Journal of Universal Computer Science, vol. 31, no. 8 (2025), 758-787
submitted: 11/6/2024, accepted: 5/2/2025, appeared: 28/7/2025

XGBoost Classifier-Based Model to Predict the Nature of Gender-Based Violence. Case Study: Santander, Colombia

Juan-Sebastián González-Sanabria
(Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia
https://orcid.org/0000-0002-1024-6077, juansebastian.gonzalez@uptc.edu.co)

Cristian Pinto
(Universidad del Cauca, Popayán, Colombia
https://orcid.org/0009-0006-0545-5700, pintocristian@unicauca.edu.co)

Jhon Zuñiga
(Universidad del Cauca, Popayán, Colombia
https://orcid.org/0009-0007-2887-5908, jhonalexsa@unicauca.edu.co)

Hugo Ordoñez
(Universidad del Cauca, Popayán, Colombia
https://orcid.org/0000-0002-3465-5617, hugoordonez@unicauca.edu.co)

Xiomara Blanco
(Universidad Internacional de La Rioja, Logroño, Spain
https://orcid.org/0000-0001-9947-748X, xiomarapatricia.blanco@unir.net)

Abstract: Gender-based violence remains a persistent social challenge in Colombia. Despite efforts to address it, statistics show a steady increase year after year. This study addresses the need for predictive solutions by introducing a Machine Learning model using XGBoost, chosen for its high performance in classification tasks with complex datasets. The model is trained on data collected from the department of Santander, Colombia, aiming to predict gender-based violence incidents based on specific socio-demographic and situational features. The motivation behind using XGBoost lies in its ability to handle diverse data types and produce accurate, interpretable results. Key influential features in the model's predictions were identified, including the context of the incidents and the relationship between victim and the perpetrator, underscoring the importance of situational as well as individual factors. The model achieved promising results, with an accuracy, precision, recall, and F1 score exceeding 84% demonstrating its potential to effectively predict and contribute to preventing gender-based violence in the region. This approach not only represents a proactive response to a critical social challenge but also offers a framework that could be applied in similar contexts at the national and international levels.

Keywords: Machine learning, Gender-based violence, Prediction, Classification model
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1
DOI: 10.3897/jucs.129515

1 Introduction

Gender-based violence is defined as "harmful acts against a person or group of people because of their gender. It has its origin in gender inequality, abuse of power, and the existence of harmful actions. The term is primarily used to underscore the fact that structural gender-based power differences place women and girls at risk of multiple forms of violence. While women and girls suffer gender-based violence disproportionately, men and boys can also be targeted" [UN Women, 2023]. This problem has been present in Colombia, as evidenced by an increase in cases in recent years. According to the Observatory of Legal Medicine, in 2021 there were 55,582 cases of Gender-Based Violence (GBV), with an increase of 19% compared to 2020 [ReliefWeb, 2022]. This increase represents a great social problem due to the repercussions it has on the victim, ranging from physical injuries, unwanted abortions, sexually transmitted infections, depression, and other anxiety disorders that have implications for other problems, such as tobacco and drug use. GBV has consequences at home and societal levels, there is evidence that children who grow up in families with violence tend to suffer behavioral and emotional disorders that lead to the inability to develop their lives in society daily [World Health Organization, 2022].

In Colombia, efforts have been made to fight gender-based violence, for instance, with the ratification of current international treaties on human rights and women's rights. Laws have also been enacted to promote gender equality and guarantee women's rights, e.g., Law 2172 of 2021, which seeks to guarantee priority access for women victims of gender-based violence to housing subsidies [Ministerio de Vivienda de Colombia, 2021]; Law 2137 of 2021, which aims to create the National Early Warning System to prevent Sexual Violence against Children and Adolescents [Congreso de Colombia, 2021]; and Law 1761 of 2015, which creates femicide as an autonomous crime [Congreso de Colombia, 2015]. While these standards provide a solid framework for advancing women's rights, challenges remain to their full implementation, as shown by data on gender gaps.

Fig. 1 shows the number of incidents of different natures (physical or psychological violence, sexual abuse, and neglect and abandonment) per month according to data collected between 2015 and 2023 [Datos Abiertos Colombia, 2023]. In February and March, a greater number of cases are reported for each of the three types of nature; on the contrary, there is a relationship of the three types of nature with the lowest number of cases in December. In addition, it is important to highlight the difference in the number of cases reported per nature; Neglect and Abandonment report the highest number, 4660 cases; followed by Physical or psychological violence, 4520 cases; and Sexual Abuse, 2149 cases in total. The relatively low number of reported cases of sexual abuse does not necessarily mean that it is the type of violence that occurs to a lesser extent; some factors inhibit the victim from reporting this type of violence, such as having a personal link with the abuser, fear of being judged or blamed, distrust of the system, among others [Sánchez-Juárez, 2023].

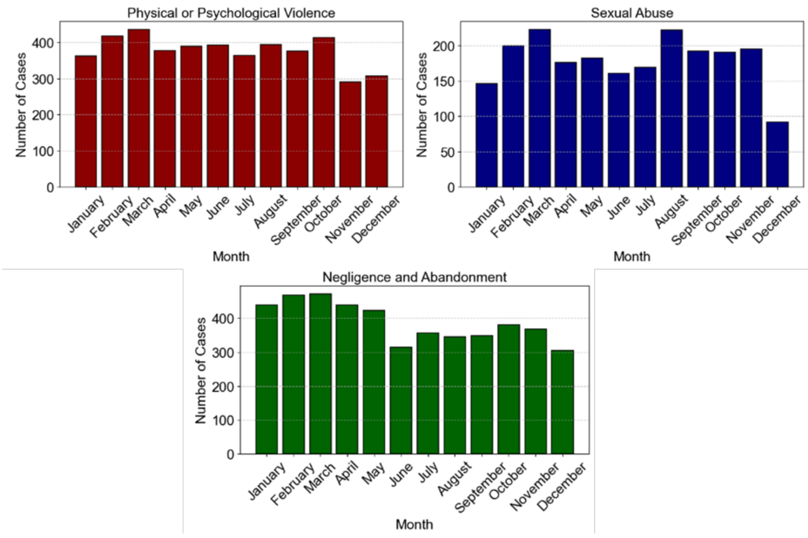

Figure 1: Number of cases per nature

Fig. 2 shows the relationship between the number of cases according to the life cycle and the victim's gender. In early childhood, the number of cases is equitable. However, in the later stages, there is evidence of a significant gap between the reported cases in women and men. The gap is most pronounced in young people, where male complaints do not exceed 500, while female complaints are approximately 2,000. Subsequently, in the older adults' stage, the number of complaints tends to balance again. This gap in the number of complaints could be due to different factors, one of them being that men have been stereotyped as a strong figure; therefore, one of the main reasons for not reporting is to feel ashamed, not being believed by the authorities or, in the case of homosexuals, fear of having to recognize their sexual identity publicly [Instituto Mexicano de la Juventud, 2017].

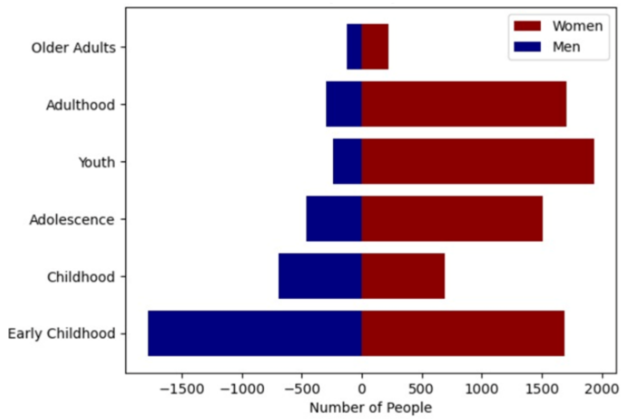

Figure 2: Distribution by Life Cycle and Gender

In the same vein, a report that highlights a significant increase in suspected cases of gender-based violence was issued in 2022, as presented in Fig. 3. The report reveals an increase in the incidence of different types of violence in that year. The category of physical violence had the highest proportion, with 47.7%, followed by sexual violence, with 26.8% of the total. In addition, 16.3% of the cases related to neglect and abandonment were identified [Instituto Nacional de Salud, 2022].

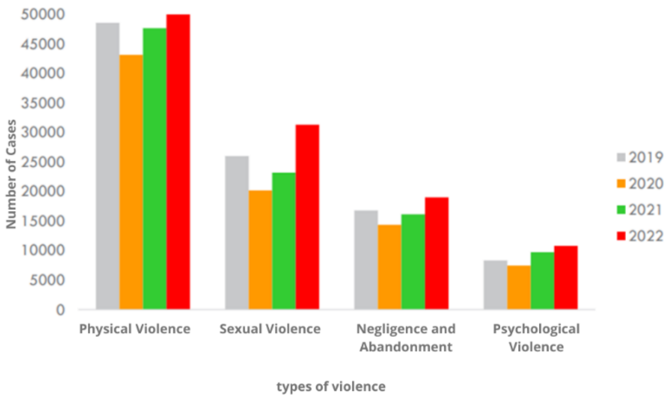

Figure 3: Suspected cases of gender-based violence in Colombia [Instituto Nacional de Salud, 2022].

By 2023, 66,742 suspected cases of gender-based and domestic violence, including attacks with chemical agents, were reported to the Public Health Surveillance System (SIVIGILA), with a weekly average of 2,781 cases. Compared to 2022, the variation in notification showed an increase of 8.3% (5117) and compared to 2021, the increase was 38.1% (18412) [Gonzalez Sarmiento, 2023].

Therefore, recognizing the importance and usefulness of applying technological tools in this context, it is necessary to implement a mechanism to offer solutions and support to government authorities, complementing the political measures that the country has proposed, and thus jointly implement effective strategies to prevent and fight GBV.

For this, using Machine Learning (ML) has proven to be a valuable tool in solving society's challenges. Although in the Colombian context technological innovation and ML have been used in various contexts, such as the prediction of theft (Ordóñez et al., 2020a) and violent homicide trends [Ordoñez-Eraso, 2020b], support for decision-making regarding the ageing of population [Ordóñez-Eraso, 2022], identification of the factors that lead people to be homeless during adolescence and adulthood [Ordóñez-Eraso, 2023], there is no record of an ML model designed to address and fight gender-based violence in Colombia that is replicable and can serve as a basis for decision-making, nor measures in this regard.

Hence, this article proposes an ML model to predict GBV actions in Santander, Colombia and contribute to the institutional strengthening of the Social Security System in Health (SGSSS by its Spanish acronym) [Ministerio de Salud de Colombia, 2023a]. The ML model is also intended to be useful for the Ministry of Equality and Equity, which seeks to protect the rights of women and girls with a gender focus by creating public policies. Given that the research is presented as a first approach from this perspective in the country in combating gender-based violence, there are limitations to generalize said model to other departments or regions due to the lack of available information or the different characteristics that promote this social phenomenon.

2 Related Studies

Gender-Based Violence (GBV) is an issue that profoundly affects public health and the well-being of society. Specifically in Colombia, the numbers presented in studies confirm the increase in GBV cases in the country compared to previous years. Therefore, it is necessary to undertake efforts to understand and combat these situations. Over the last few years, there have been multiple studies worldwide that seek to fight gender-based violence.

In the state-of-the-art review, the application of ML was identified as a crucial tool to address social problems [Matta, 2023]. This approach is used to analyze community, environmental, and social factors related to disease and mortality in groups affected by HIV or drug use. The main purpose is to understand and prevent factors that have a significant impact on the spread of HIV. Regarding Gender-based violence, one of the most promising studies [González-Prieto, 2023] showed that by applying an ML model taking as a source of information more than 40,000 complaints of gender-based violence from the official Spanish Viogen system, it achieved an improvement in the prediction of up to 25% with respect to the existing risk assessment system. Another example of the use of ML in gender-based violence [McDougal, 2021] where applying a neural network model that had an area under the curve (AUC) metric of 82%, it was determined that 7% of women who experienced marital sexual violence generally presented some of these variables as associated risk factors: sociodemographic, diet, and previous exposure to violent situations.

This model used information provided by a large national survey in India applied to 66,013 women aged 15-49. In addition, Raj [2021] using the data from the same survey, L-1 logistic regression and L-2 logistic regression models were applied to find the main risk factors for non-marital sexual violence. These models evidenced that the factors that represent a greater risk are previous exposure to violent situations, geographical location, and lack of knowledge regarding sexual and reproductive health. Investigations into complaints in the context of intimate partner violence have also been conducted. Several factors intervene in the decision of women victims of intimate partner violence to disassociate themselves from the legal process [Escobar-Linero, 2023], and, although statistical studies have been carried out, there is no evidence that ML models have been used. This is why the development of ML models was proposed to predict the disassociation of legal processes in cases of intimate partner violence. Finally, the results of the ML models were compared with works that used statistical techniques and demonstrated greater accuracy.

Continuing with the topic of gender-based violence, reference [Sower, 2021] addresses domestic violence during the COVID-19 pandemic. It indicates that such violence increased during lockdowns, as victims faced restrictions on activities that offered an escape, such as churches, schools, or shelters. This issue was even more pronounced in rural areas, where resources are limited. Therefore, it is necessary to establish a nationwide support system to assist victims of domestic violence in navigating abusive relationships or homes during a pandemic.

Another aspect to consider is the low percentage of reports regarding intimate partner violence from men. In Walker [2019], it is observed that when opportunities for anonymous reporting are provided, and language not typically associated with male-perpetrated violence against women is used, significant experiences of intimate partner violence are revealed. For example, by using the question, "Has your partner ever crossed your boundaries?" the study found that of the 258 male participants with female partners, 143 (55.4%) responded affirmatively. Additionally, instances of physical, sexual, psychological, or emotional violence—real or threatened—were detected. The findings are particularly relevant in the context of machine learning models, as factors such as the low reporting rates among men can introduce information bias into the datasets used for these models.

Predictive models prove to be highly useful in various areas of society, as shown in [Núñez, 2024], where a predictive model was used to analyze student data through four variables: motivation, attitude, knowledge, and commitment. The aim was to promote these factors, achieving a significant positive effect on the development of their competencies in sustainability. It is also used in high-impact medical applications, such as epilepsy detection. In Siddiqui [2020], demonstrates the use of a machine learning model that employs the random forest algorithm to detect seizures by analyzing brain activity records, achieving 100% seizure detection accuracy.

In Colombia, ML studies were found in other fields like the agricultural sector. Lamos-Díaz [2020], the use of models such as Random Forest, Boosting or Support Vector Machine scan be observed. It was found that the average yield of the Gradient Boosting model is higher than that of other algorithms, with the lowest values for MAE and RMSE, and the highest value for R2 to identify factors that influence crop yield. The use of ML in fields more related to gender-based violence such as crime has also been explored. Ferreira [2022] a K-nearest neighbors (KNN) model that reaches 59% completeness and more than 60% accuracy was employed. This model was developed to predict theft crimes in Bucaramanga and showed good results despite the limited information available in an intermediate city. Below, in Table 1, the main findings identified in the previously mentioned studies are presented.

Case	Main Finding
Inferred networks, machine learning, and health data [Matta, 2023]	Demonstrate the utility of visualization of results based on graphics, which identified relevant attributes in determining HIV risk. These findings provide a solid foundation for prevention efforts.
Hybrid machine learning methods for risk assessment in gender-based crime [González-Prieto, 2023]	The demonstration that the developed machine learning model significantly improved the prediction of recidivism cases in gender-based violence offenses by 25% compared to the risk assessment system used at the time of the study.
Using machine learning to explore factors associated with marital sexual violence in a	The identification of factors such as sociodemographics, previous exposure to violence, and health knowledge level as key elements in the occurrence of conjugal sexual violence events in

Case	Main Finding
cross-sectional study from India [McDougal, 2021]	India. Considering these variables provides a basis for preventive actions from multiple angles.
Machine learning analysis of non-marital sexual violence in India [Raj, 2021]	This study reveals a finding closely linked to the previous one, as associated elements for cases of non-conjugal sexual violence in India were identified, such as previous exposure to violence, geographical location, sexual behavior, and limited knowledge of sexual and reproductive health. Comparing with the previous study, highly influential factors were observed in both marital and non-marital sexual violence in India.
Using machine learning-based systems to help predict disengagement from the legal proceedings by women victims of intimate partner violence in Spain [Escobar-Linero, 2023]	The implementation of a machine learning (ML) model to identify factors influencing the withdrawal of charges after reports of intimate partner violence against women, which generated notable improvements. Compared to a previous approach based on statistical techniques, the model demonstrated improved accuracy in all cases. Furthermore, by adding a new variable, "plans to withdraw," the improvement in accuracy over the previous method was 7.5%.
Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia [Lamos-Díaz, 2020]	The utility of machine learning (ML) models as support tools in various contexts, such as cocoa farming in Colombia. Thanks to their application, factors impacting crop performance were identified, facilitating targeted actions towards improvement and productivity.
Predicting Crime in Intermediate Cities: A Machine Learning Model [Ferreira, 2022]	The application of a machine learning (ML) model to address violence in challenging contexts, such as intermediate cities with limited data availability. From this research, it was evident that with the available data, it was possible to implement a functional model with metrics that reached or exceeded 59%.

Table 1: Main Findings of Studies

Due to the results of using ML models in different areas, the use of an ML model that serves as a tool to fight gender-based violence and support the strategies to prevent and address current gender-based violence is proposed. These strategies include Line 123 Women, foster homes, psychological and legal care. In summary, the use of ML models in this area seeks to strengthen the institutional system of prevention, care, assistance, protection, prosecution, and comprehensive reparation for women victims of violence [Alcaldía de Medellín, 2023].

In addition to developing an innovative tool to address gender-based violence in Colombia, this research project aims to mark the first steps in the application of technologies like ML to combat this issue in the country. Its implementation will serve as support for the challenges currently faced in the fight against gender-based violence, considering that more countries than ever have laws against domestic violence, sexual assaults, and other forms of violence. However, challenges persist in the enforcement of these laws, resulting in limited protection and access to justice for women and girls. Additionally, not enough is being done to prevent violence, and when it occurs, it often goes unpunished [UN Women, 2023]. Faced with the challenge of insufficient prevention, the use of ML allows for the early detection of new cases and also the identification of patterns or factors associated with gender-based violence.

Nevertheless, it is crucial to consider the inherent complexities, such as the difficulty of generalizing the model to different departments or regions of Colombia due to varied socio-economic conditions, cultural norms, data collection capabilities, and other variables that directly influence the phenomenon of gender-based violence. As highlighted in [Myhill, 2017], the measurement of domestic violence is profoundly shaped by these contextual factors, which influence both the accuracy of the data collected and the applicability of findings. Variations in definitions, reporting practices, and societal attitudes can lead to discrepancies that complicate the development of reliable predictive models, emphasizing the need for localized and context-aware approaches.

Additionally, social and cultural barriers often hinder reporting and access to justice for victims, especially when gender stereotypes shape perceptions of violence. For instance, the study in Walker [2019], highlights how men experiencing intimate partner violence by women may face biased attitudes from family, friends, and authorities, potentially leading to revictimization and insufficient institutional response. This underscores the importance of designing inclusive and unbiased predictive tools that account for diverse victim experiences, challenging stereotypes and enabling equitable access to justice for all affected individuals.

Furthermore, challenges related to false allegations of violence must also be addressed [Rumney, 2017]. False accusations can affect public perception and judicial responses, diverting resources from real cases and negatively impacting both victims and the legal system. While the proposed ML model aims to mitigate these risks by improving predictions of gender-based violence, it is essential to ensure that justice systems differentiate between false and legitimate allegations, thus supporting an adequate and effective response to real victims.

These interconnected factors contextual, cultural, and systemic highlight the necessity of a multifaceted approach in the development of predictive models. This includes adapting to changes over time in the variables associated with gender-based violence, as evolving circumstances may affect the performance and relevance of such tools.

3 Methodology

Fig. 4 presents the model designed to predict gender-based violence actions. It was developed after evaluating various algorithms using the selected metrics. This process was divided into 6 stages: data set, data processing, algorithm training, performance evaluation metrics, final model, and web platform. Thus, a tool capable of processing a series of input data and making predictions was obtained.

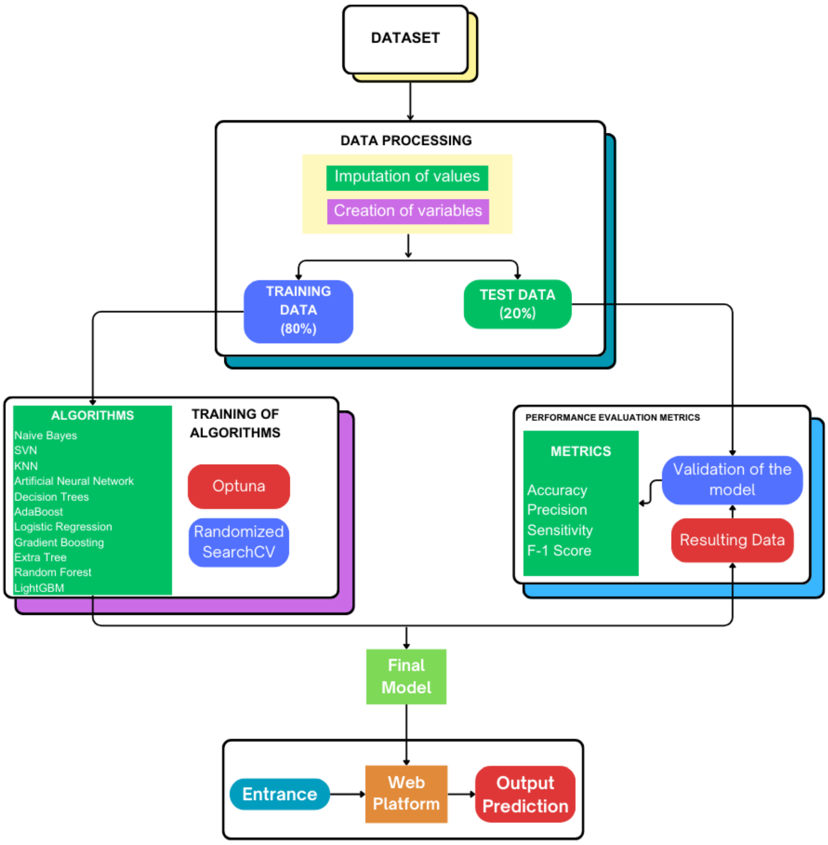

Figure 4: Proposed Model

The selected dataset contains a number of significant records with variables that have a high influence to catalog a case of gender-based violence. The information is concentrated in the Santander department and has 32 columns and 12,161 records that cover from January 2015 to March 2023. It is an open access data set available on the datos.gov.co webpage. The Ministry of Health and Environment provided it and contains cases attended by the Service Providing Institutions (IPS by its Spanish acronym) at the corresponding municipality on gender-based violence according to its nature (sexual abuse, neglect, and abandonment, and physical or psychological violence), kinship with the aggressor, life course, age, and gender.

It is worth mentioning that this data is available openly and free of charge on the platform belonging to the Colombian state and does not expose sensitive information about the victims or perpetrators of reported gender-based violence cases, such as names, addresses, email addresses, or phone numbers.

After identifying the dataset to work with, the exploratory analysis was carried out. The data are processed by taking the initial dataset and applying the necessary modifications to obtain an adequate dataset to train the ML model.

Some of the techniques used to conduct the Exploratory Data Analysis are the Imputation of null values and unwanted values in the 'age_AGR' column through the distribution calculated from existing values in it, and mapping or assignment of values to the 'nature' column from the values of the 'DEF_nature'column due to missing values. This process was carried out thanks to the direct relationship of the values of 'DEF_nature' with the variable 'nature'. Null, duplicate and repeated records were deleted to ensure the integrity of the information. Irrelevant columns were also excluded. The categories were redefined, specifically in the 'nature' column, to reduce its complexity, thus grouping the different possible values into 3 categories. New variables like 'domestic_violence', 'minor_victim', 'minor_perpetrator' and 'quarter' were created from 'vict_relationship', 'age_group', 'aggre_age' and 'Month', respectively. A dataset with 26 columns and 11,329 records was obtained at the end of the data processing (Table 2).

Variable	**Description**
Department	Name of the department of origin of the victim
Municipality	Name of the municipality of origin of the victim
Week	Epidemiological week in which the event occurred according to the current calendar
Year	Year corresponding to the epidemiological week
Age Group	Age range to which the victim belongs
Life_cycle	Classification of the patient's age according to the Ministry of Health and Social Protection
Gender	Physiological characteristics of the patient
Area	Alphanumeric value corresponding to the area code where the event occurred
Type_of_health_insurance	Type of Health Regime
Patient_hospitalized	Whether the patient was hospitalized or not
Final_condition	Final condition of the patient
Nature	Nature of the case of violence
Activity	Activity performed by the victim
Aggresor_age	Age of the perpetrator
Aggresor_gender	Gender of the perpetrator
Victim_substance	Presence of alcohol and other substances in the victim
Scenario	Scenario where the case occurred
Event name	Name of the event
Upgd_name	Name of the primary data-generating unit
Month	Month in which the acts of violence occurred
Victim_relationship	Indicates the family or personal relationship of the perpetrator to the victim
Commune	Name of the commune

Variable	Description
Quarter	Quarter of the year in which the incident occurred, possible values: 1, 2, 3, and 4
Domestic_violence	Indicates whether there is domestic violence or not 1. Yes 2. No
Minor_victim	Indicates if the victim is a minor 1. Yes 2. No
Minor_perpetrator	Indicates if the perpetrator is a minor 1. Yes 2. No

Table 2: Description of dataset

This refined dataset is used to evaluate the tested models, which are classified according to the metrics obtained. The metrics set to evaluate them are presented in the Metrics paragraph.

The variable to predict is "Nature". It was chosen as the target variable because it allows classifying a case of the dataset among the following values: 1. Physical or psychological violence; 2. Sexual Abuse; and 3. Neglect and abandonment. Each type of nature is defined as follows: Physical violence consists of causing harm to the partner through any type of physical force, such as beating or kicks [UN Women, 2023]; Psychological violence consists of intimidation or threats that, without exercising direct physical action, affect the victim's psyche and emotions and may eventually affect their physical state [Gestarsalud, 2020]; Sexual abuse involves forcing the partner to participate in the sexual act without their consent [UN Women, 2023] it has consequences ranging from unwanted pregnancies, sexually transmitted diseases, and impact on the mental health of the victim [World Health Organization, 2022]; Negligence and abandonment refers to the lack of protection and care by someone who is responsible for a person, whether child or adolescent and their needs, whether physical, psychological, educational or health, thus affecting the integrity and personal development of the victim [Arbeláez Arbeláez, 2021]. The distribution of this and the other classes in the dataset can be seen below in Table 3.

Variable	Distribution
Departament	Santander: 11193; Others: 136
Municipality	Bucaramanga: 11071; Others: 258
Week	1: 19; 2: 247; 3: 243; 4: 234; 5: 247; 6: 270; 7: 243; 8: 279; 9: 248; 10: 211; 11: 218; 12: 242; 13: 234; 14: 200; 15: 159; 16: 199; 17: 211; 18: 229; 19: 217; 20: 228; 21: 216; 22: 243; 23: 219; 24: 193; 25: 159; 26: 194; 27: 315; 28: 188; 29: 217; 30: 190; 31: 140; 32: 238; 33: 218; 34: 234; 35: 203; 36: 195; 37: 174; 38: 259; 39: 284; 40: 208; 41: 201; 42: 249; 43: 228; 44: 193; 45: 211; 46: 242; 47: 211; 48: 242; 49: 187; 50: 217; 51: 171; 52: 166; 53: 140
Year	2015: 1795; 2016: 2094; 2017: 1090; 2018: 1089; 2019: 1187; 2020: 792; 2021: 1088; 2022: 1784; 2023: 410

Variable	Distribution
Age_group	0 to 6: 3472; 7 to 11: 1386; 12 to 17: 1966; 18 to 28: 2168; 29 to 59: 1997; 60 and over: 340
Life_cycle	Adolescence: 1966; Childhood: 1386; Youth: 2168; Early childhood: 3472; Adulthood: 1997; Elderly: 340
Gender	Female: 7744; Male: 3585
Area	Municipal head: 11064; Populated center: 74; Rural dispersed: 191
Type_of_health_insurance	Contributory: 4868; Special: 138; Exception: 335; Undetermined: 124; Uninsured: 1; Non-affiliated: 723; Subsidized: 5140
Patient_hospitalized	No: 3128; Yes: 8201
Final_condition	Dead: 11303; Doesn't know – No answer: 22; Alive: 4
Nature	1. Physical or psychological violence: 4521; 2. Sexual abuse: 2149; 3. Negligence: 4659
Activity	Student: 3083; Public force: 1; Civic leaders: 8; Teachers: 5; None: 2374; Other: 5085; Person dedicated to household care: 506; Person in prostitution situation: 11; Recycler: 5; Public servant: 8; Domestic worker: 243
Aggresor_age	3: 10; 4: 92; 5: 136; 6: 103; 7: 99; 9: 64; 10: 602; 11: 610; 12: 32; 13: 91; 14: 105; 15: 261; 16: 569; 17: 243; 18: 349; 19: 39; 20: 55; 21: 444; 22: 429; 23: 514; 24: 498; 25: 140; 26: 151; 27: 385; 28: 464; 29: 305; 30: 602; 31: 286; 32: 402; 33: 302; 34: 263; 35: 370; 36: 289; 37: 216; 38: 210; 39: 18; 40: 10; 41: 323; 42: 473; 43: 105; 44: 151; 45: 142; 46: 205; 47: 68; 48: 4; 49: 21; 50: 140; 51: 99; 52: 70; 53: 32; 54: 82; 55: 99; 56: 25; 57: 46; 58: 44; 59: 64; 60: 20; 61: 45; 62: 15; 63: 16; 64: 29; 65: 15; 66: 9; 67: 57; 68: 72; 69: 2; 70: 6; 71: 4; 72: 68; 73: 6; 74: 4; 75: 15; 76: 54; 77: 10; 79: 6; 80: 3; 81: 11; 82: 46; 83: 99; 88: 5; 95: 4; 99: 7
Aggresor_gender	F: 4874; M: 6455
Victim_substance	No: 11114; Yes: 215
Scenario	Sports and recreational area: 27; Commerce and service areas (Store, shopping center): 121; Educational institution: 297; Health institution: 109; Recreation places with alcohol sales: 55; Workplace: 70; Other open spaces (forests, pastures, etc): 968; Public road: 2074; Housing: 7608

Variable	Distribution
Event_name	Public health surveillance of gender violence: 4979; Public health surveillance of intrafamily gender violence: 6350
Upgd_name	Clinica Chicamocha SA: 748; Clinica Materno Infantil San Luis SA: 725; Fundacion Oftalmologica de Sder Foscal: 596; Hospital Local del Norte: 2075; Hospital Universitario de Santander: 1410; Los Comuneros Hospital Universitario de Bucaramanga: 483; Others: 3809; Sede Gonzalez Valencia: 244; Serviclinicos Dromedica SA: 373; Uimist: 632; Universidad Industrial de Santander- UIS: 234
Month	January: 948; 02. February: 1087; 03. March: 1130; 04. April: 993; 05. May: 994; 06. June: 868; 07. July: 890; 08. August: 962; 09. September: 916; 10. October: 985; 11. November: 853; 12. December: 703
Victim_relationship	Grandparent: 70; Common-law partner: 135; Husband: 215; Ex-partner: 700; Family member: 1045; Brother: 62; Mother: 3929; Boyfriend/Girlfriend: 125; Others: 2705; Father: 920; Partner: 1423
Commune	North: 1784; 02. Northeast: 712; 03. San Francisco: 1034; 04. Western: 877; 05. Garcia Rovira: 665; 10. Provenza: 619; 11. South: 496; 13. Eastern: 684; 14. Morrorico: 514; Others: 3944
Quarter	1: 3412; 2: 2810; 3: 2639; 4: 2468
Domestic_violence	No: 8367; Yes: 2962
Minor_victim	No: 5438; Yes: 5891
Minor_perpetrator	No: 906; Yes: 10423

Table 3: Data Distribution

Having defined the distribution of the dataset, Table 4 shows an example of an input dataset for the implemented model.

Input Data	Value
department	Others
week	20
year	2020
age_group	0 a 6
gender	Male
area	MUNICIPAL DISTRICT
neighborhood	10. PROVENZA
social_security_type	non-affiliated
hospitalized patient	1

Input Data	**Value**
final_condition	1
activity	13
age_agre	40
gender_agre	F
relationship_vict	Mother
victim_substances	1
scenario	1
nom_upgd	HOSPITAL LOCAL DEL NORTE
municipality	Others
month	05. May
quarter	2
life_cycle	Primera infancia
domestic_violence	1
minor_victim	1
minor_aggressor	0
nom_eve	VIGILANCIA EN SALUD PÚBLICA DE LA VIOLENCIA DE GÉNERO E INTRAFAMILIAR
Output Data	**Value**
Nature	1

Table 4: Example of Input and Output Data with the Implemented Model

After data processing, 80% of the original records are taken from the dataset for training, and the remaining 20% to test the models. Each model has many parameters that can be configured, but at first, each algorithm was evaluated in its base state, without configuring hyperparameters, to decide which ones present the best results when evaluating the data set in question. The tested algorithms are presented in Table 5.

Model	**Description**
Naive Bayes	It is widely used in classification. It is called "Naive" because it assumes that all prediction variables are independent of each other. It is fast and scalable, calculates conditional probabilities for combinations of attributes and the target. From the training data, it establishes independent probabilities that indicate the likelihood of each target class, given a set of input variable values
Support Vector Machine	It works by correlating data to a large feature space so that data points can be categorized, even if the data cannot be linearly separated otherwise. A separator between the categories is detected and the data is transformed so that the separator can be extracted as a hyperplane. Thereafter, the characteristics of the new data can be used to predict the group to which the new record belongs
K-nearest	Also known as KNN or k-NN. It is a non-parametric supervised learning classifier, which uses proximity to make classifications or predictions about the clustering of an individual data point

Model	Description
Neural Network	It is a simplified model that mimics information processing in the human brain. It is composed of interconnected units in layers: one input, one or more hidden, and one output. These units communicate through connections with different weights. Data enters the input layer and propagates through the layers until a result is obtained at the output layer
Decision Trees	It is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the characteristics of the data. A tree can be seen as a constant approximation by parts
AdaBoost	It is a meta-estimator that starts by adjusting a classifier in the original dataset and then adjusts additional copies of the classifier in the same dataset, but where the weights of incorrectly classified instances are adjusted in such a way that subsequent classifiers focus more on difficult cases
Multinomial Logistic Regression	It is useful to classify subjects based on the values of a set of predictor variables. This type of regression is like logistic regression, but more general, since the dependent variable is not restricted to two categories
Gradient Boosting	It builds an additive model at a forward stage; this allows the optimization of arbitrary differentiable loss functions. In Regression trees of each stage fit into the negative gradient of the loss function, for example, binary or multi-class logarithmic loss. Binary classification is a special case in which only a single induced regression tree is found
Extra Tree	It implements a meta-estimator that fits a few Random Decision Trees (also known as Extra Trees) on various subsamples of the dataset and uses the average to improve predictive accuracy and control overfitting
Random Forest	It is a meta-estimator that fits a series of classifier decision trees in various subsamples of the dataset and uses an average that improves predictive accuracy and controls overfitting. The size of the subsample is controlled with the if parameter (default), otherwise the entire dataset is used to build each tree
LightGBM	It uses histogram-based algorithms, which group continuous feature values (attributes) into discrete containers. These speeds up training and reduces memory usage
XGBoost	It is based on assembled learning, that is, it consists of multiple models that allow it to have better accuracy in predictions. It tries to correct the errors made by the models by adjusting the successive models and adding weights to them. This approach ensures effectiveness in improving the model

Table 5: Tested Algorithms. Adapted by [IBM, 2021]; [Scikit Learn, 2022]

Models were evaluated according to the metrics presented in Table 6.

Metric	Description	Formula
Accuracy	The multi-class classification calculates the accuracy of the subset: the set of predicted labels for a sample must exactly match the corresponding set of labels in y_true	$\boldsymbol{Accuracy} = cp\ /\ tp$ $\boldsymbol{cp}$ number of correct predictions $\boldsymbol{tp}$ total number of predictions
Precision	It is intuitively understood as the ability of the classifier not to label as positive a sample that is negative	$\boldsymbol{Precision} = tp\ /\ (tp\ /\ fp)$ tp number of true positives fp number of false positives
Sensitivity	It is intuitively understood as the ability of the classifier to find all positive samples	$\boldsymbol{Recall} = tp\ /\ (tp\ +\ fn)$ tp number of true positives fn number of false negatives
F1	It can be interpreted as a harmonic means of precision and recall. An F1 score reaches its best value at 1 and its worst at 0. The relative contribution of precision and recall to the F1 score is equal	$\boldsymbol{F1} = 2 * (\boldsymbol{ps} * \boldsymbol{rs})\ /\ (ps + \boldsymbol{rs})$ $\boldsymbol{ps}$ Accuracy $\boldsymbol{rs}$ Recall

Table 6: Performance Evaluation Metrics [IBM, 2021]

After evaluating the algorithms, Table 7 shows the results obtained by each metric. As can be seen in the figures, the most outstanding algorithms were LightGBM and XGBoost. Hence, it is decided to continue only with these two algorithms to the next stage, where a search for hyperparameters is done for each one to obtain the configuration that fits best and choose the final model.

Models / Metrics	Accuracy	Precision	Sensitivity	F1-score
SVM	0.816	0.805	0.789	0.796
Random Forest	0.850	0.840	0.840	0.839
Gradient Boosting	0.836	0.827	0.819	0.822
LightGBM	0.850	0.841	0.840	0.840
XGBoost	0.850	0.842	0.840	0.841
Multinomial Logistic Regression	0.835	0.827	0.822	0.825
KNN	0.793	0.774	0.766	0.769
GaussianNB	0.658	0.686	0.682	0.624
Decision Trees	0.783	0.771	0.769	0.770
AdaBoost	0.825	0.817	0.812	0.814
Neural networks	0.810	0.80	0.799	0.799
Extra Trees	0.847	0.837	0.836	0.836

Table 7: Metrics of the models

4 Results

The search for hyperparameters for each algorithm was carried out in different ways, with tools such as Optuna and RandomizedSearchCV. On the one hand, Optuna is an automatic hyperparameter optimization software framework designed specifically for machine learning [Pinto-Muñoz, 2023]. On the other hand, random search through RandomizedSearchCV is based on extracting a fixed number of parameter configurations from the specified distributions. It basically refers to randomly testing different combinations of parameters and evaluating them to score the best combination of hyperparameters.

As mentioned earlier, the decision was made to continue the hyperparameter tuning process with only two algorithms: LightGBM and XGBoost. Hyperparameter tuning techniques were applied by trying various configurations. Tables 8 and 9 present some notable results with different configurations for the LightGBM and XGBoost algorithms, respectively. Additionally, for the hyperparameter configuration test for the LightGBM and XGBoost models, the calculation of the Macro F1 and Micro F1 metrics was implemented. Macro F1 calculates the F1-Score for each class and then obtains the unweighted average, while Micro F1 calculates the F1-Score considering the total of true positives, false positives, and false negatives across all classes. The interpretation of these two new metrics refers to the model's behavior with respect to each class and its distribution. The values obtained in these metrics show that the model performs well despite the presence of data imbalance, particularly in class 1, as will be seen later.

Hyperparameters	**Average Accuracy**	**Average Precision**	**Average Sensitivity**	**Average F1 Score**	**Macro F1**	**Micro F1**
'num_leaves': 40, ‘n_estimators': 100, 'min_child_samples': 10, 'max_depth': 3, 'learning_rate': 0.2	0.855	0.845	0.846	0.845	0.845	0.855
'learning_rate': 0.051417608292967786, 'num_leaves': 26, 'max_depth': 7, ‘min_child_samples': 62, 'subsample': 0.7461116830392539, 'colsample_bytree': 0.5904387690644886, 'reg_alpha': 0.09233912822725567, 'reg_lambda': 0.042356048609430104, 'n_estimators': 126,	0.852	0.842	0.842	0.841	0.841	0.852

Hyperparameters	Average Accuracy	Average Precision	Average Sensitivity	Average F1 Score	Macro F1	Micro F1
'scale_pos_weight': 0.5842230278735413						
'learning_rate': 0.01802606582997746, 'reg_alpha': 6.779674157856769e-08, 'reg_lambda': 0.053596159180180146, 'num_leaves': 23, 'max_depth': 7, 'min_child_samples': 4, 'subsample': 0.790561196083304, 'scale_pos_weight': 4.855801771130356, 'colsample_bytree': 0.972976079256426, 'n_estimators': 554	0.851	0.842	0.842	0.841	0.841	0.851

Table 8: Metrics obtained by the hyperparameter sets for the LightGBM model

Hyperparameters	Average Accuracy	Average Precision	Average Sensitivity	Average F1 Score	Macro F1	Micro F1
'objective': 'multi:softmax', 'num_class': 3, 'importance_type': 'gain', 'learning_rate': 0.030958241241393895, 'max_depth': 7, 'colsample_bytree': 0.5330554532179588, 'n_estimators': 568, 'reg_alpha': 4.1084086686936143e-08, 'reg_lambda': 0.0002428080231861165, 'min_child_weight': 1,	0.856	0.848	0.846	0.846	0.846	0.856

Hyperparameters	**Average Accuracy**	**Average Precision**	**Average Sensitivity**	**Average F1 Score**	**Macro F1**	**Micro F1**
'gamma': 0.23340955083195258, 'subsample': 0.8458562164387711						
'learning_rate': 0.044978141622494895, 'max_depth': 5, 'colsample_bytree': 0.6612006182149313, 'n_estimators': 740, 'reg_alpha': 3.462106562185482e-05, 'reg_lambda': 2.5206355109283938e-08, 'min_child_weight': 3, 'gamma': 0.000219351881425621, 'subsample': 0.704807956180106	0.853	0.845	0.842	0.843	0.843	0.853
'learning_rate': 0.016951568933621472, 'max_depth': 7, 'colsample_bytree': 0.5954988480447891, 'n_estimators': 972, 'reg_alpha': 1.8721084537933326e-08, 'reg_lambda': 0.0002344955542435387, 'min_child_weight': 5, 'gamma': 8.251699970197526e-05, 'subsample': 0.8577198024143824	0.852	0.843	0.842	0.842	0.842	0.852

Table 9: Metrics obtained by the hyperparameter sets for the XGBoost model

As presented in Table 10, the model with optimized hyperparameters that yielded the best result out of the evaluated metrics was XGBoost, said metrics were estimated by cross-validation to obtain a more robust value.

Metrics	XGBoost	LightGBM
Average Accuracy	0.856	0.855
Average Precision	0.848	0.845
Average Sensitivity	0.846	0.846
Average F1 Score	0.846	0.845
Macro F1	0.846	0.845
Micro F1	0.856	0.855

Table 10: Better metrics obtained by XGBoost and LightGBM models

Fig. 5 shows the combination of hyperparameters that gave the best results. It should be noted that this combination was achieved using Optuna. The hyperparameters configured for the XGBoost model are as follows. Regarding the speed and stability of learning, learning_rate controls the contribution of each tree, while subsample determines the proportion of instances considered in the construction of each tree. The complexity of the model is managed with max_depth, which indicates the maximum depth of the trees, and colsample_bytree, which sets the proportion of features considered in each tree. Regularization terms such as reg_alpha and reg_lambda penalize large coefficients and absolute sums of sheets, respectively, thus helping to avoid overfitting. In addition, min_child_weight, gamma, and n_estimators also affect model complexity and training speed. Together, these XGBoost settings enable controlling how the model learns (learning rate), how complex it is (depth and number of trees), and how it avoids overfitting (regularization). In addition, XGBoost performs well on datasets with a high number of records, allowing it to identify patterns and generalize better, also the size of the dataset helps mitigate overfitting in XGBoost models due to its internal regularization and ability to dynamically adjust the depth of trees avoiding excessive data noise. After evaluating the models with multiple hyperparameter configurations, the best combinations are collected and the comparison between models is made to decide the final proposed model and its hyperparameter configuration.

```
pipeline_estimador = Pipeline([
    ("procesado_variables", pipeline_procesado),
    ("estimador", xgb.XGBClassifier(
        objective= 'multi:softmax',
        num_class = 3,
        importance_type='gain',
        learning_rate=0.030958241241393895,
        max_depth=7,
        colsample_bytree=0.5330554532179588,
        n_estimators=568,
        reg_alpha=4.1084086686936143e-08,
        reg_lambda=0.00024280802318611656,
        min_child_weight=1,
        gamma=0.23340955083195258,
        subsample=0.8458562164387711
    ))
])
```

Figure 5: Hyperparameters configuration of the proposed final model

Once the model that yielded the highest performance was chosen according to the evaluated metrics, a series of graphs were created to show the predictions made by the model versus the real values according to the different classes that the target variable "nature" has. Fig. 6 shows that the model presents quite accurate predictions for each class. Specifically, it presents an accuracy of 81% for class 0 (Physical or psychological violence), 79% for class 1 (Sexual abuse), and 93% for class 2 (Neglect and abandonment). These values evidence the remarkable performance of the model to predict, especially in cases where nature has a value of 2 (Negligence and abandonment).

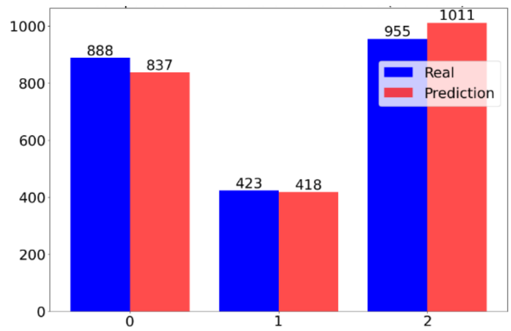

Figure 6: Comparison of Real Data and Predictions

Fig. 7 shows that the model erroneously predicted 51 cases of Physical or Psychological Violence, and 56 cases of Neglect and Abandonment. It made less mistakes regarding Sexual Abuse, counting only 5 cases. However, the figures mentioned do not imply that the Sexual Abuse nature is the one with the highest percentage of success, since it has 79%, being the lowest among the natures. This low number of erroneous cases can be attributed to the inequality of data in the dataset, which can cause a bias in the training process; since the model tends to pay more attention to the classes with greater representation, it makes fewer errors in these categories during training. This can make it more difficult for the model to set an effective decision limit for the minority class, resulting in lower prediction accuracy.

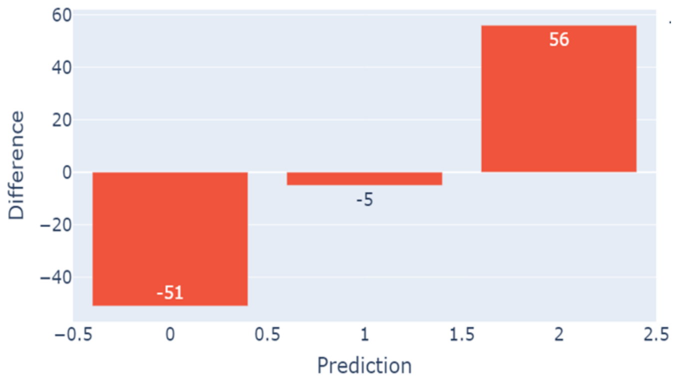

Figure 7: Difference between Real Data and Prediction

The previously noted imbalance between classes 0 and 2 with respect to class 1 affects the performance of the model. But it should be noted that due to the high metrics obtained, it was decided to maintain the simplicity of the same and avoid techniques that increase its complexity and understanding.

In the Fig. 8, the precision and recall curves by class are shown, where it can be observed that, following the trend shown in the previous graphs, classes 0 and 2 demonstrate good performance, as they maintain high precision even as recall increases, indicating a good balance. On the other hand, the curve corresponding to class 1 shows variations in the initial recall values, and as recall increases, precision decreases at a faster rate compared to the other classes. This indicates that this class is more challenging to classify correctly, and this lower performance may be due to class imbalance.

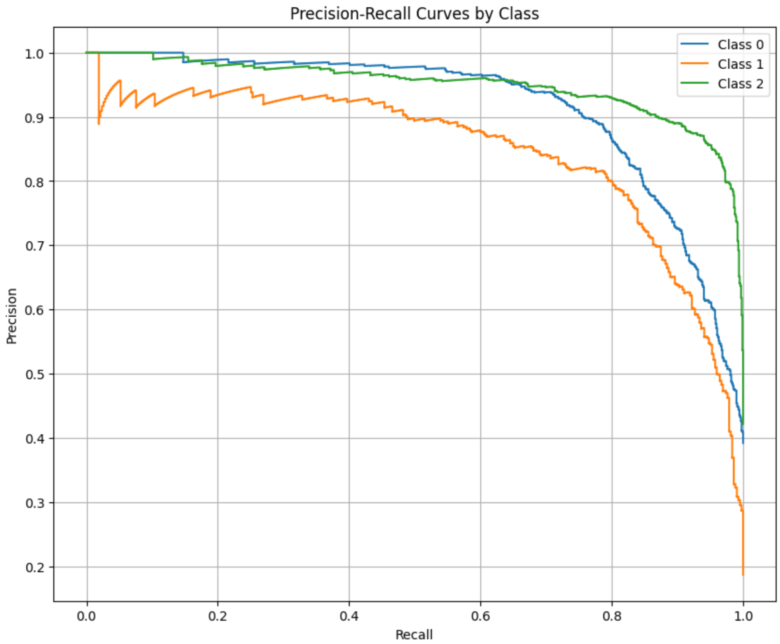

Figure 8: Precision-Recall Curves by Class

To continue with the analysis of the proposed model, Fig. 9 shows each feature used for model training and their respective importance. It is obtained thanks to the feature_importance property, which calculates the importance of each variable according to the criterion configured for the model. In this case, in blue, the configured criterion was "gain", where the features that have the most impact on reducing the error of the model are considered more important; and in green, the criterion used was "weight", which is based on the frequency with which each feature is used to divide the data in the model trees. Observing the resulting importance, it is possible to conclude that the variable Event_name has a significant importance in the accuracy of the model but is used less frequently in the construction of trees, contrary to the variables

Patient_hospitalized and Department, which are used very frequently but do not make a very large contribution to accuracy. Moreover, there are variables such as Scenario or victim_relationship that are frequently used in the construction of trees and have a considerable impact on the accuracy of the model. These variables are very relevant.

Considering the results obtained in the performance evaluation metrics, the final model obtained using the XGBoost algorithm stands out for the scores achieved, ensuring over 84% in the metrics it was evaluated on, including Accuracy, Precision, Sensitivity, and F1 Score. However, it is important to recognize that the interpretation of these metrics should also take into account a series of contextual factors that greatly influence their meaning and practical application. Aspects such as the size and quality of the dataset, the relevance of the collected information, the context of the problem being addressed, geographical or sociocultural peculiarities, can impact the results obtained. Additionally, the approach used to build the model, the selection of relevant features, the implementation of data preprocessing techniques, and hyperparameter optimization play a crucial role in the metrics and predictive capability of an ML model.

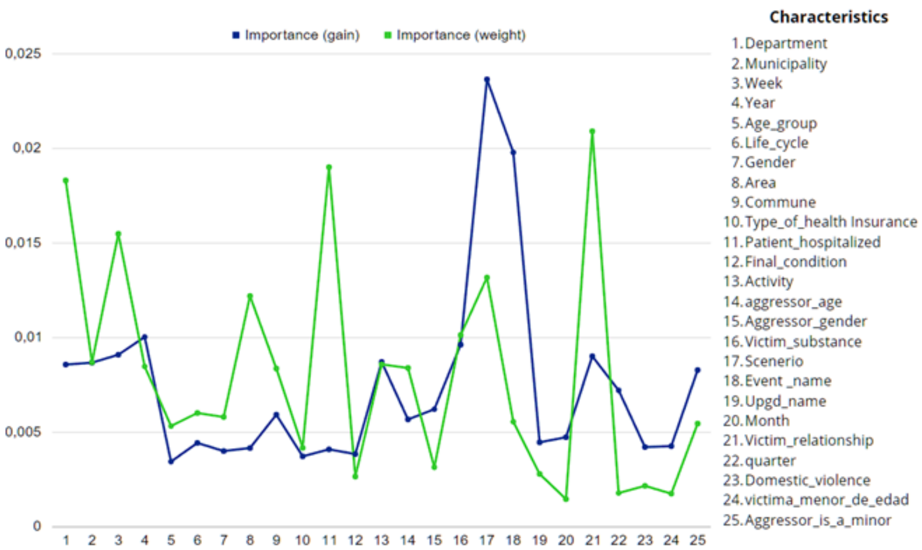

Figure 9: Importance of Features

Complementing the previous points, a graph was generated to illustrate the impact of categories within each feature, in this case, for predicting class or nature 0, using SHAP. This library allows us to analyse the contributions of categories and their impact on the prediction generated by the model. To better understand, in Fig. 10, we can observe the relevance of the features, divided by categories, in predicting class 0. In red, categories with a positive impact on the effective prediction of this class are highlighted. On the other hand, in blue, the category "Month with value 12. December" stands out, which has a negative impact and adversely affects the classification of the target variable. This result may be due to this category having the least amount of data, making it a scenario where identifying patterns and correlations is more challenging.

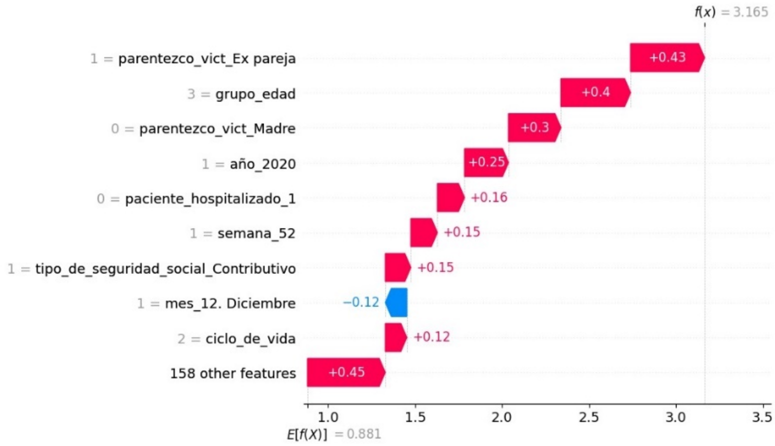

Figure 10: Feature impact with SHAP

When comparing the results obtained by the model, a variety of performances can be observed. For example, in the study [Ferreira, 2022], where a KNN model was developed to predict theft actions in the city of Bucaramanga, values of 59% and 60% were obtained in recall and accuracy metrics, respectively. The model presented in this research shows significantly higher performance metrics. However, there were also models in related works where even more remarkable performance was achieved. In the model presented by Escobar-Linero [2023], where an Artificial Neural Network (ANN) was implemented to predict victims' disengagement in legal processes related to partner violence cases, a result of 91% was obtained in the accuracy metric. These comparisons highlight the importance of considering performance metrics and contexts when selecting and evaluating machine learning models.

It is relevant to mention that the key features identified in this model, such as the incident scenario, geographic area, the relationship between the victim and the perpetrator, as well as the department or commune, show similarities with factors highlighted in other studies. For instance, similar to a study on conjugal sexual violence in India, where previous exposure to violence and geographical location were emphasized as crucial elements [McDougal, 2021], this model also underscores the importance of context in the occurrence of violent acts.

Moreover, another study on non-conjugal sexual violence in India identified factors such as geography and the victim's relationship with the perpetrator, noting that approximately 45% of cases were perpetrated by partners or ex-partners and 24% by relatives [Raj, 2021]. These findings are consistent with the results of this study, where a significant percentage of cases were perpetrated by relatives (53.17%), and records related to partners or ex-partners represented 22.91%. These similarities highlight the diversity of influential factors in violence, underscoring the importance of addressing this issue from multiple perspectives and contexts.

To use the dataset and the proposed model, a web application was created (https://clasificador-vg.web.app). It has a series of functionalities ranging from observing the variables of the dataset to creating different types of graphs such as bar, lines, and donut using the information obtained from the analysis of the dataset. These graphs utilize the information obtained from the dataset analysis with the aim of providing a better understanding of the data and the predictions made by the model. Additionally, in the tab called predictor (Fig. 11), it is possible to use the model to predict nature from variables entered by the user through the web application. The application is able to predict the gender-based violence nature with an average accuracy of more than 84%.

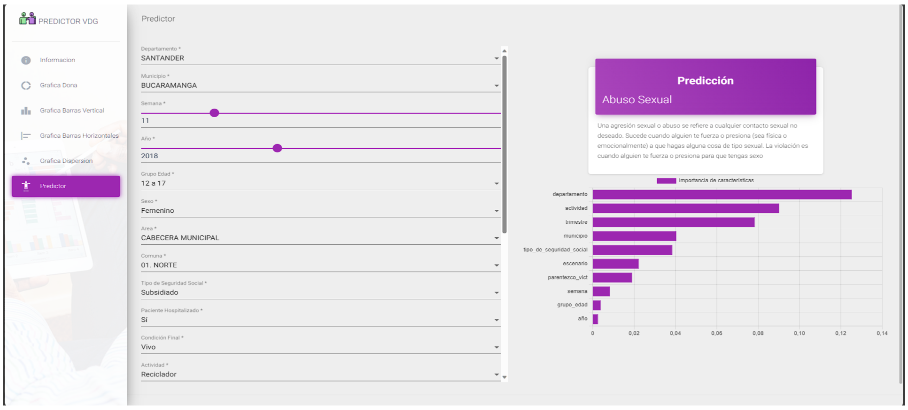

Figure 11: Prediction Website

In addition to the above, when making the prediction, the application displays the result plus a brief description and a graph showing the most important features when making the prediction. It is important to highlight that, to carry out the process described in this article thoroughly, several steps were followed. Additionally, a systematic mapping was conducted for the related works section, which delved into the existing issues and provided the basis for the development of the tool presented in this document. Furthermore, the CRISP-DM methodology was employed for its creation, which consists of several stages (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment). This methodology ensured a robust development of the tool by following a consistent methodological approach.

5 Conclusions and Limitations

The proposed model to predict cases of gender-based violence in Colombia has demonstrated high accuracy. The process of selecting the final model was meticulous and covered various stages, from the analysis and careful treatment of the original dataset to the evaluation of multiple available models. After an exhaustive search for hyperparameters, it was possible to fine-tune the performance of the model. These development and optimization phases resulted in a highly promising model, providing confidence in its ability to contribute to the identification and prevention of gender-based violence cases.

The developed tool contributes to the approach taken in Colombia to address GBV. As in the case of the Ministry of Health and Social Protection, where the implementation of a comprehensive approach to fight it is prioritized and supported by various strategies [Ministerio de Salud de Colombia, 2023b]. Within these, Knowledge Management is key, and the tool developed plays a significant role in this framework. It also aligns with the interests of the 2023 – 2026 National Development Plan, which aims at guaranteeing women's rights and proposes the creation of the National Monitoring System for Gender-based Violence and also the creation of a new national policy on sexual and reproductive rights, with the purpose that women have a life free of violence [Departamento Nacional de Planeación, 2022].

Furthermore, this tool facilitates a detailed understanding of the nature of incidents linked to gender-based violence. This understanding is essential for the formulation of appropriate measures and responses. Also, by storing the predictions made, the tool makes it possible to maintain an accurate record of the data, thus facilitating the analysis of trends in cases of gender-based violence.

Despite the rigorous process to obtain the final model, the metrics did not experience a substantial improvement compared to the base XGBoost algorithm; however, the final model has demonstrated its ability to successfully predict GBV cases based on user input. This underscores the fundamental limitation in the availability of comprehensive and extensive datasets in Colombia, which is a challenge in the context of modelling related to gender-based violence. The lack of adequate datasets with outstanding and nationally relevant attributes has been a notable obstacle in this process. Despite this, the final model represents a significant step in the right direction to address a critically important issue such as gender-based violence.

An important limitation of the developed model is that if the data used to train the model contain incorrect or biased classifications, the accuracy of the predictions made will be affected. Additionally, it is necessary to consider that the captured records do not represent the entirety of gender-based violence cases, as reflected in the data published by the Observatory of Feminicides in Colombia. Between 2019 and 2021, there were 1,826 cases of gender-based violence reported, but according to figures from the Attorney General's Office, only 677 cases were criminally reported during the same period [Prieto, 2022].

Furthermore, it must be taken into account that the developed tool is conceived as a first step in the fight against gender-based violence from a machine learning perspective. However, it is necessary to consider that the use of this technology entails inherent challenges, such as the need for significant resources like computing power, data infrastructure, and machine learning expertise, as well as specialized knowledge in the field. If these challenges are not addressed properly, they may pose obstacles to the effective integration of machine learning as a relevant tool in this fight.

Additionally, in order to achieve better integration of the developed model in the fight against gender-based violence, it is crucial to consider that predictions must be evaluated by experts in the field. Although models can identify correlations, their ability to understand underlying causes may be limited. Gender-based violence events are influenced by various external factors such as sociocultural norms and institutional policies, which can be challenging to fully capture in the model. Therefore, the involvement of gender-based violence experts is essential for properly interpreting predictions and making informed decisions in combating this issue. Additionally, it should be noted that the use of new technologies may generate resistance among those responsible for combating gender-based violence, hence the development of a user-friendly tool and activities that encourage its use and understanding are vital.

Our research has certain limitations, such as the possible bias in the data by focusing exclusively on the Santander region. Additionally, we recognize that gender violence detection could benefit from including additional data, such as gestures or non-verbal cues. However, our strengths lie in the automation and scalability of the developed tool, providing valuable support to authorities in early detection and timely intervention in cases of gender violence.

References

[Alcaldía de Medellín, 2023] Alcaldía de Medellín (2023). Implementación de estrategias para prevenir y atender las violencias basadas en género [Internet]. https://www.medellin.gov.co/es/proyectos/implementacion-de-estrategias-para-prevenir-y-atender-las-violencias-basadas-en-genero/

[Arbeláez Arbeláez, 2021] Arbeláez Arbeláez, L. M., Pulido Villamil, L., & Romero, A. N. (2021). Afectaciones Instituto Colombiano de Bienestar Familiar Equipo técnico Subdirección de Restablecimiento de Derechos. Bienestar Familiar. https://www.icbf.gov.co/system/files/procesos/pu5.p_cartilla_afectaciones_v1.pdf

[Congreso de Colombia, 2015] Congreso de Colombia (2015). Ley 1761 de 2015. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=65337

[Congreso de Colombia, 2021] Congreso de Colombia (2021). Ley 2137 de 2021. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=168047

[Datos Abiertos Colombia, 2023] Datos Abiertos Colombia (2023). Violencia de Género e intrafamiliar de enero 2015 a marzo 2023 - Semana Epidemiológica 12 [Internet]. https://www.datos.gov.co/en/Salud-y-Protecci-n-Social/04-Violencia-de-G-nero-e-intrafamiliar-de-enero-20/sq8q-pnf5/about_data

[Departamento Nacional de Planeación, 2022] Departamento Nacional de Planeación (2022). Mujeres, centro del cambio en el Plan Nacional de Desarrollo [Internet]. https://2022.dnp.gov.co/Paginas/mujeres-centro-del-cambio-en-el-plan-nacional-de-desarrollo.aspx

[Escobar-Linero, 2023] Escobar-Linero, E., García-Jiménez, M., Trigo-Sánchez, M. E., Cala-Carrillo, M. J., Sevillano, J. L., & Domínguez-Morales, M. (2023). Using machine learning-based systems to help predict disengagement from the legal proceedings by women victims of intimate partner violence in Spain. PLoS One, 18(6), e0276032. https://doi.org/10.1371/journal.pone.0276032

[Ferreira, 2022] Ferreira, J. D. G., Nieto-Rodríguez, M. P., &. Rocha-Ruiz, C. A. (2022). Prediciendo el crimen en ciudades intermedias: un modelo de 'machine learning' en Bucaramanga, Colombia. URVIO. Revista Latinoamericana de Estudios de Seguridad, 34, 83-98. https://doi.og/10.17141/urvio.34.2022.5395

[Gestarsalud, 2020] Gestarsalud (2020). Violencia psicológica: un flagelo silencioso difícil de identificar [Internet]. https://gestarsalud.com/2020/12/10/violencia-psicologica-un-flagelo-silencioso-dificil-de-identificar/

[González-Prieto, 2023] González-Prieto, A., Brú, A., Nuño, J. C., & González-Álvarez, J. L. (2023). Hybrid machine learning methods for risk assessment in gender-based crime. Knowledge-Based Systems, 260, e110130. https://doi.org/10.1016/j.knosys.2022.110130

[Gonzalez Sarmiento, 2023] Gonzalez Sarmiento, N. F. (2023). Informe de Evento Primer Semestre Violencia de Género e Intrafamiliar y Ataques con Agentes Químicos [Internet]. https://www.ins.gov.co/buscador-eventos/Informesdeevento/violencia%20de%20genero%20informe%20primer%20semestre%202023.pdf

[IBM, 2021] IBM (2021). SPSS Modeler Documentation [Internet]. https://www.ibm.com/docs/es/spss-modeler/

[Instituto Mexicano de la Juventud, 2017] Instituto Mexicano de la Juventud (2017). Violencia contra hombres. Una violencia más silenciosa [Internet]. https://www.gob.mx/imjuve/articulos/violencia-contra-hombres-una-violencia-mas-silenciosa?idiom=es

[Instituto Nacional de Salud, 2022] Instituto Nacional de Salud (2022). Casos sospechosos de violencia de género e intrafamiliar en Colombia, semanas epidemiológicas 01 a 44 de 2022-Semana epidemiológica 46. In: Boletín Epidemiológico Semanal. www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2022_Boletín_epidemiologico_semana_46.pdf

[Lamos-Díaz, 2020] Lamos-Díaz, H., Puentes-Garzón, D. E., & Zarate-Caicedo, D. A. (2020). Comparison Between Machine Learning Models for Yield Forecast in Cocoa Crops in Santander, Colombia. Revista Facultad de Ingeniería, 29(54), e10853. https://doi.org/10.19053/01211129.v29.n54.2020.10853

[McDougal, 2021] McDougal, L., Dehingia, N., Bhan, N., Singh, A., McAuley, J., & Raj, A. (2021). Opening closed doors: using machine learning to explore factors associated with marital sexual violence in a cross-sectional study from India. BMJ Open, 11(12), e053603. https://doi.org/10.1136/bmjopen-2021-053603

[Matta, 2023] Matta, J., Singh, V., Auten, T., & Sanjel, P. (2023). Inferred networks, machine learning, and health data. PLoS One, 18(1), e0280910. https://doi.org/10.1371/journal.pone.0280910

[Ministerio de Salud de Colombia, 2023a] Ministerio de Salud de Colombia (2023a). Prevención de las violencias de género y sexual, y atención integral a las víctimas [Internet]. https://www.minsalud.gov.co/salud/Paginas/Prevenci%C3%B3n-de-las-violencias-de-g%C3%A9nero-y-sexual,-y-atenci%C3%B3n-integral-a-las-v%C3%ADctimas.aspx

[Ministerio de Salud de Colombia, 2023b] Ministerio de Salud de Colombia (2023b). Abordaje integral de las violencias de género [Internet]. https://www.minsalud.gov.co/salud/publica/ssr/Paginas/Abordaje-integral-de-las-violencias-de-genero.aspx

[Ministerio de Vivienda de Colombia, 2021] Ministerio de Vivienda de Colombia (2021). Ley 2172 de 2021. https://minvivienda.gov.co/normativa/ley-2172-2021

[Myhill, 2017] Myhill, A. (2017). Measuring domestic violence: context is everything. Journal of Gender-Based Violence, 1(1), 33-44.

[Núñez., 2024] Núñez, M. E., Siddiqui, M. K., & Abbas, A. (2024). Motivation, attitude, knowledge, and engagement towards the development of sustainability competencies among students of higher education: a predictive study. Discover Sustainability, 2024, 17, e5.

[Ordóñez, 2020a] Ordóñez, H., Cobos, C., & Bucheli, V. (2020a). Modelo de machine learning para la predicción de las tendencias de hurto en Colombia. Revista Ibérica de Sistemas e Tecnologias de Información-RISTI, (E29), 494-506.

[Ordóñez, 2020b] Ordóñez, H. A., Pardo-Calvache, C. J., & Cobos-Lozada, C. A. (2020b). Detección de tendencias de homicidios en Colombia usando Machine Learning. Revista Facultad de Ingeniería, 29(54), e11740. https://doi.org/10.19053/01211129.v29.n54.2020.11740

[Ordoñez-Erazo, 2022] Ordoñez-Erazo, H. A., Ordóñez, C., & Bucheli-Guerrero, V. A. (2022). Predicción de factores clave en el aumento de la demografía en Colombia a través del ensamble de modelos de Machine Learning. Revista Científica, 44(2), 282-295. https://doi.org/10.14483/23448350.19205

[Ordoñez-Eraso, 2023] Ordoñez-Eraso, H., Ordoñez-Quintero, C., & Bucheli-Guerrero, V. (2023). A model based on the gradient boosting regressor to predict trends in the ratio of residence in relation to the age of homeless people in Colombia. Investigación e Innovación en Ingenierías, 11(1), 145-157. https://doi.org/10.17081/invinno.11.1.6572

[Pinto-Muñoz, 2023] Pinto-Muñoz, C. C., Zuñiga-Samboni, J. A., & Ordoñez-Erazo, H. A. (2023). Machine Learning Applied to Gender Violence: A Systematic Mapping Study. Revista Facultad de Ingeniería, 32(64), e15944. https://doi.org/10.19053/01211129.v32.n64.2023.15944

[Prieto, 2022] Prieto, L. V. (2022). Balance sobre la violencia basada en género en Colombia. Fundación Paz & Reconciliación (Pares). https://www.pares.com.co/post/balance-sobre-la-violencia-basada-en-g%C3%A9nero-en-colombia

[Raj, 2021] Raj, A., Dehingia, N., Singh, A., McAuley, J., & McDougal, L. (2021). Machine learning analysis of non-marital sexual violence in India. EClinicalMedicine, 39, e101046. https://doi.org/10.1016/j.eclinm.2021.101046

[ReliefWeb, 2022] ReliefWeb (2022). Colombia: Situación de la Violencia Basada en Género (VBG), Comparativo 2020-2021 (abril 2022) [Internet]. https://reliefweb.int/report/colombia/colombia-situaci-n-de-la-violencia-basada-en-g-nero-vbg-comparativo-2020-2021-abril

[Rumney, 2017] Rumney, P. N., & McCartan, K. F. (2017). Purported False Allegations of Rape, Child Abuse and Non-Sexual Violence. The Journal of Criminal Law, 81(6), 497-520.

[Sánchez-Juárez, 2023] Sánchez-Juárez, A. (2023). Un estudio analiza los motivos que inhiben la denuncia de abusos sexuales. In: UOC News https://www.uoc.edu/portal/es/news/actualitat/2016/053-denuncia-abusos-sexuales.html

[Scikit Learn, 2022] Scikit Learn (2022). scikit-learn Documentation [Internet]. https://scikit-learn.org/

[Siddiqui, 2020] Siddiqui, M. K., Huang, X., Morales-Menendez, R., Hussain, N., & Khatoon, K. (2020). Machine learning based novel cost-sensitive seizure detection classifier for imbalanced EEG data sets. International Journal on Interactive Design and Manufacturing, 14(4), 1491-1509.

[Sower, 2021] Sower, E. A., & Alexander, A.A. (2021). The Same Dynamics, Different Tactics: Domestic Violence During COVID-19. Violence and Gender. 8(3), 154-156. https://doi.org/10.1089/vio.2020.0066

[Un Women, 2023] UN Women (2023). Ending Violence Against Women [Internet]. https://www.unwomen.org/es/what-we-do/ending-violence-against-women/

[Walker, 2019] Walker, A., Lyall, K., Silva, D., Craigie, G., Mayshak, R., Costa, B., Hyder, S., & Bentley, A. (2019). Male Victims of Female-Perpetrated Intimate Partner Violence, Help-

Seeking, and Reporting Behaviors: A Qualitative Study. Psychology of Men & Masculinities. Advance online publication. https://doi.org/10.1037/men0000222

[World Health Organization, 2022] World Health Organization (2022). Violence Against Women [Internet]. In: WHO Fact Sheets. https://www.who.int/es/news-room/fact-sheets/detail/violence-against-women