A Twist in Intimate Partner Violence Risk Assessment Tools: Gauging the Contribution of Exogenous and Historical Variables

Lara Quijano-Sánchez[a,b,*], Federico Liberatore[e,b], Guillermo Rodríguez-Lorenzo[a], Rosa E. Lillo[d,b], José L. González-Álvarez[c]

[a] *Escuela Politécnica Superior, Universidad Autonóma de Madrid, Madrid, Spain*
[b] *UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, Madrid, Spain*
[c] *Gabinete de Coordinación y Estudios, Secretaría de Estado de Seguridad, Ministerio del Interior, Madrid, Spain*
[d] *Statistics Department. Universidad Carlos III de Madrid, Spain*
[e] *School of Computer Science & Informatics, Cardiff University, UK*

Abstract

Gender violence is a problem that affects millions of people worldwide. Among its many manifestations Intimate Partner Violence (IPV) is one of the most common. In Spain, a police monitoring protocol has been developed to minimize recidivism in IPV cases. This protocol is complemented by VioGén, an Intimate Partner Violence Risk Assessment Tool (IPVRAT) created by the *Spanish State Secretariat for Security of the Ministry of Interior* (SES) for risk prediction. VioGén's goal is to help the authorities determine what security and safety measures are most suitable. This paper improves on the current version of VioGén by introducing a model based on machine learning and data science and by studying the predictive value of exogenous and historical variables. The model is fitted on an anonymized database provided by SES and extracted from VioGén. This database includes the 2-year evolution of 46,047 new cases of IPV violence reported between October 2016 and December 2017, making it the largest database analyzed in the field. Obtained results show a clear improvement in the predictive capabilities of the new model against the original system, where it would have corrected more than 25% of the infra-protected cases, while improving the overall accuracy at the same time. Finally, lessons learned from the performed study and experiments are reported to aid in the design of future IPVRAT. In particular, insights show that IPVRAT should not treat cases statically as the incorporation of information regarding their evolution improves significantly the model's performance.

Keywords: Police Risk Assessment, Reassault Risk Assessment, Machine Learning, VioGén System, Intimate Partner Violence, Gender Violence

1. Introduction

Gender-based violence is one of the most common human rights violations, affecting millions of people [30]. It constitutes an attack against the freedom, integrity and dignity of its victims [28]. Intimate Partner Violence (IPV) involves stalking, sexual violence, physical assault, threats, psychological intimidation or coercion, and any abuse of control of a partner in a intimate relationship [3]. Those who have been assaulted by an intimate partner are more at risk of repeated violence or even murder [3]. Between 2003 and 2014, it is estimated that 55% of all female intimate partner homicides (IPH) were linked to IPV [33]. It is important to notice that low IPH levels in Europe do not always lead to low IPV levels [30]. The need to identify IPH and IPV risk factors to predict the phenomenon and to identify persons with the highest harm potential has been previously illustrated in [31].

Risk evaluation and management is one of the most important methods in the area of IPV prevention, where Intimate Partner Violence Risk Assessment Tools (IPVRAT) are intended to assist the competent authority in charge of each case management. In this work, we address the problem of determining the most appropriate Protection Level (PL) for an IPV victim based on the case data, with the objective of eliminating (or minimizing) the possibility of recidivism. The PL specifies the protection measures and resources that are assigned to the victim and is associated to the severity of the case and its risk of recidivism.

In Spain, there is a police surveillance protocol to overcome this issue, which seeks to reduce IPV recidivism. This protocol is complemented by VioGén, a system created by the *Secretaría de Estado de Seguridad del Ministerio del Interior* (Spanish State Secretariat for Security of the Ministry of Interior, SES), for PL definition. VioGén's

*Corresponding author

Email addresses: lara.quijano@uam.es (Lara Quijano-Sánchez), liberatoreF@cardiff.ac.uk (Federico Liberatore), guillermo.rodriguezl@estudiante.uam.es (Guillermo Rodríguez-Lorenzo), rosaelvira.lillo@uc3m.es (Rosa E. Lillo), jlga@interior.es (José L. González-Álvarez)

objective is to promote the work of the competent authority (i.e the police, the civil guard or the appropriate security force in charge) in deciding on the most suitable security measures for each case.

The forecasting model of the current version of VioGén [18] has been approached from the field of social sciences [20, 19]. Therefore, the value of each predictor and its subsequent validation have been carried out using a small number of observations. On the other hand, Data Science and Machine Learning (ML) techniques, i.e., Naive Bayes, Support Vector Classification, Multinomial Logistic Regression, K-Nearest Neighbors and Random Forests, have been effectively used in many contexts to turn vast volumes of data into information. As an example, in the field of predictive policing, these methods have been used to forecast future criminal activity [24, 1, 8, 15, 17]. Based on a broader set of data than the original VioGén's validation set, this work aims to use the potential of these tools to study the workings of the current system, identify points for improvement and propose new variables for the better assessment of PLs.

A PL is a Likert scale value, that takes different ranges according to the different existing IPVRAT. Each value corresponds to the severity of a particular IPV case in a given moment. Also, in actuarial IPVRAT, these values are associated to the measures taken by the competent authority to protect the victim and prevent recidivism, where the higher the PL the more protective measures and resources invested in the case. Due to limitations in resources available it is not realistic to assign each case the highest value just to try to ensure that there is no recidivism. Hence, IPVRAT should predict with precision the PL that ensures that there is no recidivism, where logically, within resource's limits (out of the scope of this study) it is always preferable to overprotect the victim than to harbor the possibility that the PL falls short. For this reason, the problem tackled in this paper is bi-objective in nature, as the goal is to improve on existing IPVRAT by maximizing the model's accuracy while, at the same time, minimizing the underestimation of the PLs. For our task, an anonymized database is used, extracted from the current system and provided by SES. This database contains the two year evolution of the 46,047 newly register cases of IPV between October 2016 and December 2017.

Therefore, this paper's contributions are five: i) it analyzes the data provided and research improvement points of the current system; ii) it evaluates potential new predictive factors; iii) it studies alternative models using ML techniques; iv) it introduces a new research paradigm where, differently from existing IPVRAT that assess the risk level, the most appropriate PL is directly estimated, and v) it suggests implications that can be extrapolated to other data or IPVRAT. This work extends previous research in IPVRAT in three ways: i) by comparing multiclass and ordinal classification paradigms in the context of IPV; ii) by studying the significance of exogenous variables; and iii) by introducing predictive variables that represent the entire evolution of a case.

In this work, we research the impact of taking these aspects into account while designing the model. Our estimation reveals that the new model would have corrected more than 25% of the cases infra-protected by the original VioGén, while improving the global accuracy at the same time.

Our work entails immediate implications for predictive policing systems. The lessons and insights learned from testing different approaches and techniques can be extrapolated to the existing IPVRAT. This manuscript helps to further encourage the application of data-driven intelligent decision support systems in public bodies.

The rest of this paper is structured as follows: In Section 2 we describe VioGén's current version and analyze existing IPVRAT. Next, Section 3 defines this paper's methodology by introducing the considered input and output variables and ML approaches. Following, Section 4 describes the dataset, the experimental design and obtained results. Finally, Section 5 concludes the paper and proposes future research lines.

2. Related Work

Recently, a broad range review of existing IPVRAT was performed in [11], thus, identifying methodological strengths and gaps in the current literature. Table 1 introduces those that stand-out and illustrates their main differences.

In this paper we focus on actuarial IPVRAT. These tools are typically validated in follow-up studies, by testing their ability to determine if individuals who are accused or adjudicated for IPV offenses, reoffend or not. Performance parameters of these different tools have been reviewed and published in a variety of papers [11, 19, 14] where, in summary, findings related to reliability and validity are similar to those obtained by

IPVRAT	Country	Goal	Sample Size	N^{O} Predictive variables	N^{O} PLs	Model Weights	Evolution Form
ODARA [13]	Canada	Act	581	13	7	unweighted sum	No
SARA [16]	Canada	PJ	-	24	3	-	No
B-SAFER (SARA short version) [?]	Canada	PJ	-	15	3	-	No
DVSI-R [29, 32]	USA	Act	14,970	11	3	odds ratio weighted sum	No
VP-SAFvR [21]	Australia	Act	44,436	52	10	odds ratio weighted sum	No
RVD [25]	Portugal	Act	216	20	3	odds ratio weighted sum	Yes
DA [4]	USA	Act	634	20	4	odds ratio weighted sum	No
Lethality-Screen (DA short version) [22]	USA	Act	254	11	2	odds ratio weighted sum	No
SVRA-I [7]	Israel	Act	1,133	45	3	expert assigned weighted sum	No
VioGén [18]	Spain	Act	6,613	55	5	odds ratio weighted sum	Yes

Table 1: IPVRAT overview. Column *Goal* can take value *Act* or *PJ*, standing for actuarial tool (i.e., it makes use of an algorithm and acts as a decision support system for the competent authority) and professional judgment (i.e., only used to guide the interview), respectively. *Sample Size* is the number of observations in the training dataset. *N^{o} PLs* is the number of possible outcomes. All the actuarial models consist of a weighted sum of the variables; column *Model Weights* illustrates the type of weight adopted. *Evolution Form* indicates whether the system includes a specific form for tracking a case evolution.

VioGén [20, 19].

Regarding the algorithmic complexity of these tools and current state of automation, the above-mentioned studies have been approached from the field of social sciences. Where, as shown in Table 1, in the case of ODARA the prediction is performed by the unweighted sum of all risk factors, and in the rest of them the weight of each indicator is calculated as the odds ratio of the indicator with respect to a response variable (i.e recidivism or lethality found in training cases). Next, for each case, the risk numerical value is obtained by adding the weights of the indicators present in the case and the consequent PL is assigned according to manually devised intervals. Also, the level of automation is limited (e.g. Lethality-Screen is hand-in-situ computed). Therefore, one of this paper's goals is to approach the algorithmic design of these tools from the point of view of Data Science and ML. This goes in line with the limitations identified in [11]: i) future IPV risk assessment research should focus on better delineating the function and form of risk; and ii) risk is dynamic and should be reassessed to understand the risk posed at a particular time. In other words, IPV risk assessment is a process, not an end goal. The use of ML classifiers such as SVM and Random Forests has proven to be successful in a small study (353 homeless youth subjects) where authors used participants' answers to the Revised Conflict Tactics Scale [2] to assess whether their relationship was violent or not [24]. Also, Amusa, Bengesai, and Kahn [1] used Random Forests on data merging over 1,816 South African married women with the 2016 South African Demographic and Health Survey dataset to establish factors associated with the risk of experiencing IPV. These results encourage the further study of ML techniques in bigger samples and in existing IPVRAT. In fact, to the best of the authors' knowledge, the work here presented is the first of its kind as it not only identifies women who are vulnerable to IPV and the factors associated, but it also directly predicts the most appropriate PL which, as explained above, is directly correlated to the risk the victim might face and the urgency of protection.

Next, VioGén's working process is detailed being, this work's starting point.

2.1. VioGén's Current Version

VioGén's protocol is comprised of two main tools: the VPR (that in Spanish stands for Police Risk Assessment Form) and the VPER (that in Spanish stands for Police Risk

Protection Level	Time Window to next interview
extreme	72 hours
high	7 days
medium	30 days
low	60 days
unappreciated	60 days, only if there is a protection order

Table 2: Deadline for the next review

Assessment Evolution Form). The former is an instrument designed to assess IPV risk factors present in a relationship prior to the first report, while the latter is a follow-up form that complements the VPR by assessing changes in risk factor behavior since the prior report. The Spanish procedure followed in cases of gender-based violence is as follows: When a victim first reports IPV evidence to the institution concerned, the competent authority fills out a VPR form, complementing the information given by the victim with their own inquiries. These answers are run in the current risk prediction model and the system returns a PL recommendation (VPL): *unappreciated*, *low*, *medium*, *high*, or *extreme*. The competent authority subsequently decides on the actual Assigned Protection Level (APL). The APL entails a series of protection measures and, in addition, establishes a time window for carrying out a follow-up interview of the victim [18]. Table 2 shows the review window corresponding to each level. From this moment on, each time the victim attends to one of the periodic reviews, the competent authority fills out a VPER form. Analogously to the previous case, the results are entered into a second prediction model (generated this time from the responses collected in the VPER forms), which recommends a PL (VPL). The competent authority then updates the APL assigned to the case, consequently modifying the security measures if necessary, and establishing the time window within which the next follow-up interview must be carried out (according to Table 2). A schema of this process is illustrated in Figure 1.

Note that the victim can report new events that have occurred before the next periodic review, meaning that there has been recidivism. If this occurs, as in the previous case, a VPER form is filled out with the new information collected and the PL is reassessed, modifying the security measures if necessary and establishing the new term for the next periodic review.

VioGén is an actuarial IPVRAT [18]. For its construction, the weight of each indicator was determined as the odds ratio of the indicator itself with respect to the observed six-

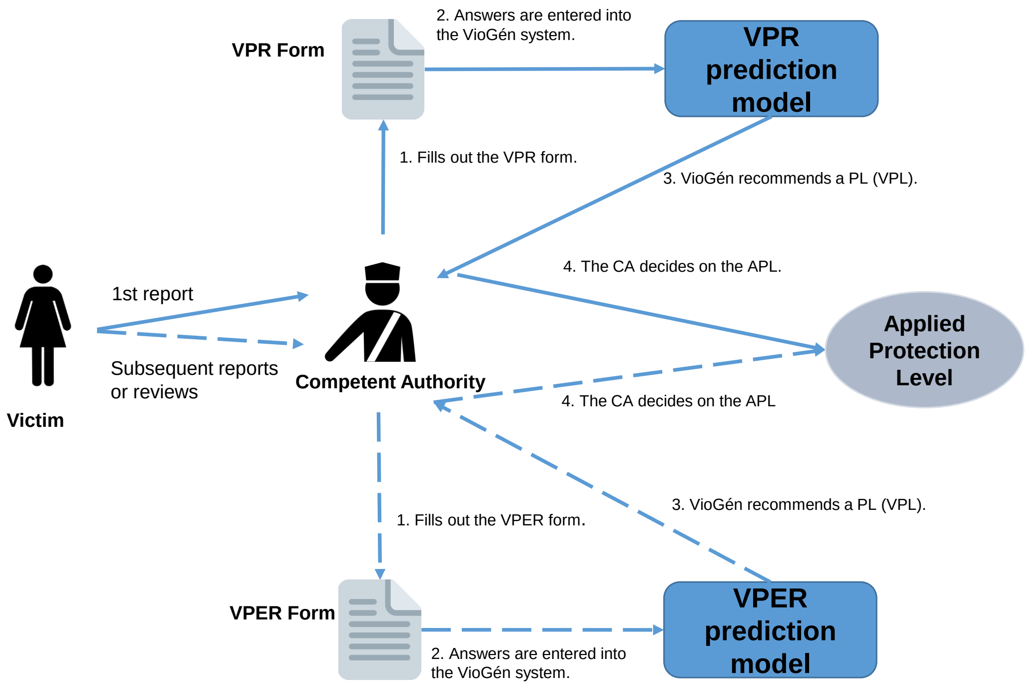

Figure 1: Schema of Viogen's working process.

month recidivism in a sample of 6,613 cases from 2015. Next, for each scenario, the numerical value of the risk was obtained by adding the weights of the indicators present in the case, as shown in Equation 1. The *score* is computed as the linear combination of the answers vector (**ans**) and their associated weights (**w**). The corresponding PL was determined according to threshold values, defined heuristically by VioGén's authors using ad-hoc rules based on their expertise. This is formalized in Equation 2, where $\mathrm{PL} = \{unappreciated, low, medium, high, extreme\}$ is the ordered set of PLs, indexed by l, and th_l are the corresponding thresholds.

$$score = \mathbf{w} \cdot \mathbf{ans} \tag{1}$$

$$\mathrm{VPL} = \arg\max_{l \in \mathrm{PL}} \{score \geq th_l\} \tag{2}$$

In this way, the Spanish protocol is among the most advanced IPVRAT currently in use. First, because of its complexity; VioGén makes use of two questionnaires, one to establish the initial PL and the another to reassess it according to the case's evolution. Only the Portuguese tool RVD resembles this functionality [6]. Second, due to its national implementation and its accuracy, comparable to ODARA, VP-SAFvR, SVRA-I, RVD and Lethality-Screen [26, 21, 7, 6, 22]. Finally, because it is developed on a computer system that allows thousands of users to connect at the same time. Only Australia (VP-SAFvR [21]) and Israel (SVRA-I [7]) employ a similar system.

3. Methodology

The limitations found in the previous section highlight: i) the static nature of the so-far developed approaches; ii) the lack of homogeneity in recidivism's definitions and the associated most appropriate PL [11]; and iii) the lack of diversity in studied prediction models that are mainly reduced to actuarial models.

Therefore, we address these limitations by: i) studying the impact of evaluating the case's history as well as exogenous variables; ii) defining a new paradigm for the computation of the most appropriate PL, by associating it to the recidivism's time windows; and iii) approaching the identification of the most appropriate PL using machine learning methods.

Following the points above, this paper focuses exclusively on the VPER prediction model. The VPR prediction model is out of scope as it cannot be extended by adding historical information on the case.

3.1. Input Variables

This subsection introduces the features used to represent a case in the VPER prediction model. In the model, each report is characterised by a vector x. The group of features that comprise this vector are summarized in the following:

- Form information, x_F. This group includes the answers to the VPER form.
- Exogenous information, x_E. This group incorporates information relative to the case that is not part of the VPER form.
- Historical information, x_H. This group includes variables that represent the case's evolution.

Therefore, $x = (x_F, x_E, x_H)$. The original VioGén system only includes x_F, whereas the other features are novel to this work. The following subsections describe in detail the content of each feature group.

3.1.1. Form Information Feature Group

This group of features includes the variables that represent the answers to the VPER forms filled by the competent authority after the interview with the victim. Table A.1 in Appendix A details the structure of the VPER form. This consists of seven different types of questions:

Type A questions : answered as "Yes" or "No".

Type B questions : answered as "Yes", "No" or "Don't know".

Type C questions : answered as "Yes", "No" or "Not applicable".

Type D questions : answered as "Slight", "Serious" or "Very serious".

Type E questions : multiple-choice answers.

Type F questions : answered as "Null", "Low" or "High".

Type G questions : answered as "Underestimate", "Overestimate" or "Equal".

The variables are encoded using one-hot, save for types D and F where we use a [0,0.5,1] Likert-scale. After the encoding, the total number of features comprising x_F is 85.

3.1.2. Exogenous Information Feature Group

This feature group represents the following information on the case:

- Institution where the complaint was filled, represented using one-hot encoding over the four possible institutions in Spain.
- Author's and victim's ages, one-hot encoding over ranges of five years for the ages' variables.

- Information on the municipality and the province where the report was taken. The locations' populations are encoded numerically using the absolute value, numerically as a normalized 0-1 value, and one-hot encoded on a discretized range. Statistics on the number of inhabitants have been obtained from the Spanish National Institute of Statistics.

More details are presented in Table A.2 in Appendix A. After the encoding, the total number of features comprising x_E is 46.

3.1.3. Historical Information Feature Group

The case history feature group incorporates: i) features representing the change in the responses to the current VPER form with respect to the previous form filled and ii) summary statistics on the case and the APL evolution.

The first set allows to understand if the condition is worsening, improving or staying stable. In fact, for each of the questions, two binary variables are introduced that represent whether the response has increased or decreased in value since the last form filled. For each type of question, the ordering of the possible answers is illustrated in the following:

Type A : "No" < "Yes".

Type B : "No" < "Does not know" < "Yes".

Type C : "No" < "Not applicable" < "Yes".

Type D : "Slight" < Serious" < "Very serious"

Type E : option not chosen < option chosen.

Type F : "Null" < "Low" < "High".

Type G : "Underestimate" < "Equal" < "Overestimate".

The summary statistics on the case and the APL evolution are captured in the following variables.

- Number of times that each APL value has been assigned to the case.

- First and last APL values assigned to the case.
- Number of VPER forms previously filled in the case.
- A binary variable that takes value one if the current VPER form is the first VPER form filled, and zero otherwise.

Overall, x_H is comprised of 179 features.

3.2. Response Variable

Below, a formal presentation of the model's response variable y is given. This hinges on the detection of recidivism in the case. Therefore, first the definition of recidivism adopted in this research is introduced, then, the response variable is formally defined.

3.2.1. Recidivism

This research adopts the definition of recidivism provided by SES. According to SES, recidivism is detected in a case when a victim suffers violence, threats, or procedure breaches from the aggressor since the last assessment of the case. The victim may report the incident before or during the next scheduled review. In either case, a VPER form is filled; in the form it is possible to specify the type and subtype of recidivism: violence (question 1, Table A.1), use of weapons (question 2, Table A.1), threats (question 3, Table A.1), or procedure breaches (question 4, Table A.1). Therefore, recidivism can be inferred from a VPER form if any of the previous questions are answered "Yes". This can be easily extended to the recidivism subtypes.

3.2.2. Optimal Protection Level

As detailed in Section 2, previous models from the literature are concerned with computing the probability of recidivism, which is then translated into a recommended PL according to manually-designed probability intervals [26, 21, 7, 6, 22, 20]. On the other hand, the focus of this paper is on directly computing the most appropriate PL for a case, referred from this point onward as the Optimal Protection Level (OPL), to avoid subjective design decisions. The rationale is assigning to a case the lowest possible PL that results in no recidivism detected before the next scheduled review. The lowest possible PL is chosen in order to efficiently use police resources and ensure a better service to all IPV's victims.

It is possible to compute the OPL for past VPER forms *a posteriori*, by considering the incumbent form's APL and if recidivism was detected as a consequence thereof. More formally, let PL $= \{unappreciated, low, medium, high, extreme\}$ be the ordered set of PLs, indexed by l. Each $l \in$ PL has an associated time window, tw_l (see Table 2). Given a form, let APL $\in$ PL be its assigned APL. The parameter rec takes value 1 if recidivism is detected in the next VPER form, according to the definition given in § 3.2.1, and 0 otherwise. In case of recidivism, tr represents the time of recidivism, that is, the number of days passed between the incumbent and the next form. The OPL for the incumbent form can be computed as follows.

$$\text{OPL} = \begin{cases} \text{APL} & \text{if } (rec = 0) \\ \min \{l \in \text{PL} | tw_l < tr\} & \text{if } (rec = 1) \wedge (\exists l \in \text{PL} | tw_l < tr) \\ extreme & \text{otherwise} \end{cases} \tag{3}$$

In other words, the OPL is set to be equal to the APL if there was no recidivism. In case of recidivism, the OPL is the lowest PL whose associated time window is smaller than the time of recidivism. If such PL does not exist (i.e., the time of recidivism is smaller than the time window associated to the *extreme* PL), then the OPL is equal to *extreme*.

As an example, if the victim was given an $APL = low$ and the case relapsed within ten days of filling in the form ($tr = 10$), the considered OPL is *high*, which according to Table 2 has $tw = 7$. Therefore, it fulfills Equation 3 as *high* is the minimum PL whose time window is strictly smaller than time of recidivism[1]. On the contrary, if a *medium* APL was given and there was no recidivism in the time window, the OPL is set to *medium*, as the APL was successful.

The OPL is used in the model as response variable y. Note that the definition of OPL given in Equation 3 can be easily extended to recidivism subtypes (see § 3.2.1) and, applied thus to specific recidivism subtypes models.

[1]It is important to notice that *extreme* also has a time window smaller than the time of recidivism. However, it is not the minimum PL, as *extreme* > *high* by definition of PL.

3.3. Model

As mentioned, our bi-objective problem consists of providing an estimation of the OPL that results in the best accuracy while, at the same time, minimizing the underestimations. Given a dataset comprised of N observations, their OPL y and the PL estimated by a model $\hat{y}_i$, the accuracy and the underestimations of the model can be computed as follows.

$$\text{acc} = \frac{|\{i = 1, \ldots, N : \hat{y}_i = y_i\}|}{N} \tag{4}$$

$$\text{und} = \frac{|\{i = 1, \ldots, N : \hat{y}_i < y_i\}|}{N} \tag{5}$$

where $i = 1, \ldots, N$ is the index used to refer to an observation. There exists a clear trade-off between these two objectives. In fact, it is possible to have no underestimations by assigning all the forms the highest possible PL. This approach, on top of being virtually inoperative, would result in a extremely low accuracy.

The problem is addressed by applying machine learning models to fit the response variable y to the corresponding inputs x. The response variable y is ordinal in nature. Therefore, two approaches are compared: multiclass classification and ordinal classification.

For the ordinal classification model, we implement the algorithm proposed by Frank and Hall [9], which is summarized in the following. Frank and Hall's methodology hinges on transforming a K-class ordinal problem to $K - 1$ binary class problems. This is achieved by converting an ordinal attribute $A^\star$ with ordered values $V_1, V_2, ..., V_K$ into $K-1$ binary attributes, one for each of the original attribute's first $K-1$ values, where the k-th binary attribute represents the test $A^\star > V_k$. Then, $K - 1$ independent probability models are fit, one for each attribute. A new observation $\hat{x}$ can be classified by predicting the probabilities of satisfying each $A^\star > V_k$ test, $Pr(\hat{y} > V_k)$. These probabilities can be used to calculate the probability of $\hat{x}$ belonging to a class V_k, $Pr(\hat{y} = V_k)$, as follows:

$$\begin{aligned} &Pr(\hat{y} = V_1) = 1 - Pr(\hat{y} > V_1) \\ &Pr(\hat{y} = V_k) = Pr(\hat{y} > V_{k-1}) - Pr(\hat{y} > V_k), \ \forall 1 < k < K \\ &Pr(\hat{y} = V_K) = Pr(\hat{y} > V_{K-1}) \end{aligned} \tag{6}$$

The class with maximum probability is assigned to the observation:

$$\hat{y} = \underset{V_k, \forall k=1,\ldots,K}{\arg\max} \{Pr(\hat{y} = V_k)\} \tag{7}$$

Apart from its simplicity, this methodology has the added benefit of allowing the direct penalization of class underestimation by applying appropriate weights to the observations when fitting each of the binary classification problems. In particular, given a value V_l, the observations that comply with $A^\star \leq V_l$ are assigned a penalization coefficient $\rho_i = 1$, while observations that satisfy $A^\star > V_l$ can be assigned a coefficient $\rho \geq 1$. A value of $\rho = 1$ implies no underestimation penalization; on the other hand, a larger value of ρ corresponds to a stronger underestimation penalization. Given an observation i, the corresponding underestimation weight, w_i^ρ, is obtained by normalization:

$$w_i^\rho = \frac{\rho_i}{\sum_{j=1}^{N} \rho_j} \tag{8}$$

Furthermore, prior to fitting both the multiclass and the ordinal model, it is possible to assign weights to the observations to balance the dataset. For all the observations i such that $y_i = V_k$, the associated balancing coefficient is

$$\beta_i = \frac{\sum_{k' \neq k} N_{k'}}{N} \tag{9}$$

where $N_{k'}$ is the number of observations whose class is $V_{k'}$ and |k|= |PL|. Given an observation i, the corresponding balancing weight, w_i^β, is obtained by normalization.

The underestimation and balancing weights can be combined by multiplying w_i^ρ and w_i^β.

Different classical multiclass and binary classification models [12] have been tested (i.e., Naive Bayes, Support Vector Classification, Multinomial Logistic Regression, K-Nearest Neighbors and Random Forests). However, initial experiments (not reported for the sake of brevity) showed that XGBoost [5] provided the best results, that are illustrated in the next section.

4. Experiments and Results

This research considers all the cases newly introduced into the VioGén system between October 2016 and December 2017 (46,047 cases) and the VPER forms corresponding to the two-year follow-up of each of them (255,425 records). To the best of the authors' knowledge, this is the largest IPV case study carried out to date [11, 14]. Given its relevance to the research community and its representativeness to the Spanish reality in the following subsection we perform a descriptive overview of the dataset which includes:

- A general description of the cleaning process, as well as the number of studied cases.
- A preliminary statistical analysis of recidivism cases.
- A study of VioGén's performance (VPL) on the dataset against the OPL.
- Analysis and insights of the APL's: distribution in the dataset, its performance against the OPL and variations with respect to the VPL.

Next, this paper's research questions and the proposed experimental design to address them are introduced in a new subsection followed by the subsequent models' results and a discussion on them.

4.1. Dataset

Prior to the dataset generation, a pipeline comprised of cleaning (e.g., checking for duplicate cases, removing incomplete cases, checking and fixing coherence issues in the forms' answers), variable encoding, and analysis was carried out. Note that all the encoding and cleaning decisions have been checked by SES for correctness and coherence.

After the cleaning step, the dataset includes 44,655 cases and 252,689 VPER forms. Of the latter, 20,864 forms are without recidivism and 231,825 are with recidivism. On average 5.66 VPER forms are registered per case.

4.1.1. Recidivism Analysis

By studying recidivism in the dataset the following is observed:

Recidivism type \ Grouped Prob.	**Total**	**Unappreciated**	**Low**	**Medium**	**High**	**Extreme**
VPER	0.0762	0.0959	0.0499	0.0814	0.0826	0.1320
VPER w/out past recidivism	0.0713	0.0947	0.0463	0.0653	0.0624	0.0597
VPER w/past recidivism	0.1531	0.2781	0.1506	0.1578	0.1150	0.1490

Table 3: Probability of recidivism in the period after a VPER depending on the APL.

Out of the 44,655 cases, there is some form of recidivism in 9,086 of them. Out of these, the average number of recidivism reports is 1.67 and the median is 1. The probability of recidivism in the period after a VPER depending on the APL is shown in Table 3. In particular, the last two rows segment the first row (VPER) according to whether the case itself is recidivist since the previous form. Also, the column 'Total' presents the probabilities for the unsegmented dataset, which correspond to the average of the APLs' probabilities, weighted by the number of cases in each group. From the analysis of the table it can be seen that for the VPER forms (first row) the probability of future recidivism tends to increase as the APL increases (being sequentially higher for all PLs with the exception of *Unappreciated*). However, by looking at the last two rows, it is possible to observe that the probability of future recidivism changes depending on past recidivism. In fact, the distribution in the second row (VPER w/out past recidivism) displays the opposite behavior, and assigns the highest probability of recidivism when an unappreciated PL is applied. On the other hand, by inspecting the last row (VPER w/past recidivism), it can be inferred that past recidivism increases the probability of future recidivism; also, the latter is largely unaffected by the APL, except when an unappreciated PL is assigned. Further studies are required to clarify the reasons behind this behavior, and this is left for future research in criminology and forensic psychology.

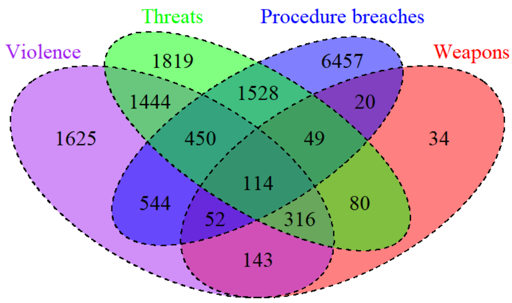

Figure 2: Venn's diagram on different types of recidivism.

Finally, Figure 2 illustrates the frequency of each subtype of recidivism collected in the forms, as well as their intersections in a single case. Where, it can be seen that, procedure breaches is the most common subtype and it is usually accompanied by threats.

4.1.2. VioGén Protection Level

We now analyze VioGén's performance against the OPL: Table 4 illustrates the confusion matrix; the main diagonal displays the cases where the VPL is exactly the OPL, the upper triangle the cases where VioGén would have overprotected the IPV victim, and the lower triangle the cases where VioGén's recommendation fell short, resulting in recidivism. According to this, VioGén's percentage of accuracy and underestimation are 80.57% and 15.54%, respectively.

OPL \ VioGén	Unappreciated	Low	Medium	High	Extreme
Unappreciated	93,553	3,768	729	57	13
Low	18,763	74,428	3,508	230	28
Medium	3,172	6,695	30,054	854	194
High	1,314	2,180	3,745	4,214	452
Extreme	577	929	1,278	616	1,338

Table 4: VPL vs OPL.

4.1.3. Applied Protection Level

Regarding the study of the APL across our dataset: Table 5 shows its distribution.

	Unappreciated	Low	Medium	High	Extreme
VPER	108,527	94,033	41,208	7,519	1,402

Table 5: APL distribution on studied datasets.

As explained in Section 2.1, the APL is the PL assigned to the victim by the competent authority after the interview. It is important to remark that this value is determined after the competent authority has received the VPL recommendation. Additionally, Table 6 compares the APL to the OPL. Following the OPL's definition given in the Section 3.2.2, it can be verified that the OPL is always equal to the APL (resulting in a 0 valued matrix upper triangle), except when there has been recidivism (corresponding to the lower triangle). Thus, the matrix lower triangle reflects the occasions where the applied PL was not sufficiently high. In summary, the APL's percentage accuracy and underestimation are 92.45% and 7.55%, respectively. It is important to notice that these results depend on the fact that the OPL is computed from the APL. Moreover, by definition, the former cannot be lower than the latter. Therefore, the two PLs are highly correlated.

OPL \ APL	Unappreciated	Low	Medium	High	Extreme
Unappreciated	98,120	0	0	0	0
Low	7,620	89,337	0	0	0
Medium	1,210	1,904	37,855	0	0
High	1,080	1,867	2,060	6,898	0
Extreme	497	925	1,293	621	1,402

Table 6: APL vs OPL.

Further on, Table 7 compares the VPL to the APL. In particular, the table can be used to observe the degree of agreement/disagreement between them. Table 8 provides a summary of the results; taking the VPL as the reference, the table illustrates the number of observations (and ratio) where the APL was lower, equal or higher than the VPL. Overall (first column), the competent authority agrees with VioGén 86.76% of the times; also, the former increases the PL (8.86%) twice as much than they decrease it (4.4%). In the following columns, the results are segmented according to the VPL's value. It can be seen that the agreement between the VPL and the APL tends to decrease as the VPL's value increases, with the competent authority favoring reducing the PL for higher values of VPL, and vice versa.

VioGén \ APL	Unappreciated	Low	Medium	High	Extreme
Unappreciated	103,079	12,294	1,934	68	4
Low	4,397	77,804	5,469	322	8
Medium	944	3647	32,544	2,147	32
High	87	257	1,021	4,497	109
Extreme	20	31	240	485	1,249

Table 7: VPL vs APL.

	Total	Unappreciated	Low	Medium	High	Extreme
Lower	11,129 (0.0440)	0 (0)	4,397 (0.0500)	4,591 (0.1168)	1,365 (0.2286)	776 (0.3832)
Equal	219,173 (0.8674)	103,079 (0.8782)	77,804 (0.8841)	32,544 (0.8278)	4,497 (0.7531)	1,249 (0.6168)
Higher	22,387 (0.0886)	14,300 (0.1218)	5,799 (0.0659)	2,179 (0.0554)	109 (0.0183)	0 (0)

Table 8: Number of times (ratio) that APL was lower, equal, or higher than VPL. The total values are given (first column), as well as the results segmented according to VPL's value (columns two to six).

Given the high percentage of agreement between the VPL and the APL, VioGén is expected to perform particularly well, as the VPL is correlated to the APL which, in turn, is correlated to the OPL. Therefore, the only opportunity for improving on VioGén's performance lies in the observations that VioGén underestimated. For this reason, this paper focuses on devising prediction models that dominate VioGén in both accuracy and underestimations.

4.2. Experimental Design

Given the problem of predicting the OPL of a VPER form, our research aims at providing an answer to the following research questions.

RQ1 Is there any significant difference between using a multiclass and a ordinal model in the problem considered?

RQ2 Does including exogenous variables (x_E) result in an improvement in the performance of the model compared to VPL?

RQ3 Does including historical variables (x_H) result in an improvement in the performance of the model compared to VPL?

To answer these questions, different models have been fit and tested, according to the following dimensions:

Model type multiclass (M) or ordinal (O).

Class-balancing weights unbalanced (U) or balanced (B).

Underestimation penalty (only for the ordinal model) $\rho = 1$ (i.e., no penalization) (1), $\rho = 2$ (2), $\rho = 4$ (4), or $\rho = 8$ (8).

Dataset full dataset (no suffix), no exogenous variables (-E suffix), no historical variables (-H suffix), or no exogenous and historical variables (-EH suffix).

The letters and numbers between brackets are used in the acronyms adopted in the rest of the paper to identify each model. For example, MU-H corresponds to a multiclass unbalanced model fitted on the dataset without historical variables, and OB2 is an ordinal model fitted on the full dataset and including both class balancing and underestimation ($\rho = 2$) weights. Overall, 40 different models have been considered, corresponding to all the combinations of the above dimensions. All models were programmed in R (version 4.1.0) and the experiments were run on a HP Z440 Workstation equipped with an Intel Xeon CPU E5-1650 v3 and 128 GB RAM, using multithreading.

As mentioned, the ML model that provided the best performance was XGBoost; the hyperparameters of the models have been tuned using Bayesian Optimization with Gaussian Processes [27]. Given its random nature, all accuracy estimates were obtained by averaging the results from 10 separate runs of randomized 10-fold cross-validation.

4.3. Model Results and Discussion

		MULTICLASS		ORDINAL P=1		ORDINAL P=2		ORDINAL P=4		ORDINAL P=8	
		acc	und	acc	und	acc	und	acc	und	acc	und
ALL	BAL.	0.8101	0.1172	0.8122	0.1152	0.8124	0.1153	0.8123	0.1152	0.8123	0.1153
	UNBAL.	0.8101	0.1171	0.8122	0.1153	**0.8124**	**0.1152**	0.8124	0.1153	**0.8124**	**0.1152**
NO EXO	BAL.	0.8092	0.1182	0.8107	0.1173	0.8107	0.1173	0.8105	0.1172	0.8105	0.1173
	UNBAL.	0.8091	0.1183	0.8105	0.1173	**0.8107**	**0.1172**	0.8106	0.1173	0.8106	0.1173
NO HIST	BAL.	0.7869	0.1541	0.7938	0.1451	0.7941	0.1449	0.7938	0.1450	0.7938	0.1450
	UNBAL.	0.7868	0.1538	**0.7940**	**0.1449**	0.7938	0.1450	0.7939	0.1450	0.7938	0.1451
NO EXO NO HIST	BAL.	0.7778	0.1551	0.7879	0.1546	**0.7878**	**0.1546**	0.7878	0.1546	0.7880	0.1544
	UNBAL.	0.7779	0.1550	0.7878	0.1546	0.7879	0.1545	0.7882	0.1545	0.7871	0.1545

Table 9: Average accuracy (acc) and underestimation (und) for all the models considered. In green best result overall models, in red best result within datasets.

Table 9 shows the average accuracy (acc) and underestimations (und) for all the models considered. By observing the table, the following general conclusions can be drawn:

- Balanced models have better (higher) accuracy, while unbalanced models have better (lower) underestimation.
- Models fitted using less variables perform worse. In particular, the historical variables have the greatest impact on the performance.
- The multiclass models perform worse than the ordinal ones.
- The underestimation penalty, ρ, does not have a significant impact on the performance of the models.

It is important to remind the reader that the goal is to identify a model with high accuracy and low underestimation. According to this, a dominance rule can be defined. A model dominates another if the former is non-worst than the latter in both criteria

Prediction	**acc**	**und**
APL	0.9245	0.0755
VPL	0.8057	0.1554
OU2	0.8124	0.1152

Table 10: Comparative performance for APL, VPL, and OU2, the best model obtained.

and is strictly better in at least one of the criteria. More formally:

$$\begin{aligned}&\mathrm{mod}_1 \succ \mathrm{mod}_2 \iff \\ &(\mathrm{acc}_1 \geq \mathrm{acc}_2) \wedge (\mathrm{und}_1 \leq \mathrm{und}_2) \wedge ((\mathrm{acc}_1 > \mathrm{acc}_2) \vee (\mathrm{und}_1 < \mathrm{und}_2))\end{aligned} \tag{10}$$

Also, two models are intransitive if they are not equivalent and they do not dominate each other. According to the definition, the best models are OU2 and OU8, which achieve equivalent performance. Following the principle of parsimony, model OU2 is chosen as the best model in the rest of the paper. Table 10 compares the average performance of OU2 to that of APL and VPL. As it can be seen, the best results are obtained by APL. This is expected, as the OPL is based on the value of the APL, as explained in detail in § 4.1.3. More interestingly, according to the results, OU2 dominates VPL. In fact, the percentage improvement with respect to the performance of VPL is 0.83% for the accuracy and 25.87% for the underestimation. Therefore, on average, OU2 improves only slightly on the VPL in terms of accuracy, while significantly reducing the underestimation.

Table 11 illustrates the confusion matrix for OU2*, i.e., the OU2 model that performed the best across the 10 repetitions of 10-fold cross validation. For this reason, the following values can be slightly different from the averages shown in Table 10. According to the confusion matrix, the accuracy of OU2* is 81.26%; its total underestimation is 11.50% and, also, OU2* underestimates with more than one level of difference just 2.83% of the cases. This result is even more impressive if we consider that VPL underestimates 15.54% of the cases, and that the difference between the models corresponds to 10,222 cases of recidivism that could have been prevented.

The disagreement in the responses of OU2* and VLP is illustrated in detail in Table 12, which highlights the difference between the confusion matrices of the two models, with respect to the OPL. Compared to the VLP, OU2* tends to overestimate more, generally erring by assigning a PL that is one class higher than the OPL. In this regard, OU2* is more conservative than VLP. Within the application context, this is a slight

OPL \ OU2*	Unappreciated	Low	Medium	High	Extreme
Unappreciated	87921	9709	464	24	2
Low	10821	80283	5701	142	10
Medium	1367	6277	31541	1671	113
High	956	2048	3663	4776	462
Extreme	414	943	1424	1134	823

Table 11: OU2* vs OPL.

OPL \ UO2 - VPL	Unappreciated	Low	Medium	High	Extreme
Unappreciated	-5632	5941	-265	-33	-11
Low	-7942	5855	2193	-88	-18
Medium	-1805	-418	1487	817	-81
High	-358	-132	-82	562	10
Extreme	-163	14	146	518	-515

Table 12: Differences of the confusion matrices of UOP2* and VPL vs OPL. In green: positive values on the main diagonal and negative values on the upper and lower triangles, indicating that UOP2* performed better than VPL. In red: negative values on the main diagonal and positive values on the upper and lower triangles, indicating that UOP2* performed worse than VPL.

mistake, as overestimations do not result in recidivism. The exception to this is the *extreme* PL, where OU2* is less accurate than VLP and underestimates more. However, the misclassified cases are assigned a *high* PL, erring only by one level.

To verify that the impact of the model's dimensions is statistically significant, a confidence interval analysis is carried out. Figure 3 is a scatter plot of the accuracy and the underestimation for all the ordinal models fitted using all the variables. Both the mean values (points) and the 95% confidence intervals (ellipses) are represented. The figure illustrates that all the ordinal models are statistically equivalent (i.e., the confidence intervals overlap), despite of differences in balancing and underestimation penalty. This same behavior is observed regardless of the dataset used (plots not provided for the sake of space).

Figure 4 represents the ordinal and multiclass models fitted using all the variables. It is possible to verify that balancing the weights does not have a significant impact on the multiclass models either. More importantly, it is possible to draw the conclusion that the ordinal models clearly dominate the multiclass model and that this result is statistically significant (i.e., the 95% confidence intervals do not overlap). Again, this conclusion is still valid regardless of the dataset used (plots not provided for the sake of space).

Figure 5 presents a graphical comparison between ordinal models fitted with dif-

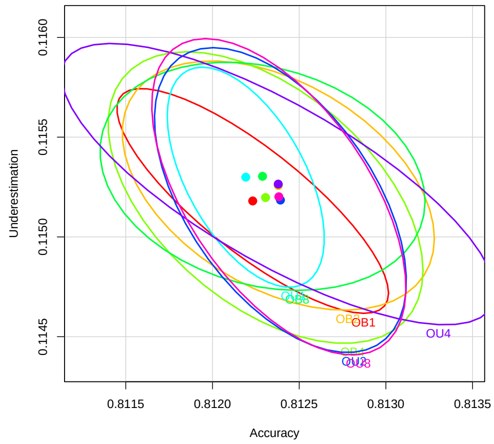

Figure 3: Scatter plot of the accuracy and the underestimation for ordinal models fitted using all the variables. The points represent the mean values, while the ellipses are the 95% confidence intervals.

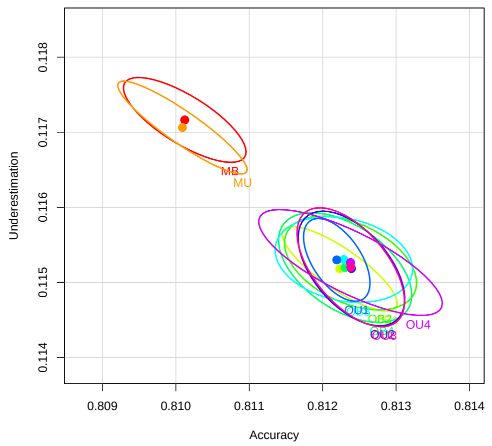

Figure 4: Scatter plot of the accuracy and the underestimation for ordinal and multiclass models fitted using all the variables. The points represent the mean values, while the ellipses are the 95% confidence intervals.

ferent datasets and VPL. For clarity, only unbalanced models with a underestimation penalty $\rho = 2$ are displayed as representative of all the ordinal models fitted using the same dataset. First, the figure allows us to make a comparison between datasets. Each dataset achieves a different performance, and the differences among them are statistically significant. Again, it is confirmed that the best results are obtained using the full dataset. Removing some of the variables invariably causes a significant reduction in both accuracy and underestimation. In particular, it is possible to observe that the historical variables contribute the most. Second, Figure 5 allows us to compare the ordinal model to the VPL and detect that OU2 and OU2-E dominate VPL, while OU2-H and OU2-EH are intransitive to VPL (i.e., they do not dominate each other). This allows us to infer that the inclusion of historical variables results in a significant improvement in the model's performance, while adding only the exogenous information does not produce a model that is significantly better than VioGén. Finally, given that model OU2 dominates model OU2-E we can conclude that, although the exogenous information by itself does not improve VioGén it does enhance the performance of a model significantly. The whole of these conclusions can be extended also to the multiclass model (plots not represented for the sake of space and clarity).

The conclusions obtained from the computational experiments are summarized in the following:

- On average, the best model is OU2.
- Given a dataset, ordinal models perform significantly better than multiclass models.
- Given a dataset and a type of model, balancing the dataset does not have a significant impact on the performance.
- Given a dataset, applying underestimation penalization does not have a significant impact on the performance of ordinal models.
- Ordinal models fitted using a dataset that includes the historical variables (i.e no suffix and -E suffix models) dominate VPL.
- Disregarding the historical variables results in an ordinal model that is irrespective to VPL.

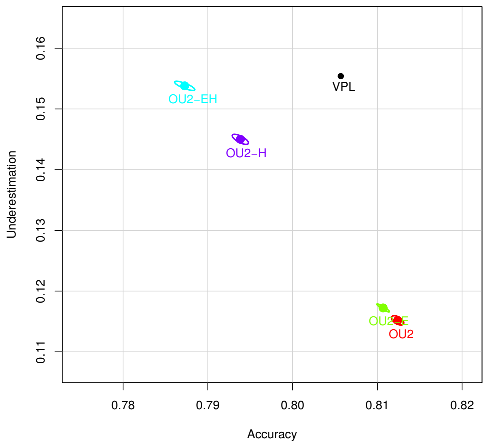

Figure 5: Scatter plot of the accuracy and the underestimation for the unbalanced ordinal models fitted using all the variables and applying an underestimation penalty $\rho = 2$. The points represent the mean values, while the ellipses are the 95% confidence intervals. For the purpose of comparison, VPL is included and represented with a black point.

It is now possible to answer our initial research questions:

RQ1 *Is there any significant difference between using a multiclass and a ordinal model in the problem considered?* Yes. Given a dataset, ordinal models perform significantly better.

RQ2 *Does including exogenous variables (x_E) result in an improvement in the performance of the model compared to VPL?* No. The resulting model is intransitive with VPL. However, it does enhance the model when coupled with historical data.

RQ3 *Does including historical variables (x_H) result in an improvement in the performance of the model compared to VPL?* Yes. The resulting model dominates VPL and the difference is statistically significant.

5. Conclusions and future work

Throughout this work, multiple advances have been made with regard to VioGén's current version. To do this: i) new exogenous variables have been studied with respect to the environment where the events take place, such as the number of inhabitants of the locality; ii) the evolution of the cases up to the moment prior to each VPER form has been included; iii) a new paradigm has been introduced when designing IPVRAT models by directly calculating the OPL instead of assigning a PL based on the probability of recidivism. This contribution is probably the most relevant in relation to the literature on actuarial IPVRAT, where classically the recidivism probability is studied with respect to the following six or 12 months, not according to time windows corresponding to OPLs. Thus, lessons learned on applying this technique serve for other IPVRAT. iv) Machine Learning techniques have been introduced when making predictions, where our model would have corrected between more than 25% of the cases that the original system infra-protected.

Various future study paths are proposed in the light of the results obtained. This research shows the importance of continuing to search for exogenous variables that represent the setting in which the case occurs, such as the rate of unemployment, the crime rate of the locality in which the incident occurs, prison reports or information of cases that are filed judicially. On the other hand, the results obtained when making predictions from the VPER forms show us the importance of representing the evolution of a case. One potential work line is to generate more detailed knowledge on the evolution of events. Also, the time windows displayed in Table 2 are arbitrary, based on the experience of experts, so our immediate future work will be to define those ranges based on data and factual information. A more comprehensive research may also be carried out on the importance of each variable in terms of recidivism. Specifically modeling via panel data. Also, future research should examine the administration of IPV/IPH risk assessment in non-Western countries and languages other than Spanish/English. When determining what tool would be most appropriate for a given setting, professionals should ensure that the tool has been tested in the target respondent's primary language [23].

Abbreviations

Acknowledgments

This research has been carried out in collaboration with SES, which gives full consent on its publication, i.e. the methodology, results, insights and data used to develop it.

IPVRAT	Intimate Partner Violence Risk Assessment Tools
IPV	Intimate Partner Violence
IPH	Intimate Partner Homicide
PL	Protection Level
VPL	VioGén Protection Level
APL	Applied Protection Level
OPL	Optimal Protection Level

Note that this paper's data complies with the GDPR, where no case can be traced. In addition, the nature of the dataset, consisting of all newly reported cases in Spain within a year, prevents potential bias of the algorithm. Also, as stated in [10], the Spanish questionnaires are action-oriented and have an automatic correction algorithm that reduces the subjectivity of the evaluators.

The research of Quijano-Sánchez was conducted with financial support from the Spanish Ministry of Science and Innovation, grant PID2019-108965GB-I00. The research of Liberatore is partially funded by the European Commission's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie, grant number MSCA-RISE 691161 (GEO-SAFE), and the Government of Spain, grant MTM2015-65803-R. All the financial support is gratefully acknowledged.

References

[1] L.B. Amusa, A.V. Bengesai, H.T. Khan, Predicting the vulnerability of women to intimate partner violence in south africa: evidence from tree-based machine learning techniques, Journal of interpersonal violence (2020) 0886260520960110.

[2] A. Bandura, The social learning perspective: Mechanisms of aggression. (1979).

[3] M. Black, K. Basile, M. Breiding, S. Smith, M. Walters, M. Merrick, J. Chen, M. Stevens, National intimate partner and sexual violence survey: 2010 summary report (2011).

[4] J.C. Campbell, D.W. Webster, N. Glass, The danger assessment: Validation of a lethality risk assessment instrument for intimate partner femicide, Journal of interpersonal violence 24 (2009) 653–674.

[5] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.

[6] O.S. Cunha, R.A. Goncalves, Severe and less severe intimate partner violence: From characterization to prediction, Violence and Victims 31 (2016) 235–250.

[7] K. Dayan, S. Fox, M. Morag, Validation of spouse violence risk assessment inventory for police purposes, Journal of family violence (2013) 811–821.

[8] P.M. Fandino, J.B. Tan Jr, Crime analytics: Exploring analysis of crimes through r programming language, Science 132 (2019) 696–705.

[9] E. Frank, M. Hall, A simple approach to ordinal classification, in: European Conference on Machine Learning, Springer, pp. 145–156.

[10] J.L. González-Álvarez, J.J. López-Ossorio, C. Urruela, M. Rodríguez-Díaz, Integral monitoring system in cases of gender violence viogén system, Behavior & Law Journal (2018) 4 (2018).

[11] L.M. Graham, K.M. Sahay, C.F. Rizo, J.T. Messing, R.J. Macy, The validity and reliability of available intimate partner homicide and reassault risk assessment tools: A review, Trauma, violence, & abuse (2019) (2019).

[12] G.Shobha, S. Rangaswamy, Machine learning, in: Elsevier (Ed.), Handbook of Statistics, volume 38, V.N. Gudivada and C.R. Rao, 2018, pp. 197–228.

[13] N.Z. Hilton, G.T. Harris, M.E. Rice, C. Lang, C.A. Cormier, K.J. Lines, A brief actuarial assessment for the prediction of wife assault recidivism: the ontario domestic assault risk assessment., Psychological assessment 16 (2004) 267.

[14] N.Z. Hilton, A.T. Pham, S. Jung, K. Nunes, L. Ennis, Risk scores and reliability of the sara, sara-v3, b-safer, and odara among intimate partner violence (ipv) cases referred for threat assessment, Police Practice and Research (2020) 1–16.

[15] C. Kadar, R. Maculan, S. Feuerriegel, Public decision support for low population density areas: An imbalance-aware hyper-ensemble for spatio-temporal crime prediction, Decision Support Systems 119 (2019) 107–117.

[16] P.R. Kropp, S.D. Hart, The spousal assault risk assessment (sara) guide: Reliability and validity in adult male offenders, Law and human behavior 24 (2000) 101–118.

[17] F. Liberatore, L. Quijano-Sanchez, M. Camacho-Collados, Applications of data science in policing, European Law Enforcement Research Bulletin (2019) 89–96.

[18] J.J. López-Ossorio, Construcción y validación de los formularios de valoración policial del riesgo de reincidencia y violencia grave contra la pareja (VPR4. 0-VPER4. 0), Ph.D. thesis, UAM, 2017.

[19] J.J. López-Ossorio, J.L. González-Álvarez, I. Loinaz, A. Martínez-Martínez, D. Pineda, Intimate partner homicide risk assessment by police in spain: The dual protocol vpr5. 0-h, Psychosocial Intervention (2020) (2020).

[20] J.J. López-Ossorio, J.L. González-Álvarez, J.M.M. Vicente, C.U. Cortés, A. Andrés-Pueyo, Validation and calibration of the spanish police intimate partner violence risk assessment system (viogén), Journal of police and criminal psychology 34 (2019) 439–449.

[21] T.E. McEwan, D.E. Shea, J.R. Ogloff, The development of the vp-safvr: an actuarial instrument for police triage of australian family violence reports, Criminal justice and behavior 46 (2019) 590–607.

[22] J.T. Messing, J. Campbell, J. Sullivan Wilson, S. Brown, B. Patchell, The lethality screen: the predictive validity of an intimate partner violence risk assessment for use by first responders, Journal of interpersonal violence 32 (2017) 205–226.

[23] J.T. Messing, J. Thaller, Intimate partner violence risk assessment: A primer for social workers, The British Journal of Social Work 45 (2015) 1804–1820.

[24] R. Petering, M.Y. Um, N.A. Fard, N. Tavabi, R. Kumari, S.N. Gilani, Artificial intelligence to predict intimate partner violence perpetration, Artificial intelligence and social work (2018) 195.

[25] C. Quaresma, Violência doméstica: da participação da ocorrência à investigação criminal, Lisboa: Instituto Superior de Ciências Sociais e Políticas (2012).

[26] D.L. Radatz, N.Z. Hilton, Determining batterer intervention program treatment intensities: an illustration using the ontario domestic assault risk assessment, Partner abuse 10 (2019) 269–282.

[27] J. Snoek, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, arXiv preprint arXiv:1206.2944 (2012).

[28] Spanish-Government, Spain, Organic Law 1/2004, of December 28, on Comprehensive Protection Measures against Gender Violence, State Agency Official State Bulletin (Agencia Estatal Boletín Oficial del Estado), 2004.

[29] R. Stansfield, K.R. Williams, Predicting family violence recidivism using the dvsi-r: Integrating survival analysis and perpetrator characteristics, Criminal Justice and Behavior 41 (2014) 163–180.

[30] UNODC, Global study on homicide (2019): Gender-related killing of women and girls. (2019).

[31] S. Vail, C. Corradi, M. Naudi, Femicide Across Europe: Theory, Research and Prevention, Policy Press, 2018.

[32] K.R. Williams, S.R. Grant, Empirically examining the risk of intimate partner violence: The revised domestic violence screening instrument (dvsi-r), Public Health Reports 121 (2006) 400–408.

[33] World-Health-Organization, et al., Violence against women: a global health problem of epidemic proportions, Geneva: WHO (2014) (2014).

Appendix A

In the following Tables we describe the variables coded (using one-hot encoding) in this paper's models and represented by vector $x = (x_{fa}, x_{gi}, x_{ch})$. Note that for each possible answer, the last option (mainly DontKnow or No) is never coded as it is taken as the default option. Multiple choices are encoded using dummies. Table A.1 describes the variables that correspond to answers in VPER forms, i.e. x_{fa}. Also, for each variable in the table two extra variables are coded, i.e Increment and Decrement in the variable with respect to the last questionnaire. As mentioned in Section 3.1.3, this is done to the reflect each case's evolution, and completes the rest of the variables described in Section 3.1.3 corresponding to "Case History", i.e. x_{ch}. Finally, the "Case General Information" exogenous variables, x_{gi}, are described in Table A.2.

Question	Possible Answers

1 Has there been any kind of violence by the aggressor	Yes/No
1.1 Humiliation, insults	Yes/No/DontKnow
1.1.a Severity level	Slight/Serious/Very serious
1.2 Physical violence	Yes/No/DontKnow
1.2.a Severity level	Slight/Serious/Very serious
1.3 Sexual violence	Yes/No/DontKnow
1.3.a Severity level	Slight/Serious/Very serious
1.4 ¿Has there been a defensive reaction from the victim to the attack?	Yes/No/DontKnow
2 Has the aggressor used weapons or objects against the victim?	Yes/ No
2.1 The aggressor employed	White-weapon/Firearm/Other
2.2 Does the aggressor have access to firearms?	Yes/No/DontKnow
3 Does the victim receive or has he received threats or plans aimed at causing physical / psychological harm?	Yes/ No/DontKnow
3.1 Severity level	Slight/ Serious/ Very serious
3.2 Types of threats	AggressorSuicide/Economic/Death/ Reputation/ChildrenIntegrityOrCustody
4 Non-compliance with precautionary judicial provisions or violation of penalties or criminal security measures since the last assessment	Yes/No
4.1 The aggressor has contacted the victim online	Yes/No
4.2 The aggressor has contacted the victim through third parties	Yes/No
4.3 The aggressor has approached the victim	Yes/No
5 Exaggerated jealousy, control, or bullying in the past 6 months	Yes/ No/DontKnow
5.1 The aggressor shows exaggerated jealousy about the victim or has suspicions of infidelity	Yes/ No/DontKnow
5.2 The aggressor shows control behaviors over the victim	Yes/ No/DontKnow
5.2.a Types of behaviours	Physical/Psychological/ social/Labor/Economic/Cybernetic
5.3 The aggressor shows harassing behaviors on the victim	Yes/No/DontKnow
6 The aggressor is on the run or missing	Yes/No
7 Evidence of behavior by the aggressor since the last assessment	
7.1 Has distanced himself from the victim	Yes/No
7.2 Shows a peaceful attitude, assumes their situation with respect to the victim, without the intention of revenge against her or her environment	Yes/No
7.3 Shows a respectful attitude towards the law and collaboration with the agents	Yes/No
7.3 Show regret	Yes/No/DontKnow
7.4 Avails itself of aid programs	Yes/No/DontKnow
7.5 Complies with the regime of separation and family charges	Yes/No/NotApplicable
8 Does the agressor have a criminal or police record?	Yes or No
8.1 There are previous violations (precautionary or criminal measures)	Yes/ No/DontKnow
8.2 There is a history of physical or sexual assault	Yes/ No/DontKnow
8.3 There is a history of gender violence against other victims	Yes/ No/DontKnow
9 Are any of these circumstances currently present in the aggressor?	
9.1 has a diagnosed mental and / or psychiatric disorder	Yes/No/DontKnow
9.2 shows suicide attempts or thoughts	Yes/ No/DontKnow
9.3 suffers from some type of addiction (abuse of alcohol, psychopharmaceuticals or narcotic substances)	Yes/ No/DontKnow
10 Factors of vulnerability of the victim. Does any of these circumstances currently exist in the victim?	
10.1 Disability	Yes/No/DontKnow
10.2 In gestation period	Yes/No/DontKnow
10.3 Serious illness	Yes/No/DontKnow
10.4 Lacks favorable family or social support	Yes/No/DontKnow
10.5 Diagnosed mental or psychiatric disorder	Yes/No/DontKnow
10.6 Shows suicidal thoughts or attempts	Yes/No/DontKnow
10.7 Addiction	Yes/No/DontKnow
11 the victim hinders police or judicial actions	Yes/No
11.1 has resumed cohabitation with the aggressor while a measure of removal is in force	Yes/No

11.2 does not declare about reportable episodes, or if it has, subsequently expresses wishes to withdraw the report or refuse protection	Yes/No
11.3 carries out activities that go against their own safety (encounters with the aggressor, refuses or leaves the foster home, etc.)	Yes/No
12 Since the last assessment, have any of the following events occurred?	
12.1 The victim is financially dependent on the aggressor	Yes/No
12.2 The victim has minors or dependents	Yes/No
12.3 Legal proceedings for separation or divorce, unwanted by the aggressor	Yes/No
12.4 the victim establishes a new romantic relationship, not accepted by the aggressor	Yes/No
12.5 The aggressor establishes a new romantic relationship	Yes/No/DontKnow
12.6 The aggressor has a stable employment and economic situation	Yes/No/DontKnow
12.7 The aggressor has social support and favorable to his reintegration	Yes/No/DontKnow
12.8 There are conflicts because of their children	Yes/No/NotApplicable
13 The victim considers her current risk level to be	Unappreciated/Low/High
13.1 Do you agree with the risk appreciated by the victim?	Overestimates/Underestimates/Equal

Table A.1: VPER form varibles.

Question	**Possible Answers**
Age Victim	Ranges: 16-20,..,56-60,61-65,66-70,71-75,..,89-90
Age Author	Ranges: 16-20,..,56,-60,61-65,66-70,71-75,..,89-90
Institution	LocalPolice/ ForalPolice/NationalPolice/ CivilGuard
Locality's Population	Numeric
Normalized Locality's Population	[0-1]
Locality's Size	isTown/isSmallCity/isMediumCity/isBigCity
Is outside Peninsula	Yes/No
Province's Population	Numeric
Normalized Province's Population	[0-1]
Province's Size	isSmallProv/isMediumProv/isBigProv

Table A.2: Case General Information variables.