persistent suicidal ideation	sit_ale	Alert situation	que_pie	Burn feet
plan_suici	Organized suicide plan	muest_toxi	Toxicology samples taken	que_pliegu	Burn folds
antec_tran	History of psychiatric disorder	tipo_muest	Type of samples requested	que_genita	Burn genitals
tran_depre	Depressive disorder	prueba	Name of toxicology test	que_tronco	Burn trunk
trast_personalidad	Personality disorders	fec_expres	Unknown	que_miesup	Burn upper limb
trast_bipolaridad	Bipolar disorder	result_pru	Fill in value result/Units	que_mieinf	Burn lower limb
esquizofre	Schizophrenia	hor_inv_br	Unknown	cla_gra	Burn grade
antec_v_a	History of violence or abuse	grupo_sust	Substance group	ext_que	Extent
abuso_alco	Alcohol abuse	cod_sust	Product code and name	fec_hecho	Date of event (dd/mm/yyyy)
ahorcamien	Hanging or asphyxiation	clasificac	Unknown	escenario	Scenario
arma_corto	Sharp element	categoria	Unknown	ambito_lug	Scope of violence by place of occurrence
arma_fuego	Firearm	nom_pro	Unknown	sp_its	Prophylaxis HIV
inmolacion	Immolation	tip_exp	Type of exposure	prof_hep_b	Prophylaxis Hepatitis B
lanz_vacio	Throw into the void	lugar_expo	Place where intoxication occurred	prof_otras	Other prophylaxis
lanz_vehic	Thrown into vehicle	fec_exp	Date of exposure (dd/mm/yyyy)	ac_anticon	Emergency contraception

References

1. Sardinha, L.; Maheu-Giroux, M.; Stöckl, H.; Meyer, S.R.; García-Moreno, C. Global, regional, and national prevalence estimates of physical or sexual, or both, intimate partner violence against women in 2018. *Lancet* **2022**, *399*, 803–813. [CrossRef] [PubMed]
2. Thurston, A.M.; Stöckl, H.; Ranganathan, M. Natural hazards, disasters and violence against women and girls: A global mixed-methods systematic review. *BMJ Glob. Health* **2021**, *6*, e004377. [CrossRef]
3. Hechos y Cifras: Poner fin a la Violencia Contra las Mujeres | ONU Mujeres. Available online: <https://www.unwomen.org/es/what-we-do/ending-violence-against-women/facts-and-figures#83918> (accessed on 4 July 2024).
4. Zamora-Moncayo, E.; Burgess, R.A.; Fonseca, L.; González-Gort, M.; Kakuma, R. Gender, mental health and resilience in armed conflict: Listening to life stories of internally displaced women in Colombia. *BMJ Glob. Health* **2021**, *6*, e005770. [CrossRef] [PubMed]

5. Svallfors, S. Hidden Casualties: The Links between Armed Conflict and Intimate Partner Violence in Colombia. *Politics Gend.* **2023**, *19*, 133–165. [CrossRef]
6. Boletín 8—Datos para la paz—Corte Octubre 2023. 2023. Available online: https://datos.paz.unidadvictimas.gov.co/archivos/datosPaz/boletin_datos_paz_octubre_fronteras.pdf (accessed on 25 December 2023).
7. Observatorio de Género de Nariño, Secretaría de Equidad de Género e Inclusión Social de Nariño y ONU Mujeres. Mujeres y Hombres: Brechas de Género en Nariño. Available online: <https://colombia.unwomen.org/es/biblioteca/publicaciones/2020/12/mujeres-y-hombres-brechas-de-genero-en-narino> (accessed on 1 February 2024).
8. García Restrepo, E.; Cardona, D.; Tirado Otálvaro, A.F. La violencia contra las mujeres en Colombia, un desafío para la salud pública en cuanto a su prevención, atención y eliminación. *CES Derecho* **2021**, *12*, 167–175. [CrossRef]
9. Ministerio de Salud y Protección Social. Resumen Ejecutivo Encuesta Nacion de Demografía y Salud. Available online: <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/ENDS-libro-resumen-ejecutivo-2016.pdf> (accessed on 1 February 2024).
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. Available online: https://observatoriogenero.udenar.edu.co/cifras_violeta_vi/ (accessed on 11 July 2024).
11. Sanín, J.R. Violence against Women in Politics: Latin America in an Era of Backlash. *J. Women Cult. Soc.* **2020**, *45*, 302–310. [CrossRef]
12. Giammarioli, A.M.; Longo, E.; Bucciardini, R. Gender-Based Violence is a Never to be Forgotten Social Determinant of Health: A Narrative Literature Review. In *Women's Health Problems—A Global Perspective*; IntechOpen: London, UK, 2023. [CrossRef]
13. Dawson, M.; Carrigan, M. Identifying femicide locally and globally: Understanding the utility and accessibility of sex/gender-related motives and indicators. *Curr. Sociol.* **2020**, *69*, 682–704. [CrossRef]
14. Castorena, C.M.; Abundez, I.M.; Alejo, R.; Granda-Gutiérrez, E.E.; Rendón, E.; Villegas, O. Deep Neural Network for Gender-Based Violence Detection on Twitter Messages. *Mathematics* **2021**, *9*, 807. [CrossRef]
15. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; Pardo-Quiles, D.J.; Heras-González, P.; Chatzigiannakis, I. Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques. *Appl. Sci.* **2020**, *10*, 8244. [CrossRef]
16. Karystianis, G.; Cabral, R.C.; Han, S.C.; Poon, J.; Butler, T. 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* **2021**, *3*, 602683. [CrossRef]
17. Bauer, G.R.; Mahendran, M.; Walwyn, C.; Shokoohi, M. Latent variable and clustering methods in intersectionality research: Systematic review of methods applications. *Soc. Psychiatry Psychiatr. Epidemiol.* **2022**, *57*, 221–237. [CrossRef] [PubMed]
18. Vicente, G.; Goicoa, T.; Ugarte, M.D. Bayesian inference in multivariate spatio-temporal areal models using INLA: Analysis of gender-based violence in small areas. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 1421–1440. [CrossRef]
19. Pinto-Muñoz, C.-C.; Zuñiga-Samboni, J.-A.; Ordoñez-Erazo, H.-A.; Pinto-Muñoz, C.-C.; Zuñiga-Samboni, J.-A.; Ordoñez-Erazo, H.-A. Machine Learning Applied to Gender Violence: A Systematic Mapping Study. *Rev. Fac. Ing.* **2023**, *32*, e15944. [CrossRef]
20. Devries, K.M.; Mak, J.Y.; Bacchus, L.J.; Child, J.C.; Falder, G.; Petzold, M.; Astbury, J.; Watts, C.H. Intimate partner violence and incident depressive symptoms and suicide attempts: A systematic review of longitudinal studies. *PLoS Med.* **2013**, *10*, e1001439. [CrossRef] [PubMed]
21. Lynn, E.; Doyle, A.; Keane, M.; Bennett, K.; Cousins, G. Drug Poisoning Deaths Among Women: A Scoping Review. *J. Stud. Alcohol Drugs* **2020**, *81*, 543–555. [CrossRef]
22. Bandara, P.; Page, A.; Senarathna, L.; Kidger, J.; Feder, G.; Gunnell, D.; Rajapakse, T.; Knipe, D. Domestic violence and self-poisoning in Sri Lanka. *Psychol. Med.* **2022**, *52*, 1183–1191. [CrossRef] [PubMed]
23. Urdinola, B.P. Modelo para la caracterización y clasificación de los tipos de violencia intrafamiliar desde los registros del sistema de salud. Ph.D. Thesis, University of California, Berkeley, CA, USA, 2004.
24. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
25. Oh, G.; Song, J.; Park, H.; Na, C. Evaluation of Random forest in Crime Prediction: Comparing Three-Layered Random forest and Logistic Regression. *Deviant Behav.* **2022**, *43*, 1036–1049. [CrossRef]
26. Guerrero, A.; Cárdenas, J.G.; Romero, V.; Ayma, V.H. Comparison of Classifiers Models for Prediction of Intimate Partner Violence. *Adv. Intell. Syst. Comput.* **2021**, *1289*, 469–488. [CrossRef]
27. Biró, T.S.; Nédá, Z. Gintropy: Gini Index Based Generalization of Entropy. *Entropy* **2020**, *22*, 879. [CrossRef]
28. Hossain, M.M.; Asadullah, M.; Rahaman, A.; Miah, M.S.; Hasan, M.Z.; Paul, T.; Hossain, M.A. Prediction on Domestic Violence in Bangladesh during the COVID-19 Outbreak Using Machine Learning Methods. *Appl. Syst. Innov.* **2021**, *4*, 77. [CrossRef]
29. Krstinić, D.; Braović, M.; Šerić, L.; Božić-Štulić, D. Multi-Label Classifier Performance Evaluation with Confusion Matrix. *Comput. Sci. Inf. Technol.* **2020**, *1*, 1–14. [CrossRef]
30. Theissler, A.; Thomas, M.; Burch, M.; Gerschner, F. ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices. *Knowl. Based Syst.* **2022**, *247*, 108651. [CrossRef]

31. Ileberi, E.; Sun, Y.; Wang, Z. Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost. *IEEE Access* **2021**, *9*, 165286–165294. [CrossRef]
32. Cuello Villamil, A.; Camelo Ciro, A.; Castro Molinares, S. Caracterización del intento de suicidio en un municipio de Colombia durante tres años. *Diversitas: Perspectivas en Psicología* **2023**, *19*, 55–73. [CrossRef]
33. Solarte, I.N.; Rodríguez, A.R.; Jiménez, D.D.; Díaz, I.D.P.; Martínez, E.; Hernández, B.P. Manual Para Análisis: Indicadores Para La Vigilancia De Eventos De Interés En Salud Pública. Available online: https://siteold.saludputumayo.gov.co/documentos/NORMAS/MANUAL_INDICADORES.pdf (accessed on 1 February 2024).
34. Duran, M.E.M.; García, O.E.P.; Carey, A.O.; Bonilla, H.Q.; Ortiz, M.; Forero, L.J.; Armenta, A.; Bolívar, N.G. Violencia de Género. Available online: <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/IA/INS/protocolo-violencia-genero.pdf> (accessed on 1 February 2024).
35. Pérez, M. Evaluación de la efectividad del sistema de información para la vigilancia epidemiológica en salud pública. *Rev. Investig. Cienc. Soc. Salud* **2023**, *7*, 1970. [CrossRef]
36. Gómez, A. Aplicación de técnicas de minería de datos en la salud pública: Un análisis de patrones de enfermedades infecciosas. *Rev. Iberoam. Sist. Tecnol. Gestión* **2022**, *51*, 84–98. [CrossRef]
37. Khan, F.; Niazi, M.A.; Muneer, S. Predicting domestic violence using machine learning. *J. Soc. Sci. Humanit.* **2019**, *28*, 45–60.
38. García, J.; López, M.; Pérez, A. Social determinants of youth violence: A predictive analysis using machine learning. *J. Soc. Violence Stud.* **2020**, *14*, 221–240.
39. Valero-Carreras, D.; Alcaraz, J.; Landete, M. Comparing two SVM models through different metrics based on the confusion matrix. *Comput. Oper. Res.* **2023**, *152*, 106131. [CrossRef]
40. Gómez, J. Análisis de componentes múltiples para la identificación de factores de riesgo en la violencia de género. *Comput. Sist.* **2023**, *25*, 123–135.
41. Zhang, W.; Liu, Z.; Xue, Y.; Wang, R.; Cao, X.; Li, J. An Improved Cross-Validated Adversarial Validation Method. In *Knowledge Science, Engineering and Management, Proceedings of the 16th International Conference, KSEM 2023, Guangzhou, China, 16–18 August 2023*; Springer: Cham, Switzerland, 2023; pp. 343–353.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.