Knowledge-Based Systems 260 (2023) 110130

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Hybrid machine learning methods for risk assessment in gender-based crime

Ángel González-Prieto [a,b,*], Antonio Brú [a], Juan Carlos Nuño [c], José Luis González-Álvarez [d,e]

[a] Facultad de Ciencias Matemáticas, Universidad Complutense de Madrid, Plaza Ciencias 3, Madrid, 28040, Spain
[b] Instituto de Ciencias Matemáticas (CSIC-UAM-UC3M-UCM), C. Nicolás Cabrera 15, Madrid, 28049, Spain
[c] Departamento de Matemática Aplicada, ETSI de Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, C. José Antonio Novais 10, Madrid, 28040, Spain
[d] Gabinete de Coordinación y Estudios, Secretaría de Estado de Seguridad, C. Amador de los Ríos 2, Madrid, 28010, Spain
[e] Instituto de Ciencias Forenses y de la Seguridad (ICFS), Universidad Autónoma de Madrid, C. Francisco Tomás y Valiente 11, Madrid, 28049, Spain

ARTICLE INFO

Article history:
Received 24 February 2022
Received in revised form 8 November 2022
Accepted 14 November 2022
Available online 21 November 2022

Keywords:
Gender-based crime
Hybrid models
Quality measures
Risk assessment
Machine learning

ABSTRACT

Gender-based crime is one of the most concerning scourges of contemporary society, and governments worldwide have invested lots of economic and human resources to foretell their occurrence and anticipate the aggressions. In this work, we propose to apply Machine Learning (ML) techniques to create models that accurately predict the recidivism risk of a gender-violence offender. We feed the model with data extracted from the official Spanish VioGen system and comprising more than 40,000 reports of gender violence. To evaluate the performance, two new quality measures are proposed to assess the effective police protection that a model supplies and the overload in the invested resources that it generates. The empirical results show a clear outperformance of the ML-centered approach, with an improvement of up to a 25% with respect to the preexisting risk assessment system. Additionally, we propose a hybrid model that combines the statistical prediction methods with the ML method, permitting authorities to implement a smooth transition from the preexisting model to the ML-based model. To the best of our knowledge, this is the first work that achieves an effective ML-based prediction for this type of crimes against an official dataset.

1. Introduction

Gender-based crime is one of the most concerning scourges of contemporary society and to foresee its occurrence appears as a top priority for governments worldwide. Even though anticipating aggressions in a gender-based crime is impossible with the current technology, it is feasible for instance to define some indicators assessing the risk of recidivism.

Due to the specificities of this kind of crimes, the victims have a high chance of been re-offended if no protection is provided. In this way, a model able to analyze, after the first occurrence of violence, the features of both aggressor and aggressed and the circumstances of the aggression is crucial to provide effective police protection to the victims. With a theoretical framework like this, it is possible to reach an accurate assessment of the risk of recidivism. This is particularly important since, in practice, the police resources are limited and there must exist a system to categorize and prioritize the cases under surveillance.

In this work, we import cutting-edge computational methodologies of Machine Learning (ML) to address the problem of prediction of recidivism in gender based-crimes. We shall focus on the particular case of Intimate Partner Violence Against Women (IPVAW), as defined by the World Health Organization (WHO) [1]. This definition excludes other offenses related to gender such as stalkings or rapes, which are crimes with a very different nature, more macroscopic and less victim-focused.

Given a large amount of structured data about IPVAW cases, we will apply ML techniques to develop novel models of risk assessment of recidivism of a victim, understood as the probability that a female victim, who has been offended and has reported her case, is aggressed again. In our case, the data will be provided by the Spanish VioGen system, a governmental program for tracking and controlling gender violence [2,3], but the approach and applied methods are general and can be straightforwardly translated to other data sources. Since its inception in 2007, this system implements a recidivism risk assessment method based on classical statistics and psico-metric analysis that has led to an outstanding reduction in the number and severity of reported gender-based crimes in Spain [4].

* Corresponding author at: Facultad de Ciencias Matemáticas, Universidad Complutense de Madrid, Plaza Ciencias 3, Madrid, 28040, Spain.
E-mail address: angelgonzalezprieto@ucm.es (Á. González-Prieto).

https://doi.org/10.1016/j.knosys.2022.110130

To empirically validate these ML-based models, we conducted several experiments. As it will be evidenced by these results, the Nearest Centroid (NC) model clearly outperforms the other risk assessment methods, including the pre-existing VioGen system. NC is a powerful similarity-based classification method that has been successfully applied in cancer class prediction from gene expression profiling [5] and in text classification [6], where it is usually known as the Rocchio classifier. In our scenario, the NC algorithm seeks to extract the main features of each of the archetypes of aggressors and, using them, analyzes new cases by computing the similarity to each of these general patterns.

In some sense, the operation of the NC model is analogue to some criminalistic methods [7,8], but the large amount of data and the variety of responses allows the ML to extract very subtle information which cannot be obtained via classical methods. A rough idea of the operation procedure of the NC algorithm is shown in Fig. 1. During the initialization phase, shown in the upper-most picture, the model is only set up with the number of categories into which the cases must be classified (three in this case) but no further information is provided. In the training step, shown in the second picture of Fig. 1, the system is fed with a large number of examples for which the intended classification is known (or it may be calculable from other sources). In our case, the model analyzes the reports collected in the database and, according to the reported number of subsequent aggressions, classifies each case into 'No' recidivism, 'Low' risk of recidivism or 'High' risk (represented in this figure by green, blue, and red, respectively). Using these data, the system computes archetypal profiles of aggressions taking into account the provided features of the cases regarding the characteristics of the aggressor and of the victim, as well as the surrounding circumstances and aggravating factors. In NC these profiles are extracted by computing the center-of-mass (the centroids) of each of the clusters according to some metric, typically the L^p-distance for some $p \geq 1$. An iterative shrinking process can also be applied so that the features which are close to the global centroid of the dataset are pulled apart. In this sense, the NC training also encompasses a feature selection procedure through which the most relevant factors for recidivism are identified and extracted.

After this training process, the system has identified the profiles of each type of case and is ready to be used for production, as shown in the third level of Fig. 1. During the exploitation phase, as shown in the forth level of Fig. 1, the system is able to classify accurately new cases of aggressions, even though very limited information is known, according to the similarity of the features of the analyzed case to each of the archetypal profiles. As by-product of this analysis, the model issues a prediction of the forecasted risk of recidivism ('No', 'Low' or 'High'). This risk assessment can be used by the police forces to gauge the preventive measures to be provided to the victim to avoid further offenses.

To test these ML-based methods, in this work we use as baseline for these experiments the preexisting risk assessment model. To this aim, we introduce two novel quality measures to evaluate the effectiveness of a prediction model: (1) provided police protection, which quantifies the extent to which the threaded victims are protected; and (2) police resources overload, which regards the amount of resources that are wasted due to an overcautious prediction of the model. A large value of the former metric points out to highly effective predictions and a low value of the later indicates an optimal use of the resources available.

The empirical results show that the ML-based method clearly outperforms the preexisting model in terms of police protection up to a 25% and, depending on the features of the police system, it may also improve the usage of the resources or lead to a slight overrun. In order to migrate from the existing system

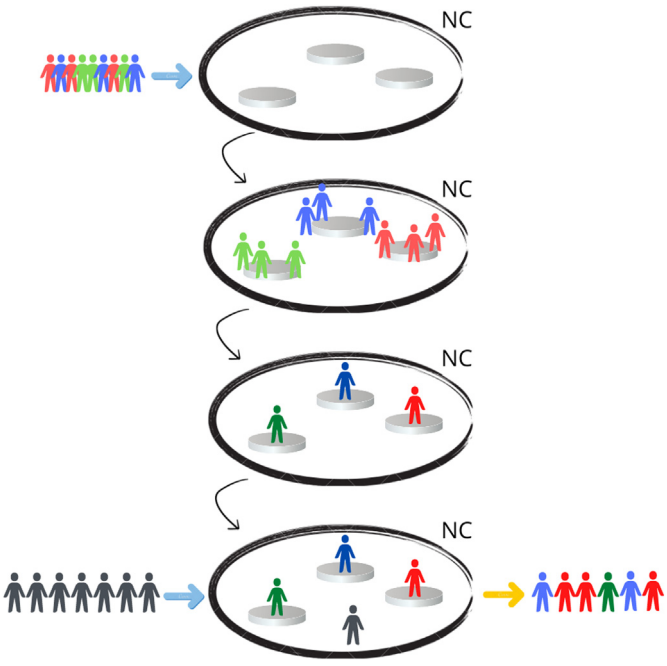

Fig. 1. Abstract representation of the Nearest Centroid (NC) prediction method. Each level in the plot corresponds to a phase of the process. From top to bottom: initialization, training, system ready and exploitation.

to the ML-based method, which may awaken initial reluctance, we also propose a hybrid model that allows decision-makers to implement a smooth progressive transition, avoiding a drastic change of paradigm. In this spirit, the aim of this work is not to radically substitute the existing risk prediction models, but to complement, extend, and refine their predictions to detect new hidden recidivist cases.

We hope that this work will have a crucial impact in the reduction of recidivism in gender violence. Only through a hand-to-hand collaboration between human and computational forces, mankind will be able to eradicate this dangerous scourge. In our opinion, the proposed method meets all the requirements to make the first move towards this important goal.

1.1. State of the art

Despite the apparent regular occurrence of crime, as it was already recognized in the 19th century [9,10], it has defied the predictability provided by mathematical modeling in the Natural Sciences. Surprisingly, it is easier to accurately predict where a rocket will be after its launch on its way to a distant planet than to foresee the next victim of an offense. The unpredictable nature of crime arises the question of whether the classical scientific method can be a solving tool instead of only a descriptive framework. The Minority Report fiction of Philip K. Dick [11] suggests that anticipating a crime is possible and discusses some consequences derived from this possibility.

In the real world, criminality can be handled by using mathematical and computational models [12,13]. Classical modeling has proven to be very accurate for some crimes with a noticeable spatial component, such as juvenile delinquency [14,15] or burglary occurrence [16–18]. However, this kind of models seems to be less adequate to study other kinds of crimes that do not present a noticeable spatial component, as it is the case of domestic or gender-based violence. Instead, new computational methodologies under the name of ML, a subfield of Artificial Intelligence (AI) [19], appear as a powerful alternative [20–24]. This approach

focuses mainly on the development of computational algorithms to solve complex problems from the data [25].

The consideration of gender-based crime as a singular case of offense is presented in many countries around the world. The USA legislation considers gender-based violence with its specificities but, nonetheless, a particularly focused tackling system does not exist. Along the last decades, several surveys have been released to get insight into the prevalence of different forms of violent crime in the USA and, in particular, they include information about the sex of the victims [26].

Across the Atlantic, most of the European countries have recognized the gender nature of domestic violence in the Istanbul Convention [27]. The signatory governments of this protocol have subsequently adjusted their policies to fulfill the Convention's requirements with adapted methods of fight and prevention. In this sense, the UK deserves special mention since they did not ratify Convention's protocols. Instead, it has taken particular measures to fight against gender-based violence as a specific crime. Data obtained from the Crime Survey for England and Wales (CSEW), and published by the Office for National Statistics (ONS), concerning the gender component of violent crimes has been analyzed to answer the question whether gender-based crime is falling or not in the last times [28]. Additionally, the UK has implemented its own model for fighting gender crime, the so-called DASH (Domestic Abuse, Stalking and Honour Based Violence) system [29]. The data gathered in this system has been analyzed in a recent paper using ML techniques [24]. The authors conclude that standard ML models are not suitable for risk assessment.

Another case of a specific prediction system for this kind of aggressions is the Israeli Spouse Violence Risk Assessment Inventory (SVRA-I), a system able to measure the likelihood of male perpetrators repeating violent behavior against their partners. The performance of this model has been evaluated in [30], but the analysis involved only 206 cases and the diagnosis was contrasted against the opinion of a board of experts. In this way, no analysis of the real outcome of the measures to prevent further aggressions in practice was conducted.

Apart from the predictive aspect of this phenomenon, the problem of detecting and avoiding gender-based crimes has been addressed from multiples viewpoints, such as the sociological perspective in [31] for 187 cases in Portugal, the point of view of social workers in [32] to identify problematic cases, or the clinical aspect [33].

Regarding the application of AI-based techniques to this problem, one of the major advances is [34], in which the authors analyzed the accuracy of ML models to predict the outcome of 452 cases occurred in Los Angeles. The experiments show that Neural Networks and Linear Regression achieved the best results in terms of accuracy and comparison with the baseline. However, the sample size is very small and the baseline is again obtained as a naïve regression in terms of basic features of the case, not analyzing the real outcome.

In the case of Spain, as analyzed in our work, there exists a dedicated governmental program for tracking and controlling gender violence, the so-called VioGen system [2,3]. This system included a predictive model based on classical statistics and psico-metric analysis, whose effectiveness to produce a remarkable drop in the number of cases has been verified several times [4,35,36]. In the latest times, some works have pointed out the necessity of applying AI techniques to this database, but the analysis has been limited to understand the influence of external factors to the risk prediction [37].

In this way, a thorough analysis of the influence of ML-based techniques in the prediction of gender-based crimes is missing in the literature. Despite several works have pointed out that this might be a promising line, these works lack of (i) large and representative sample size, (ii) the use of official databases provided by governmental protection programs, and (iii) a ground-truth based on the actual outcome of the cases and not the opinion of experts or naïve models. In this work, we aim to address these problems to obtain the first results pointing out that ML are able to provided accurate predictions for a large corpus obtained from a large official dataset, in terms both of foreseeing the actual outcome of the case and improving the pre-existing deeply tested risk-assessment model.

1.2. Our contribution

The main achievements of the present work can be summarized as follows:

1. To the best of our knowledge, this is the first work that shows that an official dataset collected by the police forces with real data can be used to effectively predict recidivism in gender-based crimes. The data used in this work is not part of an artificial dataset created from unofficial statistics, as the ones that can be publicly found. Instead, the data used was collected by the Spanish police forces through their official protocol against gender-based crime. None of the previous works was able to extract this information to provide effective predictions, and there actually are some no-go results against this possibility for other official datasets.
2. According to the experiments run, the proposed approach clearly outperforms the existing solution implemented by the Spanish police forces, which acts as baseline for this work. The police protection provided by the proposed model improves the current protection offered up to a 25%, with an exploitation of police resources that varies from a mild overload to a better use, depending on the nature of the penalty of the protection policies when a wrong prediction occurs.
3. Two new quality metrics are introduced adapted to the crime prediction problem, the so-called "police protection" and "resources overload" metrics. They are constructed to effectively capture the subtleties of crime prediction, which must combine two opposed objectives: to provide as much protection as possible and to not overload the police resources.
4. This work introduces a novel solution based on random walks to enable a smooth transition from the currently implemented statistics-based solution to our ML-based solution. This allows decision-makers to gauge the weight of the ML in the final decision so that the proposed solution can be progressively tested in-field without a drastic change. Our experiments show that, with this new hybrid model, the more prevalent the ML-based solution, the better the police protection provided.
5. Our results evidence that, for human-centered problems such as the gender-based crimes, simple ML models may be more effective than other more complicated solutions, in sharp contrast with the folklore in the ML literature. In our case, the best model for the recidivism prediction problem turns to be NC, which seeks to extract the essence of the profiles of each of the types of aggression, in contrast with other more sophisticated models like Random Forest (RF), XGBoosting or Neural Network (NN).
6. The gender-based recidivism prediction problem suffers a phenomenon of cause–effect inversions: As soon as we intervene to prevent a re-aggression, we are changing the outcome of the case so we cannot effectively determine

whether the taken measures were actually necessary. Even under this uncertainty that turns the prediction into a somehow semi-unsupervised problem, the proposed methods are able to achieve very compelling results that outperform the existing solutions.

2. Problem setting

The main objective of this work is to predict the probability of recidivism in gender-based crimes, that is, the probability that an aggressor who has exerted violence against a victim repeats any violence episode. In particular, we shall focus on the so-called Intimate Partner Violence Against Women (IPVAW), that is, a gender-based aggression in which the aggressor is a male who has exerted violence against a female victim that is also his sentimental partner (current or past).

To analyze the case, we will suppose that there are available several indicators that characterize the gender-based violence case. These features can be provided by the victim itself, or through a report to the authorities of an initial violent episode. This later scenario is the one that corresponds to the dataset analyzed in this work, but other frameworks can be addressed similarly. To be precise, we will suppose that our data source provides us, for each closed case of aggression, the circumstances that characterize the violence, understood as a combination of:

- The victim and aggressor personal profiles, such as the victim's vulnerability factors or the aggressor's police record.
- The circumstances that surrounded the first violent episode, such as the existence of physical or verbal abuse and the use of weapons.
- The socio-economical characteristics of the partner, such as the economical dependence of the victim with respect to the aggressor or the existence of children in common.
- Potential aggravating factors, such as previous reports of violence from the aggressor to other victims.

In our case, the circumstances of the crimes and the number of recidivisms were extracted from the official Spanish VioGen database with more than 44,000 real entries of gender-based crimes corresponding to the cases occurred in Spain between 1st October 2016 and 1st October 2017 (for a more detailed description, please refer to Appendix B).

The goal of the system is to predict the probability of a victim of suffering a re-aggression from the same aggressor. To analyze this risk, we set three levels of *recidivism risk*:

$$\Lambda = \{\text{No}, \text{Low}, \text{High}\}\,.$$

The meaning of these labels is as follows:

- **No**: The victim will suffer no further aggressions.
- **Low**: The victim will suffer one or two more attacks.
- **High**: The victim will suffer three or more offenses of violence after the first report.

It is worth mentioning that the threshold of 3 cases to distinguish between 'Low' risk cases and 'High' risk cases may seem rather arbitrary. To clarify its dependency, we have conducted an analysis of sensitivity for this threshold. The results showed that setting the threshold to 4 or 5 further aggressions leads to small deviations with respect to the original limit of 3. Thereby, these results evidence that the upcoming discussion is robust under small variations of this threshold.

In this manner, the aim of the system is the following: given the circumstances that characterize a IPVAW case, we seek to predict the recidivism risk $\lambda \in \Lambda$ that characterizes the number of subsequent aggressions. To train the system to extract the patterns that characterize the existence of re-aggressions, we shall suppose that our data source also provides us with a collection of closed cases of aggression. For these cases, not only the circumstances that surrounded the case are known, but also the number of subsequent aggressions that the victim suffered after the first report. In this way, to each closed aggression case, we can assign a category $\lambda \in \Lambda$ according to the number of subsequent episodes of violence reported. This collection of closed cases with a real risk prediction will be used as the training set for the model.

3. Prediction models

Throughout this work, the prediction of the recidivism risk will be treated as a classification problem. To fix the notation, let us briefly review the main ideas of a classification task. The aim of a classification problem is to understand a very complex phenomenon that assigns to each individual a certain *label* or *category*. More precisely, an individual is represented as a point $x \in \mathbb{R}^d$ for some fixed $d > 0$ (the dimensionality of the data) so that each of the components of x should be understood as interesting *features* of the individual. In our case, the vector x is the result of the encoding the characteristics of the violence case, including the circumstances of the first report of violence. For those circumstances with nominal values, any numerical encoding technique can be applied, such as one-hot encoding.

Additionally, in a classification problem there is a finite set Λ of labels in such a way that each individual is assigned to an element of Λ. In our case, $\Lambda = \{\text{No}, \text{Low}, \text{High}\}$ is the set of possible recidivisms risks, as explained in Section 2. In other words, there is a function

$$f : \mathbb{R}^d \to \Lambda$$

such that, for each individual $x \in \mathbb{R}^d$, it assigns $f(x) \in \Lambda$, the category of x.

The key problem is that, in general, this labeling function f is completely unknown or the phenomenon under study is so complex that the assignment f is intrinsically fuzzy. The aim of a classification model is thus to provide a reasonable ansatz

$$\hat{f} : \mathbb{R}^d \to \Lambda$$

such that $\hat{f}$ is as similar to the real assignment f as possible.

In the following, we propose models $\hat{f}$ for automatic violence prediction so that, when a new case x^{new} appears, the issued assessment $\hat{f}(x^{\text{new}}) \in \Lambda$ can be used as prediction to determine the measures of police protection to be taken. To evaluate these models, we will assume that the data source also provides a model of preexisting risk assessment (typically of statistical nature). In the particular case of VioGen, it implements a risk assessment model based on psychometric criteria (see Appendix A for a more detailed description). This preexisting risk model will be used as baseline for the experimental setting.

The methodology applied in this work to analyze the performance of the proposed prediction models is the following:

1. A collection of ML-based prediction systems will be selected as representatives of the different techniques in supervised ML applicable to this problem. A theoretical discussion of these models and operation procedure is provided in Section 3.1.
2. Two quality measures are designed to evaluate the performance of the selected models, enabling their mutual comparison and with the baseline (the preexisting VioGen risk assessment model). These measures are based in the estimation of the police protection provided (Section 3.2.1) and the overload in the police resources that the implementation of this solution produces (Section 3.2.2).

3. A thorough hyper-parameter tuning is conducted for each of the ML methods to improve the F_1-score and the police protection provided by their predictions (Section 4.1).
4. Based on the obtained results after hyper-parameter tuning, the best ML model is chosen and retrained to compete with the preexisting baseline. In Section 4.2, the results are analyzed to determine whether ML-based solutions can outperform the baseline.
5. To enable a smooth transition from the existing model to the ML prediction system, a stochastic hybrid model is designed that allows decision-makers to tune the impact of the ML model in the final decision. The theoretical description of this hybrid method is provided in Section 3.3.
6. A thorough testing of the evolution of the quality measures designed (police protection and resources overload) is conducted to analyze the impact basing the decisions on the ML model with respect to the baseline. The trend of these quality measure when we vary the importance of the ML model in the final decision is discussed in Section 4.3.

3.1. Supervised learning models

The proposal of this work is to construct prediction models for recidivism prediction through ML. These models are designed to automatically distil knowledge from the previously reported datxa. For the convenience of the reader, we sketch the main features of the ML-based models.

A ML model is fed with a finite set $D = \{x_1, \dots, x_N\}$, usually referred to as the *dataset*, as well as their real categories $y_i = f(x_i)$ for $1 \leq i \leq N$. In our setting, recall that the instances x_i are the 'circumstances' that characterize the violence case, whereas $y_i \in \Lambda$ is the true recidivism risk computed from the number of subsequent aggressions.

Using this dataset, the system seeks to a function $\hat{f} : \mathbb{R}^d \to \Lambda$ such that it minimizes an error function $\mathcal{E}(f)$. Several proposals exist in the literature. A common choice for this error function takes the form

$$\mathcal{E}(f) = \frac{1}{N}\sum_{i=1}^{N} \delta(\hat{f}(x_i), y_i).$$

Here, $\delta : \Lambda \times \Lambda \to [0, \infty)$ is a pre-fixed distance on the set of labels. If δ is taken to be the discrete metric ($\delta(y, y) = 0$ and $\delta(y, y') = 1$ if $y \neq y'$), then $\mathcal{E}(f)$ is the ratio of misclassified instances.

Particularly, if we focus our attention on a parametrized family of functions f_θ for parameters $\theta \in \mathbb{R}^m$, the error function becomes a function of θ,

$$\mathcal{E}(\theta) = \mathcal{E}(f_\theta) : \mathbb{R}^m \to \mathbb{R}.$$

In this context, the search of the best model reduces to an optimization problem on θ, a process called the *training* in the ML jargon. Several optimization algorithms may be applied for minimizing $\mathcal{E}$, like gradient descent, linear/quadratic programming methods, or genetic programming approaches.

Typically, the chosen optimization method requires to fix, previously to the training, some values that determine the concrete implementation of the optimization procedure (e.g. the step in the gradient descent method). These values, and any other parameters that must be fixed beforehand, are called the *hyper-parameters*, since they are at a higher level than the system parameters: they tune the model and the method applied to seek the best parameters. No general-purpose hyper-parameter optimization method is known, so experiments must be carried out by conducting an exhaustive search for the optimal hyper-parameters.

3.2. Quality measures for risk assessment

In this section, we address the problem of evaluating the quality of the risk assessments issued by prediction models. Notice that the problem of forecasting recidivism risk in crimes presents some peculiarities that turn the classical metrics unsuitable for this purpose.

- The usual coefficients used for evaluating classification models, like precision and recall, do not fully capture the complexity of this kind of predictions. This is not an issue of minimizing false positives or negatives, but something subtler: we want to minimize the recidivism without misspending police resources. Hence, any quality measure capturing the police protection provided must combine the two opposed objectives that play a role, namely, minimization of the false negatives in highly risky cases and minimization of the false positive rate in nonrisky cases.
- The amount of police resources spent with a prediction model is not absolute, but it depends on the nature of the police protection protocol. For instance, if the protocol prescribes very invasive protection measures for highly risky cases, like constant surveillance or re-allocation of the victim, then the cost of misclassifying a nonrisky case as risky is very large in terms of resource overload. However, if the police protocol only provides mild protection, like punctual surveillance of telephonic follow-up, then the cost of misclassification is much lower. In this way, any quality measure that analyzes this resources overload must depend on a parameter τ that gauges the 'penalty' of misclassifying according to the police protocol.

To overcome these difficulties, in this section we introduce two novel metrics for analyzing the quality of these risk assessments, called 'police protection' and 'resource overload'. These metrics will be later used in Section 4 to evaluate the quality of proposed ML-based prediction model against the existing risk assessment method.

3.2.1. Police protection

To analyze the performance of the proposed approach against the rule system-based previous risk prediction model, we will consider an auxiliary quality measure.

Two quality measures are standard for classification problems, the so-called precision and recall. Let us fix a classification model $\hat{f} : \mathbb{R}^d \to \Lambda$. Given a test set $x_1, \dots, x_M$, let us denote the output of the classification model as $\hat{f}(x_1), \dots, \hat{f}(x_N)$ whereas the real labels will be denoted by $y_1, \dots, y_M$. Fixed a class $\lambda \in \Lambda$, the *precision* and *recall* of $\hat{f}$ in the class λ are

$$\mathrm{Precision}_\lambda(\hat{f}) = \frac{\left|\left\{1 \leq i \leq M \mid \hat{f}(x_i) = \lambda \mathrm{and} y_i = \lambda\right\}\right|}{\left|\left\{1 \leq i \leq M \mid \hat{f}(x_i) = \lambda\right\}\right|},$$

$$\mathrm{Recall}_\lambda(\hat{f}) = \frac{\left|\left\{1 \leq i \leq M \mid \hat{f}(x_i) = \lambda \mathrm{and} y_i = \lambda\right\}\right|}{|\{1 \leq i \leq M \mid y_i = \lambda\}|},$$

where $|X|$ stands for the number of elements of the set X. In other words, $1 - \mathrm{Precision}_\lambda(\hat{f})$ is the rate of false positives and $1 - \mathrm{Recall}_\lambda(\hat{f})$ is the rate of false negatives of the class λ. In general, to combine both coefficients, it is customary to consider the F_1-*score* as the harmonic mean

$$F_1{-}\mathrm{score}_\lambda(\hat{f}) = 2\frac{\mathrm{Precision}_\lambda(\hat{f}) \cdot \mathrm{Recall}_\lambda(\hat{f})}{\mathrm{Precision}_\lambda(\hat{f}) + \mathrm{Recall}_\lambda(\hat{f})}.$$

Despite its wide range of applications, precision and recall are not fair quality measures in our case. A prediction model with a

good precision in the recidivism risk class 'High' but a bad recall might achieve an admissible F_1-score, but this is a very worthless model from the police viewpoint: it leaves most of the worst gender violence cases without police supervision. In the same vein, a bad precision in the 'No' class but a high recall is pointless, since the system is assigning no risk to very problematic cases.

To balance these opposed trends, we propose to use a novel measure of *police protection*. Given a model $\hat{f}$, it is defined as

$$\begin{aligned}\text{PoliceProtection}(\hat{f}) =& \text{Precision}_{\text{'No'}}(\hat{f}) + F_1\text{−score}_{\text{'Low'}}(\hat{f})\\ &+ \text{Recall}_{\text{'High'}}(\hat{f}).\end{aligned}$$

In this way, the higher the police protection, the better the model. Algorithms achieving a large police protection are able to identify very violent cases with a high risk of recidivism.

3.2.2. Resources overload

An increase in the provided police protection might be in conflict with the amount of police resources mobilized. Indeed, there is an obvious solution with perfect police protection, namely to assign to all the cases the maximum risk. Of course, the problem is that this model is overkilled: it is mobilizing resources to provide protection to cases that many times will not need such a deployment. This is not only an economic issue. Since the resources are limited, this indiscriminate use of resources would lead to a degradation of the quality of the service, providing worse surveillance to the most threatened victims.

To quantify the unnecessary amount of police resources mobilized, we propose the following index. Let $y_1, \dots, y_M$ be the real values of the recidivism risk and suppose that $\hat{f}(x_1), \dots, \hat{f}(x_M)$ are the issued predictions by a model $\hat{f}$. In addition, let us choose $\tau \geq 0$. Then, the *police resources overload* coefficient of $\hat{f}$ with penalty τ, denoted by $\text{PoliceResource}(\hat{f}; \tau)$, is just the weighted average of times $\hat{f}(x_i)$ is higher that y_i. Explicitly, it is given by

$$\begin{aligned}\text{PoliceResource}(\hat{f}; \tau) =& \frac{1}{2M(1+\tau)} \left(\left| \begin{Bmatrix} \hat{f}(x_i) = \text{'Low'} \\ \text{and} y_i = \text{'No'} \end{Bmatrix}_i \right| \right.\\ &+ \tau \left| \begin{Bmatrix} \hat{f}(x_i) = \text{'High'} \\ \text{and} y_i = \text{'Low'} \end{Bmatrix}_i \right|\\ &\left. + (1+\tau) \left| \begin{Bmatrix} \hat{f}(x_i) = \text{'High'} \\ \text{and} y_i = \text{'No'} \end{Bmatrix}_i \right| \right).\end{aligned}$$

The penalty τ should be understood as the extra overload that the police forces suffer when passing from a 'Low' risk surveillance to a 'High' risk surveillance. The particular value of τ depends on the police protocol applied, as well as the particular structure, organization, and role distribution of the police forces.

3.3. The hybrid model

Even though the ML-based models can reach very good results in terms of police protection, as it will be shown in Section 4, it seems reasonable that the police forces would be reluctant to drastically change the prediction method by a new untested one. Currently, VioGen is issuing fairly good predictions of risk and a drastic change could lead to a deterioration of the police efficacy if the model is not accurate in practice. In order to mitigate this reluctance, we propose to use a hybrid model between the preexisting risk assessment method and a ML-based model.

The proposed hybrid model $\hat{f}_\mu^{\text{Hy}}$ is a stochastic mixture between any two prediction models (in our case, the VioGen preexisting model and the chosen ML-based model). It depends on a real parameter $0 \leq \mu \leq 1$ in such a way that, for $\mu = 0$, the model $\hat{f}_0^{\text{Hy}}$ is the preexisting VioGen risk assessment method and, for $\mu = 1$, the model $\hat{f}_1^{\text{Hy}}$ is the ML algorithm. In the halfway, for an input $x \in \mathbb{R}^d$, the output $\hat{f}_\mu^{\text{Hy}}(x)$ is a random variable that takes values in the oriented segment starting at the prediction $\hat{f}_0^{\text{Hy}}(x)$ and finishing at $\hat{f}_1^{\text{Hy}}(x)$ with mean the μ fraction of the segment.

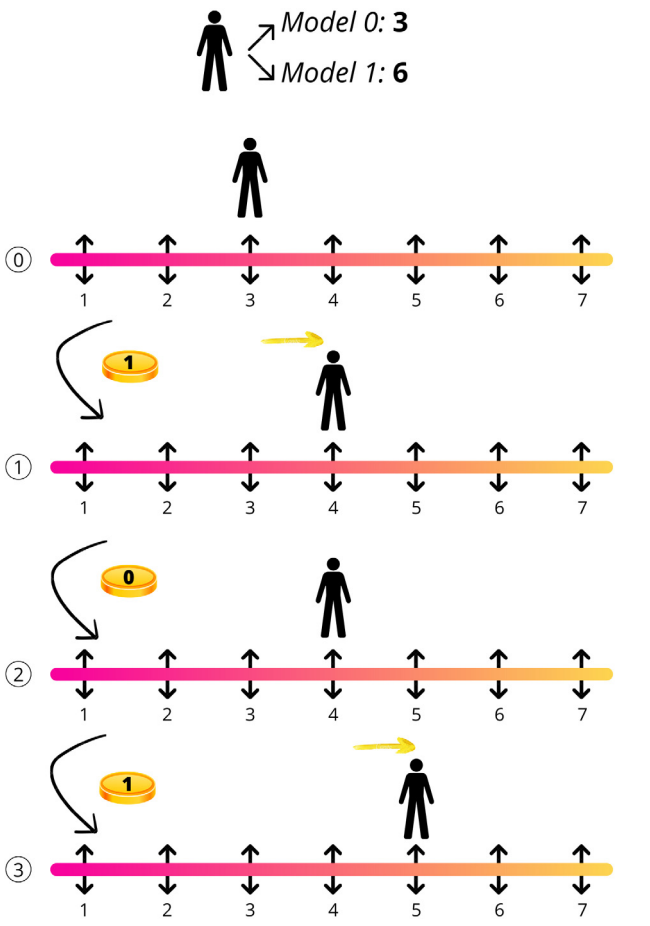

Fig. 2. Example of operation of the hybrid model. For the case x, the model 0 returned $\hat{f}_0(x) = 3$ and the model 1 returned $\hat{f}_1(x) = 6$. To conciliate both results, the hybrid model sets an initial prediction of $\hat{f}_0(x) = 3$ (step 0) and flips $|\rho| = |\hat{f}_0(x) - \hat{f}_1(x)| = 3$ biased coins, with probability of success (result 1) equal to μ. Each time that the coin returns 1, the prediction is increased one unit, and each time that the coin returns 0, the prediction remains unchanged. In this example, the first and third coins (steps 1 and 3) returned 1, but the second coin (step 0) returned 0. Hence, the final result of the hybrid model is $\hat{f}_\mu^{\text{Hy}}(x) = \hat{f}_0(x) + 2 = 5$.

To be precise, let us suppose that we have two prediction models with integral predictions $\hat{f}_0, \hat{f}_1 : \mathbb{R}^d \to \Lambda \subseteq \mathbb{Z}$. Additionally, suppose that Λ is closed under convex combinations with integer coefficients. Given $0 \leq \mu \leq 1$ and $n \geq 0$, let $X_\mu(n) \in \text{Bin}(n, \mu)$ be a binomial random variable with success probability μ and n trials i.e. $X_\mu(n) = \sum_{i=1}^n X_\mu^i$ where $X_\mu^i \sim \text{Ber}(\mu)$ are independent Bernoulli random variables with success probability μ.

The hybrid model $\hat{f}_\mu^{\text{Hy}}$ assigns to an example $x \in \mathbb{R}^d$ the value

$$\hat{f}_\mu^{\text{Hy}}(x) = \hat{f}_0(x) + \text{sign}(\rho) X_\mu(|\rho|),$$

where $\rho = \hat{f}_1(x) - \hat{f}_0(x)$. In other words, $\hat{f}_\mu^{\text{Hy}}(x)$ starts with the initial prediction $\hat{f}_0(x)$. If $\hat{f}_1(x) > \hat{f}_0(x)$, the model tosses $|\hat{f}_1(x) > \hat{f}_0(x)|$ coins with success probability μ and, for each success, the system increases prediction by one unit. On the other hand, if $\hat{f}_1(x) < \hat{f}_0(x)$, the system reduces the prediction one unit per success. If the predictions $\hat{f}_0(x)$ and $\hat{f}_1(x)$ agree, the result remains unchanged. The operation procedure of the hybrid model is illustrated in the example of Fig. 2.

Notice that, with this setup, for $\mu = 0$ we have $\hat{f}_0^{\text{Hy}} = \hat{f}_0$, and for $\mu = 1$ we get $\hat{f}_1^{\text{Hy}} = \hat{f}_1$. Hence, $\hat{f}_\mu^{\text{Hy}}$ can be seen as a stochastic interpolation between the two models or a one-sided random walk with finite length. An analogous construction can be carried out if $\Lambda \subseteq \mathbb{R}^n$ is a subset of a m-dimensional lattice, closed under convex combinations with integral coefficients, by running m parallel independent hybrid models.

In the case considered in this paper, we have codified numerically the recidivism risk labels in Λ (0 is 'No' recidivism risk, 1 is 'Low' risk, and 2 is 'High' risk). The model $\hat{f}_0$ is taken as the numerical outputs of the preexisting VioGen risk assessment through a fixed rule system and $\hat{f}_1$ is the output of the considered ML model.

4. Results

In this section, we show the results of a collection of experiments conducted to analyze the performance of the proposed ML models compared with existing risk assessment method. For this purpose, in Section 4.1 we shall analyze the performance of several ML methods for the classification task. From this analysis, NC emerges as the model that achieves the best result in terms of F_1-score and police protection. In Section 4.2 a deeper hyper-parameter tuning is conducted to obtain the best setup. With this optimal configuration, we build the hybrid model whose evolution of the police protection and resource overload metrics are analyzed in Section 4.3.

4.1. Model selection

We have considered as potential models several standard ML classification models: Decission Tree (DT) [38], Random Forest (RF) [39,40], K-Nearest Neighbors (KNN) [41], Gradient Boosting Classifier (GBC) (to be precise, in its XGBoost implementation) [42], Neural Network (NN) (to be precise, a feedforward multilayer perceptron) [43], and Nearest Centroid (NC) [5]. These models represent the state-of-the-art in multiclass classification methods. The rationale behind this choice is to select at least one algorithm from each of the main techniques in supervised learning: tree-based models (DT and RF), similarity-based models (KNN and NC), boosting algorithms (GBC and in some sense RF) and deep learning methods (NN). Hence, with this selection, we aim to cover a wide range of techniques that are typically applied in supervised prediction with proven success.

A fine hyper-parameter search has been conduced to select the best setting for each algorithm. Recall that these ML-based models may depend on some hyper-parameters $\hat{f} = \hat{f}_\vartheta$. A standard criterion is to seek to the best setup ϑ_0 such that $\vartheta_0 = \text{argmax}_\vartheta \left(F_1\text{−score}_\lambda(\hat{f}_\vartheta)\right)$. For that, a particular label λ of interest may be fixed or the weighted average of the F_1-scores among all the labels may be consider.

In our case, we shall focus on the label $\lambda =$'High' and we seek to maximize $\vartheta \mapsto F_1\text{−score}_{\text{'High'}}(\hat{f}_\vartheta)$. The dataset is randomly split into the training set (67%) and the test set (33%). For each model type and each possible combination of hyper-parameters, the model is trained with the training split and the quality score (F_1 for the class 'High') is computed for the predicted values on the test split.

The search has been performed through a grid search with the following combinations of hyper-parameters:

- DT: Criterion = Entropy, Gini; Splitter = Best, Random; Max depth = 5, 10, 50, 100, None.
- RF: Criterion = Entropy, Gini; N estimators = 1, 5, 10, 100, 500; Max depth = 5, 10, 50, 100, None.
- KNN: K = 2, 5, 10, 20, 50, 100, 200.
- GBC: Max depth = 1, 3, 5; η = 0.01, 0.1, 0.2; γ = 0, 1, 10; λ = 1, 2, 10.
- NN: Up to 3 layers with possible number of neurons = 5, 10, 50, 100; Dropout = 0, 0.1, 0.5, 1; Batch size = 10, 50,100.
- NC: Metric = Euclidean, Minkowski, Manhattan; Shrink threshold = 0.1, 0.5, 1, 10, 20, None.

The experiments were carried out by means of the Scikit-learn library [44], complemented with the XGBoost library [43] for experiments with GBC and Keras [45] for NN. For a precise description of the meaning of each of these hyper-parameters, please refer to the Scikit-learn library, XGBoost and Keras documentation.

From these experiments, we obtain the optimal models of each type, which are shown in Table 1. For each model, we show the F_1-score attained in the class 'High' (the objective function) as well as the weighted average of the F_1-score throughout all the three possible classes. The results evidence that the NC model clearly outperforms the other models in the achieved F_1-score for the class 'High'. The results for the weighted F_1 are similar in the six models (being slightly better in GBC). The scores obtained by the best models are consistent with the remaining top-three sets of hyper-parameters.

This superiority of the NC model also holds in terms of police protection level, as can be checked in the fifth column of Table 1. The police protection provided by the NC algorithm clearly outperforms the ones obtained by the other analyzed methods, with an improvement around the 30%. In is worthy mentioning that this superiority is a post-hoc fact: the best method was not selected according to police protection metric but according to the F_1-score achieved in the 'High' class, which is a standard criterion of model selection. A posteriori, we observe that this best method also outperforms regarding the police protection metric, even though it was not optimized for this task. This fact strengthens the reliability in the NC method as the best model for addressing this problem.

Furthermore, Table 1 also evidences that the NC method outperforms the baseline stated by the current VioGen implementation, both in terms of F_1-score and police protection. To improve the soundness of the results, several policies for assigning the recidivism risk have been used on the VioGen assessment method: lax, medium lax, medium cautious, and cautious. These rule systems represent an increasing level of greediness in the application of the protection measures, in such a way that 'lax' is the less conservative rule system (tends to assign low risk to doubtful cases) and 'cautious' is the most conservative system (tends to assign high risk levels to doubtful cases). A complete description of these rule systems is provided in Appendix A.3.

To reinforce these observations, auxiliary analyses have been conducted to optimize the hyper-parameters of each method with police protection as objective function. The best obtained police protections for each method are shown in Table 2. There, we observe that most of the models are bounded by a police protection of 1.1, smaller than the one achieved by the preexisting prediction method in VioGen. Particularly interesting is the case of GBC, which does not improve its results with respect to the optimization focused on F_1-score. Indeed, a closer look at the confusion matrix obtained by GBC evidences that the method misses all the true instances of the class 'High', which widely penalized the police protection score. In contrast, the NC algorithm does achieve compelling results, even better than the preexisting predictors implemented in VioGen.

Finally, it is worth mentioning that, among all the tested ML-based models, NC is the only one that can even compete with the preexisting risk assessment method as provided by VioGen. It is noticeable the poor police protection provided by other

Table 1
Quality measures for the best prediction models against the test split. For each type of prediction method, the best sets of hyper-parameters are shown ordered by rank according to their F_1-score for the 'High' class (top is better).

Method type	Hyper-parameters	'High' class F_1 (higher is better)	Weighted F_1 (higher is better)	Police protection (higher is better)
Decision tree	**Criterion = Entropy Splitter = Best Max depth = None**	**0.09**	**0.62**	**1.08**
	Criterion = Gini Splitter = Random Max depth = None	0.08	0.61	1.08
	Criterion = Entropy Splitter = Random Max depth = None	0.08	0.61	1.08
Random forest	**Criterion = Entropy N estimators = 1 Max depth = 50**	**0.08**	**0.62**	**1.06**
	Criterion = Entropy N estimators = 1 Max depth = 100	0.08	0.61	1.05
	Criterion = Entropy N estimators = 1 Max depth = None	0.06	0.61	1.05
K-Nearest neighbors	***K* = 2** $K = 5$ $K = 10$	**0.08** 0.05 0.00	**0.57** 0.65 0.65	**1.15** 0.93 0.84
Gradient boosting	**Max depth = 5 η = 0.2, γ = 0 λ = 1**	**0.00**	**0.78**	**0.78**
	Max depth = 5 η = 0.2, γ = 1 λ = 1	0.00	0.78	0.78
	Max depth = 5 η = 0.2, γ = 0 λ = 1	0.00	0.78	0.78
Neural network	**Architecture = (50, 100) Dropout = 0.1 Batch size = 100**	**0.05**	**0.65**	**0.97**
	Architecture = (50, 100) Dropout = 0.5 Batch size = 100	0.05	0.65	0.97
	Architecture = (50, 5, 5) Dropout = 0.1 Batch size = 50	0.00	0.65	1.00
Nearest centroid	**Metric = Euclidean Shrink threshold = 0.1**	**0.14**	**0.55**	**1.45**
	Metric = Euclidean Shrink threshold = 1	0.14	0.54	1.48
	Metric = Manhattan Shrink threshold = 1	0.13	0.57	1.35
VioGen risk	**Medium cautious** Cautious Lax Medium lax	**0.10** 0.10 0.6 0.6	**0.62** 0.46 0.63 0.47	**1.19** 1.28 1.12 1.20

methods such as RF, GBC or NN, which are typically presented as the state-of-the-art in classification problems, that cannot reach a compelling performance. This is compatible with the results reported in the literature, which show that many of the standard ML-based methods are unable to capture the subtleties of the gender-based violence [24]. However, it is in sharp contrast with other recent works on the application of AI techniques to VioGen, such as [37], which rely on GBC (particularly, XGBoost, as tested in this paper) as the ML solution of choice despite that our experiments point out that XGBoost obtains very low levels of police protection, even after a fine hyper-parameter tuning.

Despite its simplicity, NC turns to be the best solution to extract the underlying patterns that characterize the existence of a re-aggression. This is because NC operates by computing the archetypal profiles of each of the types of aggression, analogously to the police profiling techniques.

4.2. Fine tuning for the Nearest Centroid (NC) model

Once we selected NC as the best ML model for the risk prediction problem, we underwent a fine tuning of the hyper-parameters of the NC model. To be precise, it was conducted a more exhaustive hyper-parameter search to optimize the police protection score with possible values 'Metric' = Euclidean, Manhattan, Minkowski and 'Shrink threshold' = 0.1, 0.25, 0.5, 0.75, 1, 5, 10, 20, None. For this purpose, again the dataset was randomly split into the training set (67%) and the test set (33%). Now, on the training set, a cross-validation procedure with 10 folds is conducted with the police protection as the objective function.

Table 2
Best police protection achieved by each method.

Method	Hyper-parameters	Police protection
Decision tree	Criterion = Gini Max depth = 50 Splitter = Best	1.04
Random forest	Criterion = Gini Max depth = None N estimators = 1	1.05
K-Nearest neighbors	K = 2	**1.15**
Gradient boosting	Max depth = 5 η = 0.2 γ = 0 λ = 1	0.79
Neural network	Architecture = (100, 5, 10) Dropout = 0.0 Batch size = 50	1.06
Nearest centroid	**Metric = Euclidean** **Shrink threshold = 5**	**1.59**

Table 3
Detailed police protection score of the NC method in cross validation with 10 folds.

Parameters	Mean	Std	Rank
Metric = Euclidean Shrink threshold = 5	**1.595**	**0.037**	**1**
Metric = Minkowski Shrink threshold = 5	1.595	0.037	2
Metric = Euclidean Shrink threshold = 1	1.535	0.045	3
Metric = Minkowski Shrink threshold = 1	1.535	0.045	4

The mean score and its standard deviation along the 10 folds for the best setups of hyper-parameters are shown in Table 3. According to these results, the best model is NC with Euclidean metric and Shrink threshold = 5. This is the model that was used for the experiments of the following sections.

As shown in this table, after the appropriate hyper-parameter tuning, the NC method reaches a police protection of 1.59. In contrast, the preexisting (VioGen) risk prediction that can only achieve a police protection of 1.19, in the best scenario with the 'medium cautious' policy. Again, these results evidence that NC clearly outperforms the police protection provided by the previously implemented assessment methods, with an improvement of the 30% with respect to the baseline.

4.3. Evolution of the quality metrics with the hybrid model

In this section, we shall consider a hybrid model $\hat{f}_{\mu}^{\mathrm{Hy}}$, as described in Section 3.3. This model will interpolate between the preexisting VioGen risk assessment with the medium cautious rule ($\mu = 0$) and the purely ML-based NC method with the previously chosen hyper-parameters ($\mu = 1$), since these are the configurations that achieved the best results (c.f. Table 1).

It is worth noting that this is not another prediction model, but a mixture of two systems that allows decision-makers to conduct a smooth transition from the preexisting risk assessment protocol and the ML-based method. In particular, for $\mu = 0$ the police protection obtained by the hybrid model equals the one currently provided by VioGen, whereas for $\mu = 1$ we reach the police protection accomplished by the NC algorithm (a 33% better than the preexisting one). In this sense, the aim of this section is not to compare the previous results against a new model, but to analyze how the hybrid proposal interpolates from the existing solution

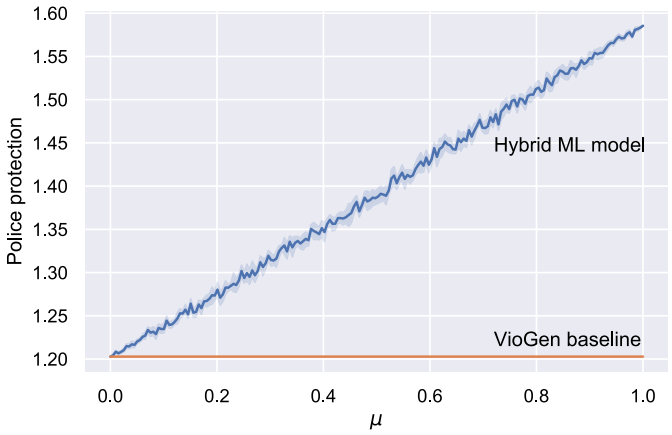

Fig. 3. Evolution of the police protection metric for the hybrid model with the best NC model with varying values of μ. For the hybrid predictor, the best NC model (Metric = Euclidean and Shrink threshold = 5) is used. The values of μ are taken from an uniform grid of the interval [0, 1] with 200 equispaced points. For each value of μ, a random sample of 10 executions is considered. The solid line shows the mean value of the police protection metric along these executions, and the shadowed region is the 0.95 confidence interval around this value. The length of this interval is short enough to provide sound evidence of behavior of the evolution. Narrower confidence intervals are obtained for larger samples, with similar trends in the evolution of the metric.

to the ML-based solution in terms of the evolution of the two proposed quality metrics.

First, we analyzed the dependence of the police protection metric, PoliceProtection($\hat{f}_{\mu}^{\mathrm{Hy}}$), with varying values of $\mu \in [0, 1]$ (Section 3.2.1). The results are shown in Fig. 3. In this plot, we observe that an increase of the importance of the ML algorithm consistently leads to an increase of the police protection. Tests with other VioGen rule systems return similar results, and even show a more prevalent outperformance for the ML method.

In the light of these results, we propose the following practical migration protocol from the existing VioGen prediction to the pure ML-based predictor as follows. Initially, the prediction model used by the police forces might be this hybrid model with a small value of μ (say $\mu = 0.1$). For some time, the evolution of the recidivism statistics is screened. If, after a prudent time, the results improve the ones of the pure preexisting risk assessment system, the value of μ can be slightly increased and subsequently screened. In this way, a falling-off in the effectiveness of the system would be rapidly detected and actions may be taken to move back to the previous useful value of μ. On the other hand, if the progression leads to a constant improvement, as predicted by Fig. 3, eventually the system would be completely migrated to ML in a smooth and controlled way.

Besides, we have analyzed the evolution of the resource overload metric (Section 3.2.2) along the hybrid model. In Fig. 4, we show the police resources overload obtained on the test split of the dataset for the hybrid model $\hat{f}_{\mu}^{\mathrm{Hy}}$ with varying $\mu \in [0, 1]$, for various values of the penalty τ. Despite the stochastic nature of the model, the results of the quality measure are consistent among executions. The obtained distributions exhibit a strong centrality with small standard deviations and 0.95 confidence interval. The dynamics of the metric with varying μ for different values of the penalty τ is shown in Fig. 4. For small values of τ, the overload of police resources is a decreasing function of μ. In other words, a stronger weight of the ML model in the hybrid model leads to a more efficient use of the resources for low penalties. When τ increases, the gain decreases and, at some intermediate point, the trend inverts. For high penalties, the ML model slightly increases the overload of the police system.

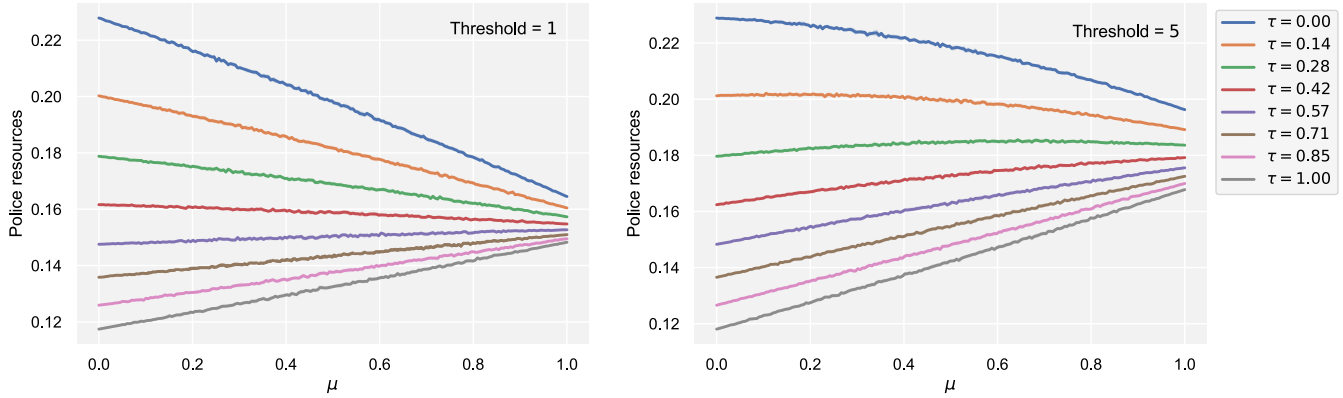

Fig. 4. Results of the police resources metric for different values of the penalty τ for the hybrid model. For the hybrid model, the NC model with Metric = Euclidean and Shrink threshold = 1 (left plot) and Metric = Euclidean and Shrink threshold = 5 (right plot) was applied. The parameter μ was uniformly sampled in the interval $[0, 1]$ with 200 sample points. Each point was computed as the mean police resources overload obtained from a random sample of 10 executions. The mean amplitude of the 0.95 confidence interval is 0.00064 and the maximum amplitude observed is 0.00152.

It is worth mentioning that the NC model set with threshold = 5 (right-hand side of the figure) tends to issue more cautious predictions. This leads to a noticeable overload in the resources for $\tau > 0.4$. However, the NC model with threshold = 1 (left-hand side of the figure), which reached the third position in the hyper-parameter tuning in terms of police protection (Section 4.3), achieved a much better performance in terms of police resources overload with only a small drop in the police protection provided. Even for quite high values of τ, with this setting, the ML-based system is able to improve the usage of resources with respect to the VioGen baseline. This evidences that, for police systems with a high penalty, it would be recommendable to choose a NC model with small threshold to achieve a better balance between provided police protection and amount of resources invested.

Summarizing, the results show that the proposed model based on NC classification does not lead to a substantial increase in the invested police resources with respect to the existing prediction method, and even reduces it for small values of the penalty τ. This points out that a progressive migration from the preexisting prediction model to a ML-based model, like NC, is highly encouraged.

5. Discussion

This work proposes a methodology based on ML techniques to help authorities involved in policy against crime. In particular, it is applied to handle a official data set, VioGen, dependent of the Spanish Government, to assess the risk of revictimization in IPVAW. The results obtained provide relevant clues to decision-makers to impose the right protection measures to victims. The ML algorithm is parametrically linked to the current VioGen system to take advantage of the previous experiences and, simultaneously, allows a successive improvement of risk assessment.

In this paper, we have introduced a stochastic hybrid model that ensembles the prediction of the preexisting risk assignment method with the proposed ML method. This hybrid solution is parametrized by a coefficient $0 \leq \mu \leq 1$ that weighs the importance of the ML model in the mixture. As aforementioned, this parameter μ allows decision-makers to smoothly transit from the existing risk prediction model to a fully ML model. But, furthermore, this parameters also enables a fine tuning of the balance between invested resources and gain in the offered police protection. As shown in Fig. 4, for small values of the penalty τ the ML method is more efficient than the existing risk assessment model, so this transition towards a fully automatized model can be carried out without overloading the invested resources.

Nevertheless, if the penalty τ is large, the ML model suffers a small increase in the use of resources that might not be affordable for the police system. In that case, the two plots of Figs. 3 and 4 may be jointly used by a decision-maker to provide an optimal solution. This process is illustrated in Fig. 5. The first step is to empirically estimate the penalty τ of the police system by conducting in-field measures. Once τ has been calculated, the decision-maker should fix a maximum amount of resources that the system is able to support. Let us call this value r_0. Now, we draw the horizontal line $y = r_0$ in the police resources plot corresponding to τ. The x-coordinate of the intersection of this line with the police resources function determines the maximum feasible value of the parameter μ, namely μ_0. For any value of the weight $0 \leq \mu \leq \mu_0$, the obtained hybrid system fulfills the imposed constraints on the used resources. Therefore, transitioning from the initial model $\mu = 0$ to the hybrid state $\mu = \mu_0$ leads to an increase in the provided police efficiency without exceeding the maximum of resources assigned.

6. Conclusions

6.1. Practical implications and future work

In this work, we have shown that ML-based solutions can be efficiently and accurately used to estimate the risk of recidivism in gender-violence, reaching to a noticeable improvement with respect to the preexisting prediction techniques. For this purpose, we focused on predicting the number of recidivisms that a case suffered, as reported in 44,000 cases extracted from the official Spanish VioGen system.

These results open a new horizon of possibilities in the application of ML-based techniques for risk prediction. An obvious next step in this idea is to test more complex ML methods to try to improve the quality of the predictions. Even though this work covers a wide range of general-purpose supervised learning algorithms, it would be interesting to handcraft a prediction model specially adapted to this kind of crimes. It is very likely that designing new methods able to exploit the psychological and spatial features of these cases would improve even more the accuracy of the predictions and thus the protection to the victims.

Another interesting future work would be to assess the risk not only in terms of number of recidivisms, but also in terms of its severity. There exists cases for which only a few further recidivisms occur, but associated to very severe aggressions, whereas other cases may lead to many aggressions of low intensity. In the current state, both the preexisting VioGen method and our ML-based solution focus on the bare number of aggressions, and it

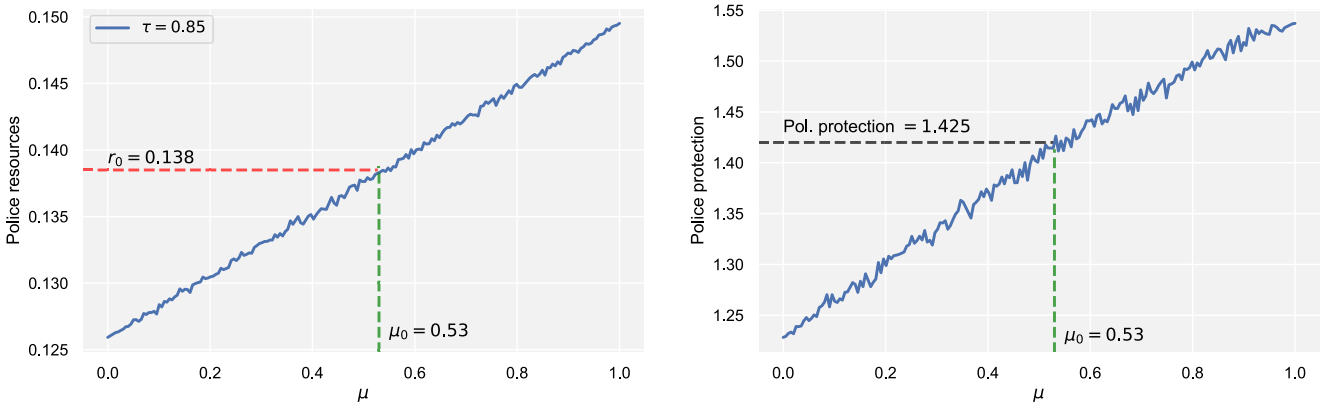

Fig. 5. Illustration of the procedure for adjusting the optimal value of μ with resource constraints. The NC model with Metric = Euclidean and Shrink threshold = 1 was applied for the hybrid model. In this plot, the penalty of the police system has been set to $\tau = 0.85$ and the maximum acceptable resources overload to $r_0 = 0.137$ (10% of increase with respect to the initial value of 0.125). The line $y = r_0$ in the police resources plot (on the left) intersects the resources function at the optimal value μ_0 ($\mu_0 = 0.53$ in this plot). The obtained gain in protection can be read from the police protection plot (on the right). In this example this value corresponds to 1.425, which amounts to a increase of almost the 18% with respect to the original value of 1.21. Therefore, an extra investment of the 10% in resources leads to an improvement of the 18% in the provided protection. For smaller values of τ, the gain is even bigger.

would be very interesting to develop new models able to also detect this very different outcomes depending on the severity.

In a similar vein, the current model is designed to assess the risk of recidivism at the moment of the first aggression. However, no follow-up of the temporal evolution of this risk is provided. In this sense, it would be very interesting to apply these ML-based methods to obtain dynamical predictions adapted to the evolution of the case.

Finally, we would also point out that this work has been carried out with data from the year 2018. In the meantime, the COVID-19 pandemic has occurred, which has deeply affected to some habits and social behaviors of the population. In this way, even though we expect that similar algorithms can be used to reach accurate risk predictions, it may happen that some features have lost importance after the pandemic, whereas others have gain predictive capability due to the changes in the social habits. In this direction, an interesting prospective work is to compare the results obtained with data before and after the pandemic, and to investigate whether new ML-models may reach to better results in the current post-pandemic situation.

6.2. Intrinsic difficulties: The quantum nature of crime prediction

Before ending, we would like to point out that the difficulties of risk prediction found in this work are not due to the dataset, which is certainly large and reliable, but to the intrinsic characteristics of the problem. The implementation of any of the measures that are deduced from the system protocol is changing the own behavior of the players in the crime scene. Unlike in other settings, such as medical research, it is not possible to set a control group where we do not intervene to detect how much this protocol improves safety. Due to obvious legal and moral concerns, it is not possible to choose not to act under the suspicious of danger.

This lack of ground truth prevents any true validation of the model, leading to an intrinsic difficulty to measure whether the protective actions have been really effective in the prevention of revictimization: if nothing wrong happens it may be a consequence of the properly taken actions or because the system failed in its prediction of danger. This cause–effect reversion is somehow analogous to the uncertainty principle of Quantum Mechanics, where measuring the system inevitably leads to its modification.

Furthermore, the dependent variable, the assessment of risk, has never been measured. Instead, it has been only estimated from an indirect measure: the relapses. Indeed, the proposed algorithm is learning to classify from an intangible asset: the risk of victimization. As a consequence of these two conceptual drawbacks, standard ML techniques are not always efficient, being Nearest Centroid (NC) an exception, as shown in this paper. Despite this apparent paradox, any policy program is going to be successful if the crime rate decreases, independently of the cause.

In parallel with the *Precog* system of Philip K. Dick, designed to eliminate criminality completely from our societies, the algorithm of machine learning proposed in this paper uses the information of previous violent episodes to provide an assessment of risk for victimization. The results obtained suggest that any security system, whatever its implementation within its national policy programs, would benefit from a rigorous ML-based risk classification following the main strategies of this work. They also evidence that further research in the line of application of AI to crimes with an important psychological component is strongly encourage. Indeed, the approach developed in this work also generalizes to other types of crimes, like juvenile delinquency and gang rivalry, provided that some similarity assumptions hold, namely distinguishable profiles of aggressors and victims.

We expect that the analysis, models, methodology and applications provided in this work will foster the transference of ML techniques in gender-based crimes to the official police risk assessment protocols and eventually will lead into a significative drop in the number and severity of intimate aggressions against women.

CRediT authorship contribution statement

Ángel González-Prieto: Conceptualization, Methodology, Formal analysis, Investigation, Software, Validation, Writing – original draft, Writing – review & editing. **Antonio Brú:** Conceptualization, Methodology, Formal analysis, Investigation, Software, Validation, Writing – original draft, Writing – review & editing. **Juan Carlos Nuño:** Conceptualization, Methodology, Formal analysis, Investigation, Software, Validation, Writing – original draft, Writing – review & editing. **José Luis González-Álvarez:** Data curation, Validation, Writing – review & editing, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors are greatly indebted to D. Gómez-Castro for reading very carefully this manuscript, his comments and suggestions to improve the exposition of this paper, and his constant encouragement throughout the development of this project. They also acknowledge the support of the staff of the Centro de Procesamiento de Datos (CPD) at Universidad Complutense de Madrid during the conduction of the experiments of this work. The first-named author also wants to thank R. Lara-Cabrera and F. Ortega for very useful conversations, as well as the hospitality of the ETSI de Sistemas Informáticos at Universidad Politécnica de Madrid and Departamento de Matemáticas at Universidad Autónoma de Madrid where this work was partially developed.

The first-named author has been partially supported by the Spanish *Ministerio de Ciencia e Innovación* through the project PID2019-106493RB-I00 (DL-CEMG). This work has been supported by the *Madrid Government (Comunidad de Madrid – Spain)* under the Multiannual Agreement with the Universidad Complutense de Madrid in the line Research Incentive for Young PhDs, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation) through the project PR27/21-029.

Appendix A. Gender-based crime dataset

The data processed in this paper was collected through the Spanish VioGen system. VioGen is a governmental computational system oriented to collect, analyze, and suggest preventive measures related to the cases of IPVAW. Currently, it has a national scope of action and supports concurrent multiuser access. It provides support for archiving and parallel processing all the reports of gender crime that take place countrywide. According to these reports, the system issues a prediction of the recidivism risk for each case.

According to official sources, since its start-up in 2007 until December 2020, 613,065 cases have been reported to the VioGen system [46]. This implies that, on average, the system adds 75 new cases per day in a total population of 48 million people. On a regular basis, around 60,000 cases remain active at any given time, which require a substantial amount of police resources.

Among other functionalities, the VioGen system gathers data about the first occurrence of a IPVAW case and its subsequent episodes. The former information is collected through a standard form called VPR, from the Spanish 'Valoración Policial del Riesgo' or Police Assessment of the Risk. The later data are collected by means of a different form called VPER, from the Spanish 'Valoración Policial de la Evolución del Riesgo' or Police Assessment of the Evolution of the Risk. In the following sections, we describe the scope and outcome of these forms, which have served as input for the training and testing of the ML models.

A.1. VPR: First aggression report

Once a woman proceeds to report the occurrence of a gender-based aggression to the police, the system offers the agent a detailed form with several questions regarding the circumstances, severity, and potentially aggravating factors surrounding the crime. In this way, after finishing the needed inquiries for gathering all the evidences, the police agent uploads these data into the VioGen system through the VPR form.

According to the input data, the system provides a risk assessment of recidivism that ranges from 'Not appreciated', with value 0, to 'Extreme', with value 4. It is worth mentioning that, in the analyzed version, this scale does not measure the risk of prospective lethal crime, but the probability of recidivism in the violent behavior, regardless of its severity, intensity or time lapse between assaults. It is worth noting that the number of cases rated with each risk level remains almost constant along time, with very small variations. Table A.4 shows the distribution of the risk assessments in the active cases reported in VioGen for the last five years as at 31st December.

Table A.4
Classification of active cases into the five risk groups in the VioGen system along the last five years. Notice that the distributions are stable over time.

Year	2016	2017	2018	2019	2020
Active cases	52,635	54,793	58,498	61,355	63,656
Not appreciated	56%	50%	43%	50%	49%
Low	36%	42%	46%	39%	41%
Medium	7%	8%	10%	10%	10%
High	0.3%	0.4%	0.4%	0.7%	0.7%
Extreme	0.02%	0.03%	0.04%	0.02%	0.01%

According to this forecast, the person in charge of the case decides to apply some preventive measures, which may range from a periodic report of the evolution of the case to preventive imprisonment of the aggressor.

A.2. VPER: Follow-up of the aggression case

It may happen that, despite the taken measures, a subsequent aggression still occurs. In that case, the new report is not added to the VioGen database through the VPR form but through a specialized form called VPER. Static indicators, like the previous police records of the criminal, lose importance in the VPER form and new dynamical indicators regarding the evolution of the aggressivity of the case are added. In addition, the VPER subsystem provides support for collecting periodic reports that the victims communicate during the surveillance period, even thought no further offense occurs.

Notice that, by means of these VPER forms, it is possible to count the number of recidivisms of a closed case. This number correlates with the recidivism risk, the amount of threat of a victim of being offended again. This risk is the target variable that has been studied in this paper, whose main goal is precisely to discuss how this recidivism risk can be predicted beforehand so that an appropriate police protection can be applied to the victim, which hopefully leads to an effective avoidance of future violent episodes.

A.3. Rule systems derived from VioGen

As mentioned above, the VioGen system automatically assigns a risk to each VPR case according to the provided answers to the VPR checklist, ranging from 'Not appreciated' (0) to 'Extreme' (4). This VioGen prediction is generated through classical statistical methods. Briefly, it operates as follows. The possible responses to each of the items of the questionnaire were assigned with a weight depending on their observed importance based on psychometric criteria. Hence, given a report of gender violence, its score of recidivism risk is computed as the weighted sum of all its responses according to the pre-fixed weights. To discretize this score, thresholds on each level were imposed in such a way that the number of cases in each of the five risk classes agrees with the empirical distribution of these values found in a pilot study, as classified by experts. The scoring weights have been updated throughout the different versions of VioGen (see [4] for the latest version). For accessibility and completeness, this work applies

Fig. A.6. Correlation plot of some of the collected answers to the VPR form. The plot compares the answers of 1000 randomly chosen cases from the VioGen dataset. The diagonal plots show the distribution of the answers of the chosen questions. The off-diagonal plots compares two-by-two the answers to the questions. For each plot, the answers are placed on a rectangular grid with as many columns (resp. rows) as response options have the question displayed horizontally (resp. vertically). The left-most points of each plot for the horizontal axis (resp. bottom points for the vertical axis) correspond to 'No'/'Very mild' responses, whereas the right-most points (resp. upper points) correspond to 'Yes'/'Very severe' responses. Missing responses were assigned to a medium value. A small random noise was added to improve visualization.

the coefficients corresponding to the VioGen prediction method described in [35] (see also [47]).

Using this assessment of risk, there is a direct translation into a rule system for predicting the recidivism risk. This rule system maps the five VioGen classes into the three classes of $\Lambda = \{\text{No}, \text{Low}, \text{High}\}$. If we restrict ourselves to maps preserving the ordering of severity of the VioGen classes, a rule system is thus implemented just by setting up two thresholds that define the changes of class. To cover all the possibilities, in this work we considered four rule systems, ranging from a *lax system*, which tends to predict low risk of recidivism, to a *cautious system*, which tries to provide a more aggressive police response by assigning higher levels of risk. The particular chosen thresholds for these rule systems can be found in Table A.5. These assignments agree with the last version of the VioGen risk assessment [48].

Notice that, instead of the five risk assessment classes provided by VioGen, for the risk classification problem we use only the three labels of Λ. From a methodological point of view, labeling the cases with only three classes allows us to distinguish more effectively between highly concerning cases and not so risky aggressors. With more classes, the boundaries become very fuzzy and some clusters contain only a few instances, in agreement with what happens when the threshold between the 'Low' and 'High' classes is increased. This collapsing of the five classes of VioGen into a three-levels scale is also typical in the literature, as in the works screening the performance of the VioGen risk assessment for recidivism avoidance [48]. Additionally, we have conducted experiments using more prediction classes and the results obtained are similar to the ones exposed in this work with only three classes.

In spite of this reduction, the classification problem for risk assessment is still very hard. In Fig. A.6 we show the correlation

Table A.5
Rule systems for the VioGen risk assessment considered in this work. Each VioGen class is mapped into a class of Λ. In bold, the assignments that vary depending on the imposed thresholds.

Rule system	VioGen risk assessment class				
	Not appr.	Low	Medium	High	Extreme
Lax	No	**No**	Low	**Low**	High
Medium lax	No	**Low**	Low	**Low**	High
Medium cautious	No	**No**	Low	**High**	High
Cautious	No	**Low**	Low	**High**	High

matrix between some of the most relevant items in the VPR questionnaire with respect to each of the classes in Λ. Each point of the plot corresponds to a IPVAW case and is colored according to the observed recidivism with the codes explained in Section 2: Blue stands for 'No' (no other offenses were reported), Orange stands for 'Low' (1 or 2 later offenses), and Green stands for 'High' (more than 3 subsequent offenses). The used indicators correspond to the following items of the VPR questionnaire: 'Verbal abuse' refers to question F01-I01 (*Did verbal abuse take place in the reported offense?*), 'Physical abuse' refers to question F01-I02 (*Did physical abuse take place in the reported offense?*), 'Sexual abuse' refers to question F01-I03 (*Did sexual abuse take place in the reported offense?*), 'Used weapons' refers to question F02 (*Did the aggressor use weapons or any other threatening objects against the victim in the reported offense?*), 'Incr. violence' refers to question F04 (*Does there exist an increasing severity and/or frequency in the aggressions or theats of violence in the last six months?*), and 'Prev. violence' refers to question F08-I19 (*Does there exist previous records of gender violence of the aggressor against other victims?*). This figure depicts the highly non-separable nature of the dataset, with many overlapping cases presenting similar responses to the main indicators.

It is worth mentioning that the threshold of 3 re-aggressions to distinguish between 'Low' and 'High' risk cases may seem rather arbitrary. Preliminary experiments show that the results with a threshold of 3, 4 or 5 are similar. However, for bigger thresholds $\geq$ 6, the ML models start losing accuracy in their predictions. The reason for this lost is twofold: first, the number of cases with 6 or more offenses is very small ($\sim$ 100 in our VioGen dataset) so the ML method is not able to extract relevant features from such a limited dataset. Second, the selected NC method operates by identifying 'characteristic profiles' of IPVAW cases. For high values of the threshold, the 'Low' class aggregates several profiles of cases and this confuses the model, pointing out that a further refinement of the 'Low' class is needed.

Appendix B. Materials and methods

B.1. Data preprocessing

The analyzed dataset is an excerpt of the Spanish VioGen database. The data correspond to 44,463 cases of Intimate Partner Violence Against Women (IPVAW) reported between 1st October 2016 and 1st October 2017. Each case was stored as an array containing the 58 responses provided to the VPR form by the police agent when the aggression was reported. These questions collect information about the features of the violent episode, the psico-social features of the aggressor, the potential vulnerabilities of the victim, the circumstances related to minors, and possible aggravating factors. A thorough description of these indicators and possible responses can be checked in [35]. Demographic and temporal information, not corresponding to the VPR form but also collected in VioGen, was removed during the preprocessing phase to avoid bias.

Missing values in the data were substituted by a special character. Multiple choice answers were codified through one-hot encoding, in such a way that each case is finally codified as a real 250-dimensional vector that corresponds to an entry in the generated dataset. To compute the number of recidivist offenses, we counted the number of VPER entries associated to each case. Specifically, for each case number in the VPR database, i.e. each entry of the dataset, we counted the number of VPER occurrences corresponding to re-aggressions of that case.

B.2. Reproducibility and data availability

The authors of this paper are committed to reproducible science. All the experiments have been conducted with the open source Python library Scikit-learn library [44]. Due to privacy constraints, data are available under demand by contacting with the corresponding author.

References

[1] W.H. Organization, Responding to intimate partner violence and sexual violence against women: WHO clinical and policy guidelines, World Health Organization, 2013.

[2] B. Sanz-Barbero, C. Linares, C. Vives-Cases, J.L. González, J.J. López-Ossorio, J. Díaz, Intimate partner violence in Madrid: a time series analysis (2008–2016), Ann. Epidemiology 28 (9) (2018) 635–640.

[3] J.L. Torrecilla, L. Quijano-Sánchez, F. Liberatore, J.J. López-Ossorio, J.L. González-Álvarez, Evolution and study of a copycat effect in intimate partner homicides: A lesson from spanish femicides, PLoS One 14 (6) (2019) e0217914.

[4] 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, Psychosoc. Interv. 30 (1) (2020) 47–55.

[5] R. Tibshirani, T. Hastie, B. Narasimhan, G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression, Proc. Natl. Acad. Sci. 99 (10) (2002) 6567–6572.

[6] J. Rocchio, Relevance feedback in information retrieval, Smart Retr. Syst.-Exp. Autom. Document Process. (1971) 313–323.

[7] S. Vettor, J. Woodhams, A.R. Beech, Offender profiling: A review and critique of the approaches and major assumptions, J. Curr. Issues Crime, Law Law Enforc. 6 (4) (2013).

[8] M. del Mar Pecino-Latorre, J. Santos-Hermoso, M. del Carmen Pérez-Fuentes, R.M. Patró-Hernández, J.L. González-Álvarez, The action system model: a typology of Spanish homicides, Front. Psychol. 11 (2020).

[9] E. Durkheim, Le Crime, Phénoméne Normal, J.-M. Tremblay, 2006.

[10] L.A.J. Quetelet, Recherches Statistiques sur Le Royaume Des Pays-Bas, Tarlier, 1829.

[11] P.K. Dick, The Minority Report: And Other Classic Stories, vol. 30, Citadel Press, 2002.

[12] J.C. Nuño, M.A. Herrero, M. Primicerio, A triangle model of criminality, Phys. A 387 (12) (2008) 2926–2936.

[13] S.D. Johnshon, A brief history of the analysis of crime concentration, European J. Appl. Math. 21 (4–5) (2010) 349.

[14] P.J. Brantingham, B. Yuan, D. Herz, Is gang violent crime more contagious than non-gang violent crime? J. Quant. Criminol. (2020) 1–25.

[15] M.B. Short, P.J. Brantingham, A.L. Bertozzi, G.E. Tita, Dissipation and displacement of hotspots in reaction-diffusion models of crime, Proc. Natl. Acad. Sci. 107 (9) (2010) 3961–3965.

[16] A.B. Pitcher, S.D. Johnson, Exploring theories of victimization using a mathematical model of burglary, J. Res. Crime Delinquency 48 (1) (2011) 83–109.

[17] C. Wang, Y. Zhang, A.L. Bertozzi, M.B. Short, A stochastic-statistical residential burglary model with independent Poisson clocks, European J. Appl. Math. (2020) 1–27.

[18] P. Boqué, L. Serra, M. Saez, 'Surfing' burglaries with forced entry in Catalonia: Large-scale testing of near repeat victimization theory, Eur. J. Criminol. (2020) 1477370820968102.

[19] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, third ed., Prentice Hall Press, USA, 2009.

[20] R.A. Berk, S.B. Sorenson, Algorithmic approach to forecasting rare violent events: An illustration based in intimate partner violence perpetration, Criminol. Public Policy 19 (1) (2020) 213–233.

[21] J.G. Cabello, Intimate partner violence: A novel warning system in which the victims' environment alerts to the danger, Heliyon 6 (1) (2020) e03211.

[22] N. Hassan, A. Poudel, J. Hale, C. Hubacek, K.T. Huq, S.K.K. Santu, S.I. Ahmed, Towards automated sexual violence report tracking, in: Proceedings of the International AAAI Conference on Web and Social Media, 14, 2020 pp. 250–259.

[23] I. Rodríguez-Rodríguez, J.-V. Rodríguez, D.-J. Pardo-Quiles, P. Heras-González, I. Chatzigiannakis, Modeling and forecasting gender-based violence through machine learning techniques, Appl. Sci. 10 (22) (2020) 8244.

[24] E. Turner, J. Medina, G. Brown, Dashing hopes? The predictive accuracy of domestic abuse risk assessment by police, Br. J. Criminol. 59 (5) (2019) 1013–1034.

[25] Z.R. Yang, Machine Learning Approaches To Bioinformatics, first ed., World Scientific Publishing Co., Inc., USA, 2010.
[26] Bureau of Justice Statistics, Statistics about different types of crime in the USA, 2021, https://www.bjs.gov/index.cfm?ty=pbtp&tid=3. (Accessed: 2021-02-05).
[27] Council of Europe, The Council of Europe Convention on Preventing and Combating Violence Against Women and Domestic Violence, 2014.
[28] S. Walby, J. Towers, B. Francis, Is violent crime increasing or decreasing? A new methodology to measure repeat attacks making visible the significance of gender and domestic relations, Br. J. Criminol. 56 (6) (2016) 1203–1234.
[29] L. Richards, Domestic abuse, stalking and harassment and honour based violence (DASH, 2009) risk identification and assessment and management model, Assoc. Police Off. (ACPO) (2009).
[30] K. Dayan, S. Fox, M. Morag, Validation of spouse violence risk assessment inventory for police purposes, J. Fam. Violence 28 (8) (2013) 811–821.
[31] O.S. Cunha, R.A. Goncalves, Severe and less severe intimate partner violence: From characterization to prediction, Violence Vict. 31 (2) (2016) 235–250.
[32] J.T. Messing, J. Thaller, Intimate partner violence risk assessment: A primer for social workers, Br. J. Soc. Work 45 (6) (2015) 1804–1820.
[33] P.R. Kropp, Some questions regarding spousal assault risk assessment, Violence Against Women 10 (6) (2004) 676–697.
[34] R. Petering, M.Y. Um, N.A. Fard, N. Tavabi, R. Kumari, S.N. Gilani, Artificial intelligence to predict intimate partner violence perpetration, Art. Intell. Soc. Work (2018) 195.
[35] J.J. López-Ossorio, J.L.G. Álvarez, S.B. Pascual, L.F. García, G. Buela-Casal, Risk factors related to intimate partner violence police recidivism in Spain, Int. J. Clin. Health Psychol. 17 (2) (2017) 107–119.
[36] J.J. López-Ossorio, J.L. González-Álvarez, J.M. Muñoz Vicente, C. Urruela Cortés, A. Andrés-Pueyo, Validation and calibration of the spanish police intimate partner violence risk assessment system (Viogén), J. Police Crim. Psychol. 34 (4) (2019) 439–449.
[37] L. Quijano-Sánchez, F. Liberatore, G. Rodríguez-Lorenzo, R.E. Lillo, J.L. González-Álvarez, A twist in intimate partner violence risk assessment tools: Gauging the contribution of exogenous and historical variables, Knowl.-Based Syst. 234 (2021) 107586.
[38] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[39] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, IEEE, 1995 pp. 278–282.
[40] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[41] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Amer. Statist. 46 (3) (1992) 175–185.
[42] 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, 2016, pp. 785–794.
[43] D. Svozil, V. Kvasnicka, J. Pospichal, Introduction to multi-layer feed-forward neural networks, Chemometr. Intell. Lab. Syst. 39 (1) (1997) 43–62, URL https://xgboost.readthedocs.io/en/stable/.
[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[45] F. Chollet, et al., Keras, 2015, URL https://github.com/fchollet/keras.
[46] Government Office against Gender-based Violence, Monthly statistical newsletter about gender-based crimes in Spain, 2021, https://violenciagenero.igualdad.gob.es/en/violenciaEnCifras/boletines/boletinMensual/home.htm. (Accessed: 2021-02-05).
[47] 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) del Ministerio del Interior de España (Ph.D. thesis), Facultad de Psicología. Universidad Autónoma de Madrid, 2017.
[48] J.J. López-Ossorio, I. Loinaz, J.L. González-Álvarez, Protocolo para la valoración policial del riesgo de violencia de género (VPR4. 0): revisión de su funcionamiento, Rev. Española de Med. Leg. 45 (2) (2019) 52–58.