applied sciences

MDPI

Article

Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques

Ignacio Rodríguez-Rodríguez [1,2,*], José-Víctor Rodríguez [3], Domingo-Javier Pardo-Quiles [3], Purificación Heras-González [4] and Ioannis Chatzigiannakis [5]

1 Departamento de Ingeniería de Comunicaciones, ATIC Research Group, Universidad de Málaga, 29071 Málaga, Spain
2 Instituto Universitario de Investigación de Estudios de Género, Universitat d'Alacant, 03080 Alicante, Spain
3 Departamento de Tecnologías de la Información y las Comunicaciones, Universidad Politécnica de Cartagena, 30202 Cartagena, Spain; jvictor.rodriguez@upct.es (J.-V.R.); domingo.pardo@upct.es (D.-J.P.-Q.)
4 Departamento de Ciencias Sociales y Humanas, Universidad Miguel Hernández de Elche, 03202 Elche, Alicante, Spain; p.heras@umh.es
5 Dipartimento di Ingegneria Informatica Automatica e Gestionale 'Antonio Ruberti', Sapienza Università di Roma, 00185 Roma, Italy; ichatz@diag.uniroma1.it
* Correspondence: ignacio.rodriguez@ic.uma.es

Received: 28 October 2020; Accepted: 18 November 2020; Published: 20 November 2020

Abstract: Gender-Based Violence (GBV) is a serious problem that societies and governments must address using all applicable resources. This requires adequate planning in order to optimize both resources and budget, which demands a thorough understanding of the magnitude of the problem, as well as analysis of its past impact in order to infer future incidence. On the other hand, for years, the rise of Machine Learning techniques and Big Data has led different countries to collect information on both GBV and other general social variables that in one way or another can affect violence levels. In this work, in order to forecast GBV, firstly, a database of features related to more than a decade's worth of GBV is compiled and prepared from official sources available due to Spain's open access. Then, secondly, a methodology is proposed that involves testing different methods of features selection so that, with each of the subsets generated, four techniques of predictive algorithms are applied and compared. The tests conducted indicate that it is possible to predict the number of GBV complaints presented to a court at a predictive horizon of six months with an accuracy (Root Median Squared Error) of 0.1686 complaints to the courts per 10,000 inhabitants—throughout the whole Spanish territory—with a Multi-Objective Evolutionary Search Strategy for the selection of variables, and with Random Forest as the predictive algorithm. The proposed methodology has also been successfully applied to three specific Spanish territories of different populations (large, medium, and small), pointing to the presented method's possible use elsewhere in the world.

Keywords: gender-based violence; machine learning; information and communication technologies; multi-objective evolutionary search; random forest; time series forecasting

1. Introduction

Right now, Intimate Partner Violence (IPV) is a significant issue for a large number of women around the globe. Its impact incorporates physical, sexual, and mental mischief by a current or previous partner, in any form or means. As per UN (United Nations) reports, practically 35% of women around the globe have encountered some sort of physical or sexual violence [1], while similar insights find that some 75% of women face physical and sexual hostility. This paper is looking to

highlight this issue. Unfortunately, in 2017, nearly 87,000 women were executed globally, of whom 58% (50,000) were murdered by their other half or by different family members (https://www.unodc.org/).

Lately, a great deal of research has focused on IPV and its association with many related issues. The scope of this analysis incorporates the resources utilized by the victims (perhaps better referred to as 'survivors') [2], as well as examining the boundaries involved, and also the public policies focused on preventing and overseeing Gender-Based Violence (GBV).

The Council of the Europe Convention on preventing and combating violence against women, including domestic violence (opened in 2011 and in effect since 1 August 2014)—better known as the Istanbul Convention—emphasizes in Article 17 that countries shall encourage the Information and Communication Technologies (ICT) sector especially to participate in policies to prevent violence against women.

In 2020, ICT and Machine Learning (ML) have clearly progressed and had their impact felt throughout all avenues of society, and all around the world. In the mid-1990s, Haraway [3] predicted the social changes and the impact, particularly for sex-related issues, that would come with ICT.

Fortunately, both ICT and ML strategies offer additional opportunities for preventing and dealing with these kinds of violence. Innovative advances, software, and new ideas—for example, the Internet of Things (IoT) and cloud-computing methodologies—offer a wide scope of opportunities for managing violence against women [4], especially after being effectively integrated with different fields, for example, e-health [5]. Systems for advanced data processing—with Machine Learning (ML) and Big Data [6] as two key examples—can likewise be utilized to battle gender violence.

Luckily, in recent years, governments have recognized the power of data analysis and its potential for policy planning, which has resulted in efforts to systematically collect information on a wide range of topics spanning many decades. In this regard, Spain has been attentive to these trends and, since 2003, has been collecting valuable data via the National Institute of Statistics (*Instituto Nacional de Estadística, INE*), structured as a time series related to violence against women that now makes it possible to analyze GBV and its relationship to other variables.

This work puts ML techniques into practice in order to model and forecast the incidence of GBV according to a predictive horizon of half a year, achieved by extracting the variables that have the highest influence on the existence of such violence from a total of more than 30 features extracted from a Spanish national database. In addition, the possibility of forecasting GBV is analyzed using four predictive algorithms so that governments can improve their policy planning on this issue, thereby optimizing and maximizing strategies.

To fulfill this paper's objectives, Section 2 describes previous contributions based on ML applications for improving public policies and actions against GBV. Section 3 explores different techniques in the field of features selection and forecasting in time series. Section 4 explains the nature of the collected database that will be analyzed with the methodology proposed in Section 5. The results, in Section 6, include the performance of the modeling stage under different approaches added to a test of forecasting GBV and its accuracy under certain methods. Finally, Section 7 draws conclusions, suggests future works, and closes the document.

2. Related Works

ML is an application of Artificial Intelligence (AI) that provides ICT-based systems with the ability to automatically learn and improve from experience without being explicitly programmed. These algorithms are able to extract knowledge from data and then, after a learning phase, to develop a complex task. Diagnosis, analysis, and forecasting are among the possible applications, alongside many others. Consequently, therefore, raw data can be utilized to form new knowledge.

ML has been widely applied in the field of GBV. For example, smart speakers implement AIs that are equipped with modules for voice recognition based on ML. This is fundamental for protecting survivors, as voice recognition can distinguish the orders and keywords given by women suffering from IPV—or by the offenders—then obtain the requested information or warn the emergency

channels. Islam et al. [7] partially based their work's proposal on this idea. ML can also recognize abnormal physical activity thanks to classification techniques. Hegde et al. [8] used the data collected from wearable sensors to identify the activities of daily life. This could also be useful for detecting unexpected situations, such as potential aggression.

ML techniques are also powerful in modeling data and for making predictions that can be very helpful for improving the management of all kinds of daily life issues. Glaeser et al. [9] used ML to predict educational (and other) outcomes in a database of dysfunctional families, producing predictive algorithms for city governments and envisioning the predictions as a first step toward generating new insights. So, ML has the capability of analyzing years of collected data and model behaviors, with which it is possible to adapt future strategies and, for example, to reorganize resources, or adapt and optimize public budgets [10].

Research has applied ML to various questions posed by political science, demographics, economics, and criminology [11], via which the limitations of the linear modeling framework and the criteria applied for evaluating findings are discussed. Kleinberg et al. [12] use an ML model to predict the patients who will benefit the most from joint replacement surgery from a dataset of possible beneficiaries. Cederman and Weidmann [13] investigate whether ML can predict armed conflict, while Beck et al. [14] previously used neural networks to forecast militarized international disputes. Furthermore, Brandt et al. [15] employed automated coding to forecast Palestinian–Israeli conflicts, and Perry [16] applied the Random Forest technique to predict violent episodes in Africa. These scholars use their predictions as a starting point for disentangling the process in question and for pushing existing theories.

Kleinberg et al. [17], for example, illustrate how these kinds of predictions can help us to understand the process underlying judicial decisions. The authors began by training a model to predict judges' bail-or-release decisions in New York City, USA. Their findings show that judges can overweigh current charges, releasing high-risk cases if their present charge is minor and detaining low-risk ones if the present charge is serious. From a policy standpoint, the authors' predictive model, if used in practice, promises significant welfare gains over human decisions without eroding important social values (e.g., racial equality): reducing reoffending rates by 25% with no increase in jailing rate or, alternatively, pulling down the jailing rate by 42% with no increase in reoffending rate. In the same sense, Coglianese and Lehr [18] introduced the idea of 'cyberdelegation', as part of a debate over whether AI can be introduced as a support for court processes.

So, the possibilities of ML becoming applied to public policies and as support for decision-making are clear. Indeed, specifically, ML algorithms have also been applied to the field of violence and crime, with some efforts to do so going back more than 30 years, e.g., analysis via linear regressions time series to reflect the number of arrests per day [19]. Ozkan [20] studied the possibility of future recidivism in offenders, applying neural networks, and achieving good results. The insights offered by algorithms can be used to decide about parole in interpersonal violence situations [21]. For example, Berk et al. [22] used Random Forest algorithms and concluded that approximately 20% of those released after an arraignment for domestic violence are arrested within two years for a new domestic violence offense. Their results also proposed an important ranking of risk factors for multiple assaults.

In order to obtain a prediction accurate enough, it is necessary to start with an adequate database. Fortunately, the rise of Big Data has led to the collection of all kinds of data over the last few decades, in particular regarding how society develops in terms of wealth, social stability, employment, culture, etc. Although GBV itself has been included as a feature to forecast crime, as shown in the work of Holcomb and Sharpe [23] where police calls were forecasted, violence against women has also been the variable studied in relation with other crucial factors, like unemployment [24] and the increasing recidivism among unemployed suspects. However, other less direct circumstances have also become the subject of study, such as in the work of Cohn, where the influence of seasonal factors on domestic violence incidence was analyzed [25].

It is important, therefore, to bear in mind the intersectional nature of GBV. Many aspects can have a hand in influencing the course of violence, including poverty or health status [26]. Thus, although social wealth and the family's economic situation are drawn upon as the main characteristics involved, the wide variety of additional factors that can be found make a multidisciplinary approach necessary, incorporating environment factors, education, safety and security, health, and also, correlatively, the interaction of professionals in each sector [27]. Many of these variables are already taken into account when planning policies related to violence management [28].

Some previous papers have focused on the specific field of forecasting GBV by applying ML. In 2017, Thornton [29] tried to forecast domestic homicides and serious violence by using a database reflecting these situations in the county of Dorset (United Kingdom) and evaluating the police protocol. In doing so, he found that predicting deadly domestic violence dependent on insight from earlier police contacts did not seem possible at present, given the discovery that less than 50% of these cases had occurrences of earlier police contact and that, when contact occurred, the connections were evaluated by the protocol as not being of a high hazard in 89% of cases.

Chalkley and Strang [30] reproduced the methods used by Thornton and found false-negative risk assessments in 67% of the deadly violence cases that had prior contact with the police but were not classified within the existing protocol as high risk. They proposed that possible alternative predictors regarding sex, health, and other descriptors could improve the performance of the prediction. But this related data needs to be collected over a long period in order to obtain knowledge. Delgadillo-Alemán et al. [31] used the data provided by the Mexican Women's Institute, combined with other local organizations, and developed a mathematical deterministic model, which took into account variables like violence index, violence in childhood, the acceptance of machismo, and external factors, among others. By utilizing mathematical models, the authors showed their model's capability for diagnosing GBV risk in a certain couple. Although an interesting approach, its focus was on differential equations, so it does not explore the whole of society in a certain territory.

Spain, as with many other countries, has been gathering compelling and interesting data for decades relating to many aspects of society. In this country, we can find the previously mentioned INE which, in its current form, was founded in 1945, but its predecessor, the Kingdom's Statistical Commission, dates back to 1856. On its webpage, a range of time-series data is freely available and ready to be downloaded. Some authors have taken advantage of this availability in order to study the GBV phenomenon, using this database combined with other sources. De la Poza, Jódar, and Barreda [32], for example, proposed a mathematical model to infer hidden GBV incidence. Such work includes factors like the social awareness of men, age, drug consumption, and statistics of murdered women, all of which went into building a deterministic model and estimating the hidden population of aggressors. Unfortunately, however, this is not quantified by official statistics via which the accuracy of the model can be compared.

In conclusion, this review of the existing literature allows us to determine that the possibilities of ML have been widely proven as useful in making decisions related to social management, considering that these techniques are able to predict incidences of some public problems. We can also assess that the utility of ML in GBV is beyond doubt and, in this sense, some remarkable works have been identified. Despite this, however, we feel that the potential of ML in domestic violence forecasting for society as a whole is still unexplored and that the power of collected data is still insufficiently exploited. We think that, as previously shown regarding other disciplines, ML can be utilized to make useful GBV predictions for a certain territory, thereby optimizing the use of public resources. In this sense, to the best of the authors' knowledge, no ML-based study for the specific forecasting of GBV has been previously published that analyzes the features that most influence its appearance in a social group, that carries out a fair comparison between predictive ML algorithms applied to the same extensive database, and that considers differently populated territories. In any case, we have the feeling that many studies not have made a deep comparison of different methods of machine learning, both for selecting the most important variables and predictive techniques, and in this work we want to go

beyond the works presented and check, not only if it is possible to predict the incidence of gender violence, but also what technique would be more appropriate, making a comparison. Regarding these deficiencies, in this work, the issues mentioned are addressed through the use of a vast Spanish GBV database disaggregated by territories, which should allow for the proposed methodology to be applied to any other country/region/city.

3. Feature Selection and Forecasting Time Series

3.1. Feature Selection Techniques

Feature selection (FS) is the process of choosing the most relevant and pertinent features from an arrangement of features in a certain given dataset. For a dataset with d input features, the feature selection process brings about k features to such an extent that k < d, where k is the smallest arrangement of critical and applicable features [33]. This results in quicker ML algorithm training, the reducing of a model's complexity so it is simpler to decipher, better forecasting power, and the decreasing of overfitting by choosing the correct arrangement of features, among others.

There are three types of feature selection procedures [34]:

- Wrapper methods
- Filter methods
- Embedded methods

Wrapper methods use factor combinations to decide forecasting force. Normal wrapper strategies include: Subset Selection, Forward Stepwise Selection, and Backward Stepwise Selection (Recursive Features Elimination—RFE) [35]. The wrapper technique will locate the best mix of features, testing each variable against test models it builds with them to assess the outcomes [36]. Of the three strategies, this is more demanding computationally. In the Subset Selection strategy, we fit the model with every potential combination of N features [37]. With Forward Stepwise Selection, however, we first begin with a null model, i.e., beginning with one model variable. At this point, features are added one at a time with the best model picked depending on a metric (i.e., a valuation of the error) [36]. In this strategy, once the predictor is chosen, it never drops in the second step. This is done until the best subset of features is chosen, following a stopping criterion that establishes when the feature selection process must finish. In Backward Stepwise Selection (or Recursive Feature Elimination), the method works the opposite in that it wipes out features. As they are not run on each combination of features, they are less computationally concentrated by a significant degree when compared to straight Subset Selection [38]. Fundamentally, this is the inverse of Forward Stepwise selection. It begins with all predictors and, afterward, drops one feature at a time before selecting the best model. Likewise, the computational effort is fundamentally the same as that of Forward Selection. Filter and Wrapper strategies have been used and compared in some studies [39].

Filter methods are likewise considered as a Single Factor Analysis. By utilizing this technique, the predictive power of each individual variable (feature) is assessed, while different statistical methods can be utilized to decide predictive force [40]. One such pathway is achieved by correlating the feature with the objective (i.e., what we are foreseeing), with the features with the highest correlation being the most effective.

In contrast, Embedded Method (Shrinkage) is an inbuilt variable selection strategy, within which the features are not chosen or dismissed. With this approach, some value parameter controls (weights) are carried out, making it possible to name the LASSO (Least Absolute Shrinkage and Selection Operator) Regression. With this technique, regularization is carried out and some coefficients of a regression tend to be zero [41]. Therefore, as a portion of the coefficients tends to be equivalent to zero, we can drop or reject such variables. Another example is that of Ridge Regression (Tikhonov regularization), which includes a punishment that rises to the square of the greatness of coefficients [42]. All coefficients are shrunk by the same factor (so no single predictor is eliminated).

Some of these techniques will be applied in our work, in which we use a Multi-Objective Evolutionary Search Strategy [43] and also a Ranker Strategy [44], minimizing the metric that could be the Root Mean Squared Error (RMSE) and also reducing the features set. The two different types of approaches in these two groups are univariate and multivariate. Univariate methods are faster and easily scalable but ignore variable dependencies. On the other hand, multivariate techniques are able to model feature dependencies but are slower and less scalable than univariate ones [45]. The chosen techniques will be exposed in detail in the Methodology Section. By minimizing the metric, it is possible to improve the forecasting stage.

3.2. Forecasting

After the FS is complete, the forecasting task in time series can be deployed. In 1996, Wolpert [46] stated that, without deep information about the underlying model, there is no certain model that will always achieve better performance than any other. As a result, a proper approach can be made by trying out various techniques, then determining which model operates better. Consequently, we have compiled linear and nonlinear techniques, with a focus on the most promising algorithms.

Linear Regression is one of the easiest approaches. This family of models attempts to find an estimation of the model parameters so that the sum of the squared errors is minimized [47]. Some modifications include partial least squares and penalized models such as Ridge Regression or LASSO.

A significant advantage of these models is that they are highly interpretable. The coefficients indicate relationships and they are usually easy to compute, so the use of several features is affordable. On the other hand, they can be limited in their performance [48]. They achieve good results when the relationship between the predictors and their response falls along a hyperplane. However, if there are relations of a higher order, like quadratic, cubic, and alike, then the nonlinear relationships may not be properly captured with these models and so other approaches are required [49].

Some other models are capable of understanding nonlinear trends and, fortunately, the exact form of nonlinearity is not required to be known before building the model. Support Vector Machines (SVM) is of one the most popular examples in this category. These are dual learning algorithms that process data merely by computing their dot-products [50], and these dot-products between variable arrays can be properly computed by a kernel function [51]. Given this function, the SVM learner attempts to find a hyperplane that separates the examples while maximizing the separation (margin) between them. SVMs are well known to be resilient to over-fitting and to keep a good generalization performance due to the max-margin criterion used in the optimization process. In addition, while other solutions may only provide a local optimum, SVMs are guaranteed to converge to a global optimum because of the corresponding convex optimization formulation [52].

Besides this, Regression Trees make up a family of modeling algorithms that is getting a lot of attention in recent years. Tree-based models use one or more 'if-then' statements for the predictors that will subsequently partition the data. Within these subsets, a model is used to forecast the outcome [53]. From a statistical point of view, reducing correlation among predictors can be achieved by adding randomness to a tree construction process, which is the basis of the Random Forest (RF) technique [54]. Each model in the set is then used to build a prediction for a new dataset, with these predictions then being averaged to provide the final forecast.

An RF model performs a variance reduction by selecting complex and strong learners that exhibit low bias. This leads to an improvement in error rates and, in addition, RF is robust to a noisy response [55].

Other comparative strategies, for example, Gaussian Processes (GPs) with Radial Basis Function Kernels (RBF) [56]—which permit an overall consistency and a non-limited number of basic functions—are infrequently utilized, albeit a few previous approaches have used this strategy with promising conclusions [57].

GPs represent a nonparametric methodology focused on modeling perceptible reactions from different training data points (function values) as multivariate normal random features [58]. A supposition is made of a priori distribution for such function data values, which will ensure the function's smoothness properties. To be explicit, there will be a high correlation between the two function values when there is closeness (in the feeling of Euclidean separation) between the comparing input vectors and if they decay as they diverge. Later, the distribution of unpredicted function data may be calculated from the use of an assumed distribution with the application of simple probability manipulation.

4. Database, Available Features, and Target to Be Forecasted

In order to study the relationship between different time series with data regarding GBV and other variables, we have accessed the database of Spanish INE (www.ine.es), where it is possible to freely find data from several decades related to demography, economy, employment market, education, energy, and so on. The data series are usually grouped by population groups (always disaggregated by sex), and also by territorial units. The frequency of data reporting can be monthly, quarterly, or annually. For our purposes, we have assembled the data by provinces and also the country's total. Spain has 50 provinces and 2 autonomous cities. We decided to select some examples as study cases then compare them with the evolution of the total country. With regard to the timescales, we decided to divide the data monthly as, on the one hand, this is the usual way of presenting the data in our database and, on the other hand, it offers sufficient granulometry to show the variable evolution. In Table 1, a brief description of the variables and the units utilized can be observed. All variables have been referred to population units (per capita) in order to make a fair comparison between territories. Although the Spanish Government has been collecting GBV casualties' data since 2003, we begin our database in 2009 in order to obtain a complete overview by avoiding any gaps in the early data of some variables. We also seek to reflect the changes introduced by the Organic Law on Measures Regarding Comprehensive Protection against Gender-Based Violence (LO 1/2004, December 28, 2004), some of whose measures were not fully implemented until some years later. Likewise, the most recent data may have a delay in their incorporation and be subject to revision, so full series data up to March 2020 have been used. With this approach, we have studied more than a decade of data.

The chosen features (among the available ones) are related to:

- Territorial: We study the time-series data for the entire country but also some provinces as examples, in order to test the validation of our purpose.
- Date and season: We will explore the evolution of GBV within years, month by month. We will also include the quarter to evaluate the influence of the season, as indicated by previous works [59].
- Demography and population: Considering population can offer insights into the influence of big population areas, but some changes in demography can also provide explanations of the course of couples [60]. In this manner, marriages, separations, and births are included, but also the proportion of men vs. women.
- Specific variables related to GBV: In this sense, there are some interesting variables available, such as:
    - Calls to the special number 016. This is a phone number dedicated to providing information to survivors, but also to manage assistance (imperative or not).
    - Complaints: In particular, we will study the number of complaints presented to a court as the independent variable to be modeled and forecasted. Ultimately, we feel that complaints express the incidence of worst cases.
    - Security devices for tracking offenders: This kind of device is proposed by a judge in high-risk cases.
    - Protection orders: Also ordered by a judge in cases of high risk.

- Level of risk of aggression for the survivor: After a police evaluation, the cases are classified as unappreciated, low, medium, high, and extremely high.
- Fatalities: Murdered victims of GBV.

- Wealth and employment: The level of wealth in a region can be related to the levels of crime and violence. Similarly, levels of unemployment (male and female) can give an idea of the level of economic stability [61]. We differentiate between the inactive population (retired, disabled) and also the employed and unemployed population.
- Education level: The relationship of illiteracy (male and female) and other educational levels (primary, secondary, university) with violence will also be studied, as previous literature indicates this point [62].

Table 1. Description of the features.

Variable	Description	Units
PROVINCE	Spanish province under study (or the whole country)	(Categorical)
DATE	Data collection date	Month
QUARTER	Quarter of the year	Quarter
YEAR	Year of data collection	Year
POP_TOT	Total population of the province	Units
RATIO_MvsW	Ratio Population of men/women	Adimensional
MARRIAGES	Number of new weddings	Units/10,000 pop
SEPARATIONS	Number of separated marriages	Units/100,000 pop
BIRTHS	Number of newborn children	Units/1000 pop
CALLS	Calls to special telephone number 016 (requests for information and assistance)	Units/10,000 pop
COMPLAINTS	Complaints made to a Court	Units/10,000 pop
DEVICES	Security devices for tracking offenders	Units/100,000 pop
PROTECTION_ORDER	Restraining order for survivors decreed by a judge	Units/10,000 pop
RISK_UN	Survivors with unappreciated risk after police valuation	Units/10,000 pop
RISK_L	Survivors with low risk after police valuation	Units/10,000 pop
RISK_M	Survivors with medium risk after police valuation	Units/10,000 pop
RISK_H	Survivors with high risk after police valuation	Units/10,000 pop
RISK_EH	Survivors with extremely high risk after police valuation	Units/10,000 pop
FATALITIES	Murdered victims of GBV	Units/1,000,000 pop
GDP	Gross Domestic Product per capita	€/10,000 pop
EMPL_MEN	Employed men	Units/100 pop
UNEMPL_MEN	Unemployed women	Units/100 pop
INACT_MEN	Inactive women	Units/100 pop
EMPL_WOM	Employed women	Units/100 pop
UNEMPL_WOM	Unemployed women	Units/100 pop
INACT_WOM	Inactive women	Units/100 pop
ILLIT_MEN	Illiterate men	Units/100 pop
ILLIT_WOM	Illiterate women	Units/100 pop
PRIM_ED_MEN	Primary education men	Units/100 pop
SEC_ED_MEN	Secondary education men	Units/100 pop
HIGH_ED_MEN	Higher education men	Units/100 pop
PRIM_ED_WOM	Primary education women	Units/100 pop
SEC_ED_WOM	Secondary education women	Units/100 pop
HIGH_ED_WOM	Higher education women	Units/100 pop

GBV: Gender-Based Violence.

With this, we have built a database of almost 250,000 data examples, taking into account all the months from January 2009 to March 2020 and the 52 Spanish territories plus the whole country. Using all this data, we will carry out a feature selection and then forecast future GBV complaints.

5. Methodology

5.1. Territories under Study

As previously stated, a large amount of data is under consideration. The purpose of this work is to study the possibility of forecasting GBV complaints so as to provide reliable information to optimize public resources and to schedule actions in advance, thereby being useful for other countries—with the necessary adjustments. In this sense, instead of proving our methodology in each and every Spanish province, we will test our proposed procedure in some particular cases: the whole country (Spain) and the three representative provinces of Madrid, Alicante, and Segovia—representing large, medium, and low populations, respectively. In addition, each of them has its own and differentiated characteristics in terms of location and economy, as well as idiosyncrasies and cultural aspects.

- Spain: A Mediterranean country and member of the European Union. The total population consists of 47,329,981 people.
- Madrid: locating the homonymous capital city of Spain, with a population of 6,661,949 people, is centered on the country's map and has a dynamic economy.
- Alicante: In the east of Spain with a population of 1,858,683 people. It has a marked open and Mediterranean character, medium-range age inhabitants, and a flourishing economy.
- Segovia: An inland province located in the west of Spain with a population of only 153,342 people and an aging population.

5.2. The Waikato Environment for Knowledge Analysis (WEKA)

The Waikato Environment for Knowledge Analysis (WEKA v.3.8) is free software developed at the University of Waikato, New Zealand (https://waikato.github.io/weka-wiki/) and licensed under the GNU (*GNU's Not Unix*) General Public License. WEKA contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces that offer easy access to these functions. The software also supports several standard data-mining tasks—more specifically, data preprocessing, clustering, classification, regression, visualization, feature selection, modeling, and forecasting.

The use of WEKA facilitates data entry, algorithm execution, and visual context in the management of the entire process. This software has been successfully applied many times before and is still being applied in recent literature. Thus, Hussain et al. used it in 2018 to study educational aspects with data mining techniques [63], or Kiranmai et al. to classify electrical power problems [64]. WEKA software is booming, and new modules are developed every year, like the one presented by Lang et al. in their 2019 work on deep learning [65].

5.3. Computer Hardware

Due to the computational demands of the ML algorithms, they have been executed using a computer equipped with an AMD Ryzen 7 1700X processor, operating at 3.8 GHz with 32 GB DDR4 RAM at 2666 MHz CL19 and a Solid-State Disk Samsung 970 Evo Plus M.2 1000 GB PCI-E 3.0.

5.4. Data Cleaning, Regularization, and Lagged Variables

The above-mentioned database should be transformed in due course to provide the proper inputs of feature selection algorithms. The values are cleaned, and some gaps have been completed. In order to later make a fair comparison between the whole country and some provinces (which will become study cases), all the features were divided by the population (hundreds, thousands, tens of thousands).

As some features could have a delayed influence, they were given six lagged values (which means taking into account the last six months), except for the categorized and date features. The *TimeSeriesLagManager* routine of WEKA allows for easily creating as many lagged variables as required.

5.5. Features Selection

WEKA offers an intuitive graphical environment for carrying out a feature selection. The module *AttributeSelection* allows for specifying different Search Methods and Attribute Evaluators, with some combinations being tested and then evaluated at the forecasting stage. The features set that provides a more accurate prediction will become the chosen features. A short introduction to FS methods was presented in Section 3.1.

5.5.1. Search Methods

As stated, we use two searching methods: The Multi-Objective Evolutionary Search Strategy and also a Ranker Strategy.

- Multi-Objective Evolutionary Search Strategy (MOES): In particular, we execute the multi-objective evolutionary algorithm known as the Evolutionary NOn-dominated Radial slots-based Algorithm (ENORA) as a selection strategy for a random search method, which minimizes the selected features and also the RMSE [66].
- Ranker: This search strategy makes ranks of features one by one by utilizing their evaluations [67].

5.5.2. Attribute Evaluators

From the feature selection methods offered by WEKA, we will choose the two most popularly used:

- Wrapper methods. The *WrapperSubsetEval* routine implemented in WEKA will allow us to evaluate some approaches via multivariate techniques. For univariate ones, we need to instead use the *ClassifierAttributeEval* procedure. We will execute the following predictors:
    - Linear Regression: This offers fast computation, fixing the coefficients for each feature.
    - Random Forest [68]: As stated earlier, this is a tree-based algorithm well-known for classification purposes.
    - Instance-Based K-nearest neighbor algorithm (IBk) [69]: A K-nearest neighbors classifier, this algorithm allows for selecting an appropriate value of K based on cross-validation but is also able to carry out distance weighting.
- Filter Method. On the side of the univariate methods, we will use the *Ranker* operation according to the below predictors:
    - Relief Attribute (Rlf) [70]: Relief feature selection is based on scoring by the identification of feature value differences between the nearest neighbor instance pairs.
    - Principal Component Analysis (PCA) [71]: With this technique, a new set of orthogonal coordinate axes is introduced, and, at the same time, the sample data variance is maximized. This leads to the scenario that the other directions, in which the variance is minor, are less important and, hence, can be removed from the dataset. PCA offers a very effective way of transforming the data in a lower dimensionality, while also being able to reveal some simplified patterns that often underlie the data.

5.5.3. Generated Subsets

With the exposed techniques, combined as indicated in Table 2, we can generate seven subsets of reduced data that will be under evaluation in the forecasting task. In all the exposed cases of FS,

the metric to be optimized is the RMSE. In addition, we will study the prediction strategies for the original dataset, which are exposed in the following subsection. Table 3 compiles the different commands used in WEKA, presenting the used parameters.

Table 2. Applied Features Selection techniques.

Search Method	Attribute Evaluator	Predictor	Acronym
MOES	Wrapper	Linear Regression	MOES-LR
		Random Forest	MOES-RF
		IBk	MOES-IBk
Ranker	Wrapper (Classifier)	Linear Regression	Rnk-LR
		Random Forest	Rnk-RF
	Filter	Relief	Rnk-Rlf
		PCA	Rnk-PCA

MOES: Multi-Objective Evolutionary Search Strategy; *LR*: Linear Regression; *RF*: Random Forest; *IBk*: Instance-Based K-nearest neighbor; *Rlf*: Relief Attribute; *PCA*: Principal Component Analysis.

Table 3. WEKA commands for Feature Selection.

Technique	Command
MOES	*weka.attributeSelection.MultiObjectiveEvolutionarySearch -generations 20 -population-size 100 -seed 1 -algorithm 0 -report-frequency 20 -log-file "C:\\Program Files\\Weka-3-8"*
Ranker	*weka.attributeSelection. Ranker -T -1. 8 -N -1*
Wrapper LR	*weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.functions.LinearRegression -F 5 -T 0.01 -R 1 -E RMSE – -S 0 -R 1.0E-8 -num-decimal-places 4*
Wrapper RF	*weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.trees.RandomForest -F 5 -T 0.01 -R 1 -E RMSE – -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 -num-decimal-places 4*
Wrapper IBk	*weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.lazy.IBk -F 5 -T 0.01 -R 1 -E RMSE – -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\"" -num-decimal-places 4*
Classifier LR	*weka.attributeSelection.ClassifierAttributeEval -execution-slots 1 -B weka.classifiers.functions.LinearRegression -F 5 -T 0.01 -R 1 -E RMSE – -S 0 -R 1.0E-8 -num-decimal-places 4*
Classifier RF	*weka.attributeSelection.ClassifierAttributeEval -execution-slots 1 -B weka.classifiers.trees.RandomForest -F 5 -T 0.01 -R 1 -E RMSE – -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1 -num-decimal-places 4*
Relief	*weka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10*
PCA	*weka.attributeSelection.PrincipalComponents -R 0.95 -A 5*

MOES: Multi-Objective Evolutionary Search Strategy; *LR*: Linear Regression; *RF*: Random Forest; *IBk*: Instance-Based K-nearest neighbor; *Rlf*: Relief Attribute; *PCA*: Principal Component Analysis.

5.6. Data Modeling and Forecasting

Once the seven reduced data subsets have been achieved, plus the original dataset, we will try to make a prediction of future values, taking into account the past time series collected in each dataset. We will attempt to forecast the complaints regarding GBV for a predictive horizon of six months but, in order to have the real data to evaluate/validate the prediction, we will apply a Cross-Validation (CV) method designed for time series [72]. We will train with a subset and then forecast the next six months/steps (for which the data is available but not included in the training dataset).

For this purpose, we use the *time series Forecasting* (http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka) module of WEKA (v. 1.027).

For each dataset, we use the following approaches as forecasting algorithms, always indicating the accuracy in terms of RMSE.

- Linear Regression (LR).
- Support Vector Machines (SVM).
- Random Forest (RF).
- Gaussian Process (GP).

A description of each method can be found in Section 3.2. Table 4 expresses the WEKA commands and the parameters of each technique.

Table 4. WEKA commands for forecasting.

Technique	Command
LR	*weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4*
RF	*weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1*
SVM	*weka.classifiers.functions.SMOreg -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001 -W 1" -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007"*
GP	*weka.classifiers.functions.GaussianProcesses -L 1.0 -N 0 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -S 1*

LR: Linear Regression; *RF*: Random Forest; *SVM:* Support Vector Machines; *GP:* Gaussian Process.

6. Results and Discussion: Forecasting Performance

The FS procedure generates seven subsets which, together with the original complete dataset, will be tested using four forecasting techniques. In this way, we will execute 32 prediction tests for the whole country (Spain), forecasting six months of the time series of GBV-complaints. In each experiment, the predictive algorithm will first be trained with a subset of the data and then a predictive horizon of the next six months/steps will be forecast, executing a CV. An example of this first phase resulting in a trained model is shown in Figure 1, for the specific case of the subset MOES-RF and using RF as a predictive technique. At a glance, it can be seen that there is a cyclic stationary behavior combined with a certain tendency.

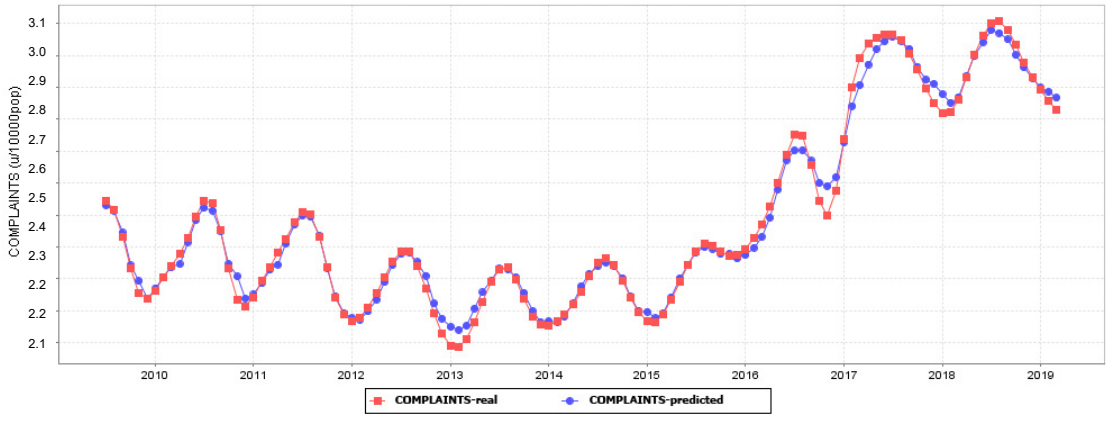

Figure 1. Training phase with RF algorithm of the subset MOES-RF for GBV complaints data series.

The results of the forecasting task can be found in Table 5. With each predictive algorithm, and for each subset of data, we calculate the accuracy of the next six months/steps of the GBV complaints series. Using the CV technique, we can obtain the RMSE for each future step, then, as a measure of performance,

obtain an average of the six values of RMSE regarding each FS technique ($\overline{\text{RMSE}}$). The standard deviation is also estimated in each forecasted series in order to infer the accuracy's variability.

Table 5. RMSE to 6-step GBV complaints forecasting in Spain.

RMSE								
Subset FS	**1 Step**	**2 Step**	**3 Step**	**4 Step**	**5 Step**	**6 Step**	$\overline{\textbf{RMSE}}$	**Standard Deviation**
Forecasting technique: LR								
No F.S.	0.2309	0.4502	0.6627	0.8719	1.0802	1.2785	0.7624	0.3922
MOES-LR	0.2229	0.2784	0.2938	0.3059	0.3224	0.3418	0.2942	0.0413
MOES-RF	0.1273	0.1784	0.1985	0.2015	0.2069	0.2093	0.1870	0.0312
MOES-IBk	0.0914	0.1878	0.2853	0.3761	0.4576	0.5227	0.3202	0.1638
Rnk-LR	0.1033	0.2034	0.2932	0.3589	0.4012	0.4039	0.2940	0.1203
Rnk-RF	0.1198	0.2058	0.2632	0.3014	0.3292	0.3493	0.2615	0.0861
Rnk-Rlf	0.3860	0.4017	0.4195	0.4362	0.4442	0.4253	0.4188	0.0217
Rnk-PCA	0.2289	0.3227	0.3619	0.3859	0.4081	0.4315	0.3565	0.0729
						$\overline{\textbf{RMSE}}$	**0.3618**	
Forecasting technique: RF								
No F.S.	0.2083	0.2407	0.2513	0.2635	0.2748	0.2747	0.2522	0.0253
MOES-LR	0.1680	0.1808	0.1867	0.1943	0.2047	0.2117	0.1910	0.0160
MOES-RF	0.1489	0.1586	0.1646	0.1714	0.1806	0.1876	0.1686	0.0143
MOES-IBk	0.1803	0.1941	0.2012	0.2104	0.2214	0.2275	0.2058	0.0176
Rnk-LR	0.1824	0.1919	0.1978	0.2055	0.2166	0.2246	0.2031	0.0157
Rnk-RF	0.1605	0.1801	0.1866	0.1930	0.2034	0.2125	0.1894	0.0183
Rnk-Rlf	0.1919	0.2047	0.2121	0.2219	0.2333	0.2395	0.2172	0.0179
Rnk-PCA	0.1820	0.1960	0.2047	0.2144	0.2258	0.2343	0.2095	0.0193
						$\overline{\textbf{RMSE}}$	**0.2046**	
Forecasting technique: SVM								
No F.S.	0.3825	0.4632	0.4879	0.5067	0.5307	0.5600	0.4885	0.0618
MOES-LR	0.1706	0.2204	0.2372	0.2489	0.2621	0.2722	0.2352	0.0365
MOES-RF	0.0987	0.1580	0.1913	0.1999	0.2082	0.2118	0.1780	0.0434
MOES-IBk	0.1782	0.2765	0.3288	0.3603	0.3838	0.4056	0.3222	0.0837
Rnk-LR	0.0759	0.1420	0.1929	0.2198	0.2314	0.2292	0.1819	0.0618
Rnk-RF	0.1320	0.1648	0.1850	0.1922	0.1997	0.1997	0.1789	0.0264
Rnk-Rlf	0.1292	0.2462	0.3460	0.4273	0.4974	0.5583	0.3674	0.1605
Rnk-PCA	0.1321	0.2579	0.3762	0.4865	0.5869	0.6711	0.4185	0.2032
						$\overline{\textbf{RMSE}}$	**0.2963**	
Forecasting technique: GP								
No F.S.	0.3402	0.3823	0.3922	0.4005	0.4156	0.4383	0.3949	0.0332
MOES-LR	0.1540	0.2160	0.2325	0.2344	0.2403	0.2531	0.2217	0.0353
MOES-RF	0.1325	0.1648	0.1735	0.1813	0.1898	0.1913	0.1722	0.0219
MOES-IBk	0.1694	0.2171	0.2325	0.2443	0.2560	0.2611	0.2301	0.0337
Rnk-LR	0.1513	0.2118	0.2317	0.2373	0.2469	0.2603	0.2232	0.0388
Rnk-RF	0.1720	0.2125	0.2220	0.2276	0.2374	0.2512	0.2205	0.0272
Rnk-Rlf	0.3075	0.3525	0.3649	0.3758	0.3927	0.4161	0.3683	0.0371
Rnk-PCA	0.2479	0.2989	0.3120	0.3229	0.3396	0.3592	0.3134	0.0384
						$\overline{\textbf{RMSE}}$	**0.2680**	

F.S.: Feature Selection; *LR*: Linear Regression; *RF*: Random Forest; *SVM*: Support Vector Machines; *GP*: Gaussian Process; *MOES*: Multi-Objective Evolutionary Search Strategy; *Rnk*: Ranker; *IBk*: Instance-Based K-nearest neighbor; *Rlf*: Relief Attribute; *PCA*: Principal Component Analysis.

A Shapiro–Wilk test was used to determine whether the data presented a normal distribution for each 6-steps prediction. The results indicated that the data was normally distributed (p-values > 0.05).

To compare the similarity of the eight forecasted series per prediction technique, we performed the parametric Welch's *T*-test. The results indicated that the 6-steps evolution of the original dataset differed significantly from the other results (p-value $< \alpha = 0.05$) in the four cases (LR, RF, SVM, and GP).

As can be seen in Table 5, the lower RMSE averaged between steps ($\overline{\text{RMSE}}$) is obtained using RF as a foresight algorithm with the MOES-RF dataset ($\overline{\text{RMSE}}$ = 0.1686 u/10,000 pop). Other FS approaches are also promising. Figure 2 shows different predictions using RF from all of the datasets. Rnk-RF also provides an accurate result in most of the predictive situations, which can indicate that the use of RF as a predictor in each attribute evaluator is an interesting choice. MOES-LR closely follows the performance in this particular situation under analysis.

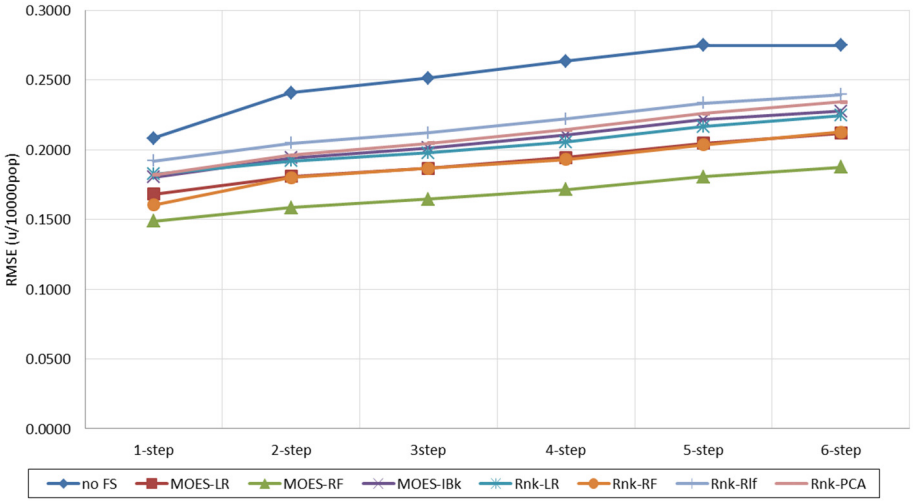

Figure 2. Comparative evolution of GBV complaints' RMSE evolution obtained via an RF predictive algorithm with different FS techniques.

But, as stated in the results, MOES-RF seems to be the better dataset for utilization. According to Table 5, with all the four different predictive techniques, the best accuracy is achieved with this subset. When using this FS combination, the best predictive algorithm is RF, as depicted in Figure 3, where it can be seen that RF offers the best average accuracy because of the low standard deviation in the RMSE in each step—which results in performance stability. SVM is able to achieve a better performance in short prediction, as well as LR, but they soon increase errors in future steps—which can be inferred by their bigger standard deviation.

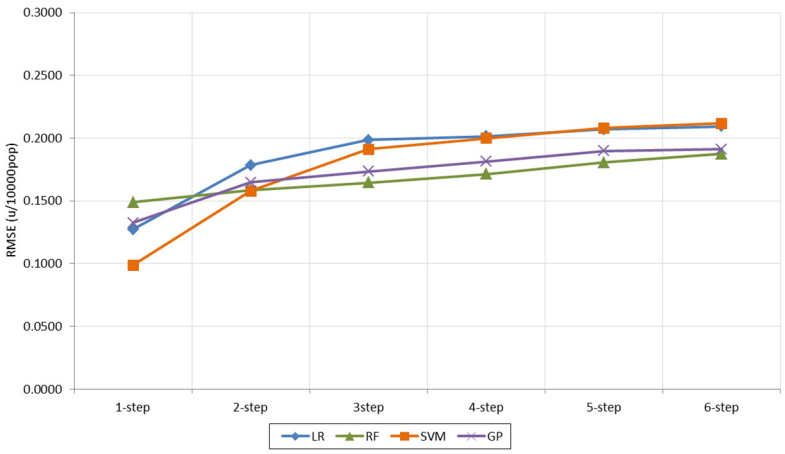

Figure 3. Comparative evolution of GBV complaints' RMSE evolution obtained with MOES-RF feature selection technique for different predictive algorithms.

The better forecasting results with RF can be also inferred by averaging the averaged RMSE ($\overline{\text{RMSE}}$), expressed by $\overline{\overline{\text{RMSE}}}$, and then by making a comparison between prediction methods. RF then offers an $\overline{\overline{\text{RMSE}}}$ = 0.2046 u/10,000 pop, followed by GP with $\overline{\overline{\text{RMSE}}}$ = 0.2680 u/10,000 pop. SVM results in variable performance (as shown in Figure 3), averaging 0.2963 u/10,000 pop. LR cannot follow the accuracy of the other algorithms even closely.

We need to bear in mind that, for a population of 47,329,981 inhabitants, the achieved $\overline{\overline{\text{RMSE}}}$ for RF is equal to a RMSE of 968.37 complaints for all the country (per month). Taking into account that the average of complaints in 2019 per month in Spain was 14,014 petitions, the error is around 6% and good enough for our purposes.

Taking into account these results, it is possible to make a practical check of the prediction evolution in the months from October 2019 to March 2020 using MOES-RF with RF as a forecasting technique, before comparing the results with the real evolution. As can be observed, the prediction is accurate enough to identify trends, while logically being more separated from the real curve as we progressively expand the predictive horizon (Figure 4).

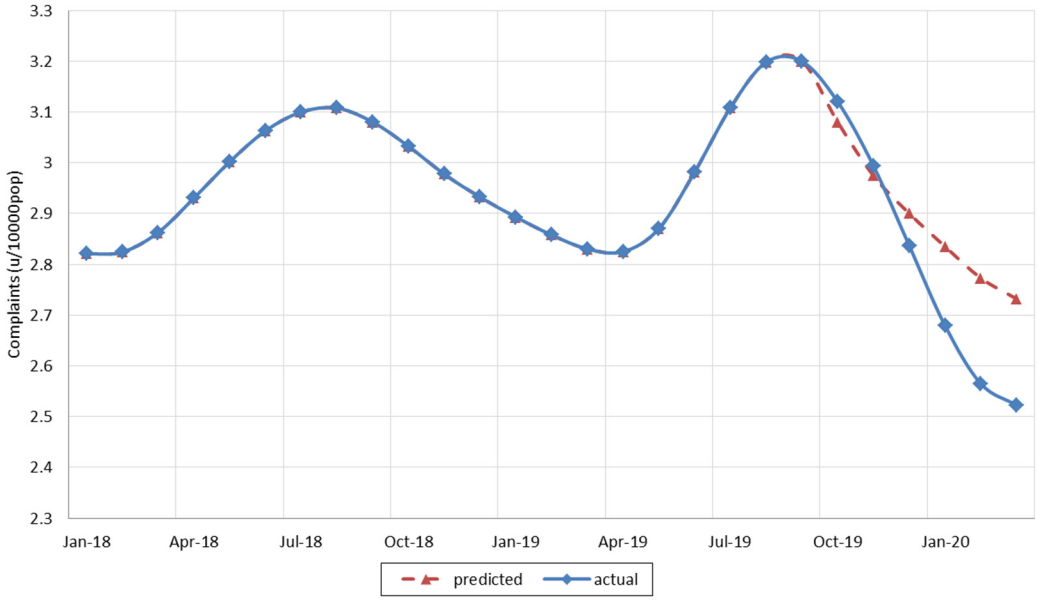

Figure 4. Prediction for the months October 2019 to March 2020 and real data comparison. Feature selection: MOES-RF. Forecasting technique: RF.

Following the same procedure with the selected Spanish provinces of different populations (Madrid, Alicante, and Segovia), we can confirm that similar results can be found. Table 6 summarizes the $\overline{\text{RMSE}}$ of the 6-step predictions carried out with the four predictive techniques and with each of the eight data subsets.

Table 6. Average RMSE to 6-step GBV complaints forecasting in different territories.

Subset FS	Spain		Madrid		Alicante		Segovia	
	($\overline{\text{RMSE}}$ 6-Steps)	**Standard Deviation**	**($\overline{\text{RMSE}}$ 6-Steps)**	**Standard Deviation**	**($\overline{\text{RMSE}}$ 6-Steps)**	**Standard Deviation**	**($\overline{\text{RMSE}}$ 6-Steps)**	**Standard Deviation**
Forecasting technique: LR								
No F.S.	0.7624	0.3922	1.1144	0.5175	1.1849	0.2454	1.7304	0.5124
MOES-LR	0.2942	0.0413	0.3430	0.1167	0.4724	0.1636	0.5129	0.1208
MOES-RF	0.1870	0.0312	0.2709	0.0066	0.3974	0.0704	0.3104	0.1108
MOES-IBk	0.3202	0.1638	0.5108	0.2423	0.9960	0.3369	0.8974	0.1692
Rnk-LR	0.2940	0.1203	0.3519	0.1406	0.8525	0.0843	0.3370	0.0107
Rnk-RF	0.2615	0.0861	0.3153	0.0493	0.4081	0.0047	0.2756	0.1074
Rnk-Rlf	0.4188	0.0217	0.5127	0.1821	1.1618	0.3647	1.2120	0.1902
Rnk-PCA	0.3565	0.0729	0.6136	0.0748	1.0383	0.3725	1.1588	0.7191
$\overline{\overline{\text{RMSE}}}$	**0.3618**		**0.5041**		**0.8139**		**0.8043**	
Forecasting technique: RF								
No F.S.	0.2522	0.0253	0.3198	0.0208	0.4245	0.0384	0.7339	0.0178
MOES-LR	0.1910	0.0160	0.3047	0.0224	0.3922	0.0360	0.6163	0.0118
MOES-RF	0.1686	0.0143	0.2928	0.0258	0.3776	0.0297	0.5756	0.0127
MOES-IBk	0.2058	0.0176	0.3051	0.0245	0.3978	0.0350	0.6592	0.0144
Rnk-LR	0.2031	0.0157	0.2990	0.0171	0.3804	0.0418	0.6195	0.0160
Rnk-RF	0.1894	0.0183	0.2921	0.0151	0.3704	0.0303	0.5798	0.0154
Rnk-Rlf	0.2172	0.0179	0.3158	0.0155	0.4026	0.0348	0.7262	0.0288
Rnk-PCA	0.2095	0.0193	0.3113	0.0218	0.4047	0.0324	0.6665	0.0231
$\overline{\overline{\text{RMSE}}}$	**0.2046**		**0.3051**		**0.3938**		**0.6471**	
Forecasting technique: SVM								
No F.S.	0.4885	0.0618	0.8038	0.2394	1.4879	0.4664	1.7207	0.7698
MOES-LR	0.2352	0.0365	0.3276	0.1057	0.4940	0.0888	0.4714	0.0987
MOES-RF	0.1780	0.0434	0.2508	0.0329	0.3407	0.0620	0.3418	0.1222
MOES-IBk	0.3222	0.0837	0.4588	0.1980	0.5229	0.1560	0.5738	0.1021
Rnk-LR	0.1819	0.0618	0.4583	0.1478	0.4866	0.1523	0.5341	0.4749
Rnk-RF	0.1789	0.0264	0.2620	0.0866	0.2975	0.0681	0.4103	0.2492
Rnk-Rlf	0.3674	0.1605	0.5824	0.0747	0.8858	0.3994	1.0350	0.1614
Rnk-PCA	0.4185	0.2032	0.4866	0.1357	0.7613	0.0778	0.6652	0.3641
$\overline{\overline{\text{RMSE}}}$	**0.2963**		**0.4538**		**0.6596**		**0.7190**	

Table 6. *Cont.*

Subset FS	Spain		Madrid		Alicante		Segovia	
	($\overline{\text{RMSE}}$ 6-Steps)	**Standard Deviation**	**($\overline{\text{RMSE}}$ 6-Steps)**	**Standard Deviation**	**($\overline{\text{RMSE}}$ 6-Steps)**	**Standard Deviation**	**($\overline{\text{RMSE}}$ 6-Steps)**	**Standard Deviation**
Forecasting technique: GP								
No F.S.	0.3949	0.0332	0.4966	0.0806	0.6799	0.0794	0.5454	0.0686
MOES-LR	0.2217	0.0353	0.3650	0.0664	0.3271	0.0738	0.3644	0.0682
MOES-RF	0.1722	0.0219	0.3013	0.0318	0.2602	0.0377	0.3033	0.0376
MOES-IBk	0.2301	0.0337	0.3709	0.0353	0.5322	0.0766	0.3788	0.0513
Rnk-LR	0.2232	0.0388	0.3136	0.0378	0.3299	0.0442	0.3116	0.0321
Rnk-RF	0.2205	0.0272	0.2762	0.0541	0.3192	0.0493	0.2899	0.0384
Rnk-Rlf	0.3683	0.0371	0.4226	0.0426	0.6502	0.1130	0.4225	0.0826
Rnk-PCA	0.3134	0.0384	0.3945	0.0716	0.6471	0.0606	0.4656	0.0939
$\overline{\overline{\text{RMSE}}}$	**0.2680**		**0.3676**		**0.4682**		**0.3852**	
Average among techniques $\overline{\overline{\text{RMSE}}}$	**0.2826**		**0.4076**		**0.5839**		**0.6389**	

F.S.: Feature Selection; *LR*: Linear Regression; *RF*: Random Forest; *SVM:* Support Vector Machines; *GP:* Gaussian Process; *MOES*: Multi-Objective Evolutionary Search Strategy; *Rnk:* Ranker; *IBk*: Instance-Based K-nearest neighbor; *Rlf*: Relief Attribute; *PCA*: Principal Component Analysis.

From the results shown in Table 6, valuable conclusions can be obtained. MOES-RF appears to be the best FS technique, with RF being the best forecasting algorithm. This performance is consistent and to be expected as the time-series data from each province makes up the subsets of the whole country or, seen in a different light, the analyzed data from Spain makes up the sum of the provinces. Nevertheless, we have to highlight that $\overline{\overline{\text{RMSE}}}$ is higher for each predictive technique when the province is less populated. This can be explained because, when we are considering the whole of Spain or Madrid (millions of people), the consistency of some data is flush, and in some way occasional and punctual situations are retailed averaged in a big population, showing a smoother evolution of the social variables. On the other hand, with a low population, every single fluctuation stands out more, resulting in more variability in the prediction and, hence, higher error and a bigger standard deviation. This particular case can be easily appreciated by studying the table, showing that $\overline{\overline{\text{RMSE}}}$ corresponding to each technique is around three times bigger than the whole country when applied to Segovia—0.2963 to 0.7190 u/10,000 pop (Spain and Segovia, respectively) when using SVM, or 0.2680 to 0.3852 u/10,000 pop (Spain—Segovia) when using GP. To deepen this idea, we average the $\overline{\overline{\text{RMSE}}}$ of each technique by territory one more time (referred to as $\overline{\overline{\overline{\text{RMSE}}}}$ in Table 6), allowing us to appreciate this evolution clearly, growing from 0.2826, 0.4076, 0.5839, and 0.6389 u/10,000 pop (Spain, Madrid, Alicante, and Segovia, respectively).

Although, for the sake of simplicity, predictions are not detailed in every step of the provinces' comparison. Figure 5 shows the instance evolution or RMSE in each territory forecasting with MOES-RF (chosen as FS) and RF. As can be seen, all of the forecasting steps show the stability of RF as a predictive algorithm, although higher values and a more oscillating prediction can be observed in the case of Segovia.

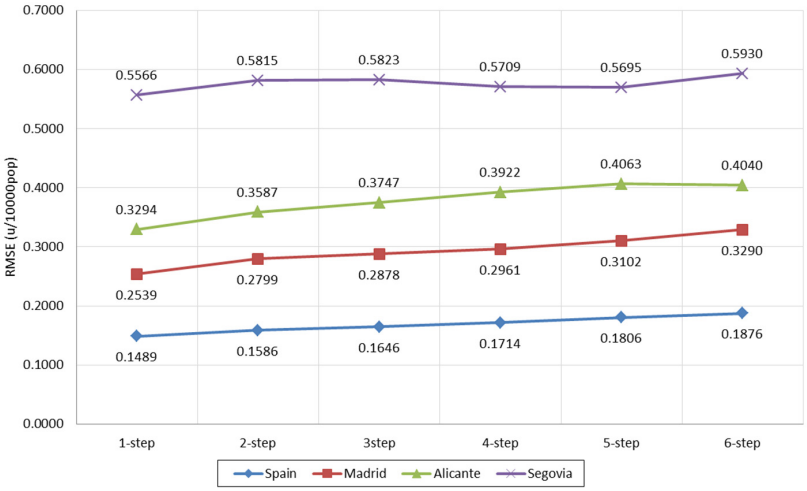

Figure 5. Evolution of RMSE in 6-step GBV complaints' forecast, performed by FS dataset MOES-RF and RF as predictive techniques for different territories.

Considering these results and the discussion, some important conclusions can be made in the next section.

7. Conclusions and Future Works

GBV makes for one of the great unresolved problems of our time that require urgent attention. Allocating resources of all kinds is essential for tackling these situations before they occur, allowing authorities to anticipate and act before the aggressions take place.

It is necessary, therefore, to consider the extent to which we can predict the incidence of this violence in order to optimize resources and allocate the necessary means in the most appropriate manner,

both in time and space. Until now, achieving such a forecast has proven complex, but the periodic collection of data that reflects the state and evolution of society, together with the increased knowledge and applicability of ML algorithms, provides a new avenue for addressing the challenge of predicting the temporal evolution of gender violence.

In this work, the possibility of predicting the reports of gender violence with acceptable accuracy has been proven, with the most appropriate technique for selecting variables and the best predictive algorithm performance having been discussed. After testing eight sets with four known predictive techniques, it has been found that the most appropriate technique is one that combines MOES-RF as a variable selection with RF to predict future values. This conclusion has been obtained by using the data corresponding to the whole of Spain from January 2009 to September 2019, which has been corroborated by comparing it with certain provinces of the country with differing populations, such as Madrid, Alicante, and Segovia—each with a particular casuistry.

Given the difference in population of the provinces studied, as well as their different geographical situation, an adequate prediction per province allows for a correct distribution of the available state resources, so that awareness campaigns, police intervention, as well as economic resources and other social policies are distributed over time in a more efficient manner. Although it can be inferred from the study that there is seasonality in general, the maximum incidence from one province to another may differ, which, thanks to the results of this work, will allow for more adjusted planning in the provincial distribution of resources.

Other combinations of FS and predictive algorithms are also promising and may also be useful. Although there is consistency between the behavior of ML techniques in each territory, it has been shown that errors increase when the population decreases, as well as the error dispersion (greater variation), giving an impression that, the larger the population, the greater consistency in the data collected, which will reflect not a particular circumstance in time, but the presence of underlying circumstances with predictable cause–effect relationships. A smaller study population will mean isolated circumstances marking the oscillation of certain variables more significantly—a dynamic that will be attenuated in larger populations.

In any case, this work intends, rather than showing concrete results in a specific period of time in Spain and some of its provinces, to present a specific methodology and to study its viability. With the conclusions drawn, we aim to serve as a basis for studies similar to ours in other countries/territories with comparable (or other) variables to be taken into account. In this sense, some other public databases could validate the proposed methodology. The European Institute for Gender Equality, in its Gender Statistics Database (https://eige.europa.eu/gender-statistics/dgs), provide several data that can be used for validation purposes.

For this reason, future work should look to test other combinations of attribute selection and prediction, as well as replicating our method to address other social issues involving a large number of people (migration, education, consumption, economy, etc.), and continuing to check the performance of the work described here.

Author Contributions: Conceptualization, I.R.-R., J.-V.R., P.H.-G. and D.-J.P.-Q.; methodology, I.R.-R., P.H.-G., J.-V.R. and D.-J.P.-Q.; software, I.R.-R. and D.-J.P.-Q.; validation, I.R.-R., J.-V.R., P.H.-G. and D.-J.P.-Q.; formal analysis, I.R.-R., J.-V.R. and P.H.-G.; investigation, I.R.-R., J.-V.R., P.H.-G. and D.-J.P.-Q.; resources, J.-V.R. and I.C.; data curation, I.R.-R., J.-V.R. and D.-J.P.-Q.; writing—original draft preparation, I.R.-R., P.H.-G. and D.-J.P.-Q.; writing—review and editing, I.R.-R., P.H.-G., J.-V.R. and I.C.; visualization, I.R.-R., J.-V.R. and D.-J.P.-Q.; supervision, J.-V.R. and I.C.; project administration, J.-V.R. and I.C.; funding acquisition, I.C. All authors have read and agreed to the published version of the manuscript.

Funding: Ignacio Rodríguez-Rodríguez would like to thank the support of Programa Operativo FEDER Andalucía 2014–2020 under Project No. UMA18-FEDERJA-023 and Universidad de Málaga, Campus de Excelencia Internacional Andalucía Tech.

Acknowledgments: The authors would like to thank to Instituto Nacional de Estadística−INE (Spain) for its availability of the data and to the Instituto Universitario de Investigación Estudios de Género of Universidad de Alicante (Spain) for its support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Devries, K.M.; Mak, J.Y.; Garcia-Moreno, C.; Petzold, M.; Child, J.C.; Falder, G.; Pallitto, C. The global prevalence of intimate partner violence against women. *Science* **2013**, *340*, 1527–1528. [CrossRef] [PubMed]
2. Hyman, I.; Forte, T.; Mont, J.D.; Romans, S.; Cohen, M.M. Help-seeking rates for intimate partner violence (IPV) among Canadian immigrant women. *Health Care Women Int.* **2006**, *27*, 682–694. [CrossRef] [PubMed]
3. Haraway, D. A manifesto for cyborgs: Science, technology, and socialist feminism in the 1980s. In *Feminism/Postmodernism*; Routledge: New York, NY, USA, 1990; pp. 190–233.
4. Rodríguez-Rodríguez, I.; Rodríguez, J.V.; Elizondo-Moreno, A.; Heras-González, P.; Gentili, M. Towards a Holistic ICT Platform for Protecting Intimate Partner Violence Survivors Based on the IoT Paradigm. *Symmetry* **2020**, *12*, 37. [CrossRef]
5. Rodríguez-Rodríguez, I.; Zamora-Izquierdo, M.Á.; Rodríguez, J.V. Towards an ICT-based platform for type 1 diabetes mellitus management. *Appl. Sci.* **2018**, *8*, 511. [CrossRef]
6. Bryant, R.; Katz, R.H.; Lazowska, E.D. Big-data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society. In *Computing ResearchInitiatives for the 21st Century, Computing Research Association*; Version 8; Washington, DC, USA, 2008; Available online: http://www.cra.org/ccc/docs/init/Big_Data.pdf (accessed on 11 August 2020).
7. Islam, A.; Akter, A.; Hossain, B.A. HomeGuard: A Smart System to Deal with the Emergency Response of Domestic Violence Victims. *arXiv* **2018**, arXiv:1803.09401.
8. Hegde, N.; Bries, M.; Swibas, T.; Melanson, E.; Sazonov, E. Automatic recognition of activities of daily living utilizing insole-based and wrist-worn wearable sensors. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 979–988. [CrossRef]
9. Glaeser, E.L.; Hillis, A.; Kominers, S.D.; Luca, M. Crowdsourcing city government: Using tournaments to improve inspection accuracy. *Am. Econ. Rev.* **2016**, *106*, 114–118. [CrossRef]
10. Cranmer, S.J.; Desmarais, B.A. What Can We Learn from Predictive Modeling? *Political Anal.* **2017**, *25*, 145–166. [CrossRef]
11. Molina, M.; Garip, F. Machine learning for sociology. *Ann. Rev. Sociol.* **2019**, *45*, 27–45. [CrossRef]
12. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; Obermeyer, Z. Prediction policy problems. *Am. Econ. Rev.* **2015**, *105*, 491–495. [CrossRef]
13. Cederman, L.E.; Weidmann, N.B. Predicting armed conflict: Time to adjust our expectations? *Science* **2017**, *355*, 474–476. [CrossRef] [PubMed]
14. Beck, N.; King, G.; Zeng, L. Improving quantitative studies of international conflict: A conjecture. *Am. Political Sci. Rev.* **2000**, *94*, 21–35. [CrossRef]
15. Brandt, P.T.; Freeman, J.R.; Schrodt, P.A. Real time, time series forecasting of inter-and intra-state political conflict. *Confl. Manag. Peace Sci.* **2011**, *28*, 41–64. [CrossRef]
16. Perry, C. Machine learning and conflict prediction: A use case. *Stab. Int. J. Secur. Dev.* **2013**, *2*, 56.
17. Kleinberg, J.; Liang, A.; Mullainathan, S. The Theory is Predictive, But is it Complete? An Application to Human Perception of Randomness. In Proceedings of the 2017 ACM Conference on Economics and Computation, Cambridge, MA, USA, 26–30 June 2017; pp. 125–126.
18. Coglianese, C.; Lehr, D. Regulating by robot: Administrative decision making in the machine-learning era. *Geo LJ* **2016**, *105*, 1147.
19. Lawrenz, F.; Lembo, J.F.; Schade, T. Time series analysis of the effect of a domestic violence directive on the number of arrests per day. *J. Crim. Justice* **1988**, *16*, 493–498. [CrossRef]
20. Ozkan, T. Predicting Recidivism through Machine Learning. Doctoral Dissertation, University of Texas, Dallas, TX, USA, 2017.
21. Ward-Lasher, A.; Sheridan, D.J.; Glass, N.E.; Messing, J.T. Prediction of Interpersonal Violence: An Introduction. *Assess. Danger.* **2017**, *1*, 1–23.
22. Berk, R.A.; Sorenson, S.B.; Barnes, G. Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *J. Empir. Leg. Stud.* **2016**, *13*, 94–115. [CrossRef]
23. Holcomb, J.P.; Sharpe, N.R. Forecasting police calls during peak times for the city of Cleveland. *Case Stud. Bus. Ind. Gov. Stat.* **2006**, *1*, 47–53.
24. Sherman, L.W. Policing domestic violence 1967–2017. *Criminol. Public Policy* **2018**, *17*, 453–465. [CrossRef]

25. Cohn, E.G. The prediction of police calls for service: The influence of weather and temporal variables on rape and domestic violence. *J. Environ. Psychol.* **1993**, *13*, 71–83. [CrossRef]
26. Goodman, L.A.; Smyth, K.F.; Borges, A.M.; Singer, R. When crises collide: How intimate partner violence and poverty intersect to shape women's mental health and coping? *Trauma Violence Abus.* **2009**, *10*, 306–329. [CrossRef]
27. Hilton, N.Z.; Eke, A.W. Assessing risk of intimate partner violence. *Assess. Danger.* **2017**, *207*, 139–178.
28. Heras-González, P.; Nardi-Rodríguez, A. Respuesta institucional a la Violencia de Género en la Comunidad Valenciana (España). Institutional response to Gender-based Violence in the Valencian Community (Spain). *General. Valencia. Serv. Publ.* **2020**, *1*, 1–30.
29. Thornton, S. Police Attempts to Predict Domestic Murder and Serious Assaults: Is Early Warning Possible Yet? *Camb. J. Evid.-Based Policy* **2017**, *1*, 64–80. [CrossRef]
30. Chalkley, R.; Strang, H. Predicting domestic homicides and serious violence in Dorset: A replication of Thornton's Thames Valley analysis. *Camb. J. Evid.-Based Policy* **2017**, *1*, 81–92. [CrossRef]
31. Delgadillo-Aleman, S.; Ku-Carrillo, R.; Perez-Amezcua, B.; Chen-Charpentier, B. A mathematical model for intimate partner violence. *Math. Comput. Appl.* **2019**, *24*, 29. [CrossRef]
32. Poza, E.; Jódar, L.U.C.A.S.; Barreda, S. Mathematical Modeling of Hidden Intimate Partner Violence in Spain: A Quantitative and Qualitative Approach. In *Abstract and Applied Analysis*; Hindawi: New York, NY, USA, 2016; Volume 2016.
33. Guyon, I.; Elissee, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
34. Sheikhpour, R.; Sarram, M.A.; Gharaghani, S.; Chahooki MA, Z. A survey on semi-supervised feature selection methods. *Pattern Recognit.* **2017**, *64*, 141–158. [CrossRef]
35. Hastie, T.; Tibshirani, R.; Tibshirani, R.J. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv* **2017**, arXiv:1707.08692.
36. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
37. Karegowda, A.G.; Jayaram, M.A.; Manjunath, A.S. Feature subset selection problem using wrapper approach in supervised learning. *Int. J. Comput. Appl.* **2010**, *1*, 13–17. [CrossRef]
38. Yang, K.; Yoon, H.; Shahabi, C. A supervised feature subset selection technique for multivariate time series. In Proceedings of the Workshop on Feature Selection for Data Mining: Interfacing Machine Learning with Statistics, New Port Beach, CA, USA, 23 April 2005; pp. 92–101.
39. Crone, S.F.; Kourentzes, N. Feature selection for time series prediction—A combined filter and wrapper approach for neural networks. *Neurocomputing* **2010**, *73*, 1923–1936. [CrossRef]
40. Sánchez-Maroño, N.; Alonso-Betanzos, A.; Tombilla-Sanromán, M. Filter Methods for Feature Selection—A Comparative Study. In *International Conference on Intelligent Data Engineering and Automated Learning*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 178–187.
41. Fonti, V.; Belitser, E. Feature selection using lasso. *VU Amst. Res. Pap. Bus. Anal.* **2017**, *30*, 1–25.
42. Zhang, H.; Zhang, R.; Nie, F.; Li, X. A Generalized Uncorrelated Ridge Regression with Nonnegative Labels for Unsupervised Feature Selection. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2781–2785.
43. Zitzler, E.; Thiele, L. Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE Trans. Evol. Comput.* **1999**, *3*, 257–271. [CrossRef]
44. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **2013**, *34*, 483–519. [CrossRef]
45. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. Distributed feature selection: An application to microarray data classification. *Appl. Soft Comput.* **2015**, *30*, 136–150. [CrossRef]
46. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]
47. Brockwell, P.J.; Davis, R.A.; Calder, M.V. *Introduction to Time Series and Forecasting*; Springer: New York, NY, USA, 2002; Volume 2, pp. 3118–3121.
48. Faloutsos, C.; Gasthaus, J.; Januschowski, T.; Wang, Y. Forecasting big time series: Old and new. *Proc. Vldb Endow.* **2018**, *11*, 2102–2105. [CrossRef]
49. Kalekar, P.S. Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi Sch. Inf. Technol.* **2004**, *4329008*, 1–13.
50. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin, Germany, 2013.

51. Schölkopf, B.; Smola, A.J. A Short Introduction to Learning with Kernels. In *Advanced Lectures on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 41–64.
52. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2002; Volume 26.
53. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
54. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
55. Oshiro, T.M.; Perez, P.S.; Baranauskas, J.A. How Many Trees in A Random Forest? In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 154–168.
56. Williams, C.K.; Barber, D. Bayesian classification with gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1342–1351. [CrossRef]
57. Ortmann, L.; Shi, D.; Dassau, E.; Doyle, F.J.; Leonhardt, S.; Misgeld, B.J. Gaussian process-based model predictive control of blood glucose for patients with type 1 diabetes mellitus. In Proceedings of the 2017 11th Asian Control Conference (ASCC), Gold Coast, QLD, Australia, 17–20 December 2017.
58. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006; Volume 2, p. 4.
59. Landau, S.F.; Fridman, D. The seasonality of violent crime: The case of robbery and homicide in Israel. *J. Res. Crime Delinq.* **1993**, *30*, 163–191. [CrossRef]
60. Bowlus, A.J.; Seitz, S. Domestic violence, employment, and divorce. *Int. Econ. Rev.* **2006**, *47*, 1113–1149. [CrossRef]
61. Anderberg, D.; Rainer, H.; Wadsworth, J.; Wilson, T. Unemployment and domestic violence: Theory and evidence. *Econ. J.* **2016**, *126*, 1947–1979. [CrossRef]
62. Brahmapurkar, K.P. Gender equality in India hit by illiteracy, child marriages and violence: A hurdle for sustainable development. *Pan Afr. Med. J.* **2017**, *28*, 178. [CrossRef]
63. Hussain, S.; Dahan, N.A.; Ba-Alwib, F.M.; Ribata, N. Educational data mining and analysis of students' academic performance using WEKA. *Indones. J. Electr. Eng. Comput. Sci.* **2018**, *9*, 447–459. [CrossRef]
64. Kiranmai, S.A.; Laxmi, A.J. Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. *Prot. Control Mod. Power Syst.* **2018**, *3*, 29. [CrossRef]
65. Lang, S.; Bravo-Marquez, F.; Beckham, C.; Hall, M.; Frank, E. Wekadeeplearning4j: A deep learning package for weka based on deeplearning4j. *Knowl.-Based Syst.* **2019**, *178*, 48–50. [CrossRef]
66. Jiménez, F.; Sánchez, G.; García, J.M.; Sciavicco, G.; Miralles, L. Multi-objective evolutionary feature selection for online sales forecasting. *Neurocomputing* **2017**, *234*, 75–92. [CrossRef]
67. Novaković, J. Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav J. Oper. Res.* **2016**, *21*, 119–135. [CrossRef]
68. Nicodemus, K.K. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Brief. Bioinform.* **2011**, *12*, 369–373. [CrossRef] [PubMed]
69. Aha, D.W.; Kibler, D.; Albert, M.K. Instance-based learning algorithms. *Mach. Learn.* **1991**, *6*, 37–66. [CrossRef]
70. Kononenko, I. (1994, April). Estimating Attributes: Analysis and Extensions of RELIEF. In *European Conference on Machine Learning*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 171–182.
71. Abdi, H.; Williams, L.J. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459. [CrossRef]
72. Bergmeir, C.; Benítez, J.M. On the use of cross-validation for time series predictor evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [CrossRef]