Early Warning System Based on Data Mining to Identify Crime Patterns

Jesús Silva[1(✉)], Stefany Palacio de la Cruz[2],
Jannys Hernández Ureche[2], Diana García Tamayo[2],
Harold Neira-Molina[2], Hugo Hernandez-P[3], Jairo Martínez Ventura[3],
and Ligia Romero[2]

[1] Universidad Peruana de Ciencias Aplicadas, Lima, Perú
jesussilvaUPC@gmail.com

[2] Universidad de la Costa, St. 58 #66, Barranquilla, Atlántico, Colombia
{spalacio1, jhernand4, dgarcia34, hneira, lromero11}@cuc.edu.co

[3] Corporación Universitaria Latinoamericana, Barranquilla, Colombia
hhernandez@ul.edu.co, jairoluis2007@hotmail.com

Abstract. The analysis of criminal information is critical for the purpose of preventing the occurrence of offenses, so the crime records committed in the past are analyzed including perpetrators. The main objective was to identify crime patterns in the city of Bogota, Colombia, supported using Early Warning System based on data mining (CRISP-DM method). The research results show the identification of 12 different criminal profiles demonstrating that the Early Warning System is applicable since it managed to significantly reduce the time devoted to the processes of registering complaints and searching for criminal profiles.

Keywords: Criminal patterns · Early warning · Data mining · Data grouping

1 Introduction

Most of the studies about violence in Colombia seek to understand the causes of the criminal acts, as well as their implications, to offer recommendations of public policies. Many of these studies consider homicide rates, or some of the individual variables related to crime, as violence indicators [1]. Murder is widely recognized as the most serious crime and the most homogeneous in time, and the one that enables more reliable comparisons. But it is important to note that there are other manifestations of violence and criminality that deeply affect the population and must be included within the same index in order to have a complete view about the behavior of security problems. In this sense, the measurement of crimes allows to identify the main characteristics and trends of several crimes in order to develop policies for their prevention and repression and mitigate the effects of these problems on society [2, 3].

The criminal statistics analysis can be divided into two broad areas. The first one considers the official figures obtained from the registry the police department and justice institutions in each country [4]. This information offers advantages such as its

Y. Tan and Y. Shi (Eds.): DMBD 2019, CCIS 1071, pp. 259–268, 2019.
https://doi.org/10.1007/978-981-32-9563-6_27

national coverage, which normally follows a relatively constant accounting method and constitutes one of the main sources for the international comparative analysis. But it also presents disadvantages like the fact that people not always report crimes to authorities (mainly robberies, common injuries, and sexual offenses), and there may be errors or manipulations when entering the information to the system. In addition, regulatory changes can modify the definition of crimes, affecting the respective series.

The second area involves the victimization surveys which are intended to characterize issues related with crime, based on information collected directly from the population to give input to the authorities, improving the decision-making process. These surveys have an impact on subjects as territorial control, prevention, and follow-up of the crime, and the measurement of not denounced criminality [5].

Among the advantages of these surveys, the method applied enables to capture part of the cases that were not reported, assess the involved institutions and obtain observations that were not considered in official statistics. Additionally, these surveys allow to measure the perception of victims about crime, criminality issues, and the performance of responsible institutions. However, the victimization surveys also present disadvantages since they do not extend to the entire country and mainly focus on urban centers, they are not systematically and regularly carried out (usually take more than six months), and may have biases in so far as they are not effective for capturing, among others, crimes like murder and sexual crimes because of the survey nature [6, 7].

In this sense, the purpose of this research is the use of an early warning system for the exploitation of criminal information to obtain behavior patterns that facilitate the generation of prevention strategies. At the same time, the research seeks to study the added value of the use of data mining in the detection of criminal patterns in order to characterize them managing to extract conclusions in the prevention of crimes, that is, to apply data mining for explaining the past through historical information, understand the present, and predict future information.

2 Theoretical Review

Currently, there is no universally accepted definition of what criminal analysis is. In some police departments, it is considered as the study of police reports and the information extraction to enables the capture of criminals, in particular, serial killers. In other agencies, the criminal analysis consists of extracting statistical data from the bases of criminal acts that occur within an area and divide them into criminal families and times of the year. Whatever the definition of criminal analysis, its objective is to find relevant information within the data contained in each of the criminal acts and disseminate it among officers and investigators to assist in the capture of potential criminals, as well as stop criminal activity [8, 9].

The formal definition of criminal analysis employed in this work is the following: Criminal Analysis is a set of processes and analysis techniques aimed to provide timely and relevant information concerning the facts and correlations of criminal tendency to operational and administrative staff during the planning of actions to prevent and avoid criminal activities, and to clear up the cases [1].

2.1 Legislation and Policies Against Crime in Colombia

In the context of an internal armed conflict that continues bleeding the country with a lengthy trace of deaths, abductions, disappearances, displacement, etc., with the complex social reality that generates a large mass of excluded population and high rates of criminality caused by the drug-trafficking mafias and common delinquents, and with the low results presented by the criminal justice system, the State has chosen to apply legal reforms for a solution. In the last twenty years, Colombia has redacted four criminal procedure codes (Decree 050/1987, Decree 2700/1991, Law 600/2000, and law 906/2004), two penal codes (decree 100/1980 and law 599/2000), two codes of minors (Decree 2737/1989 and Law 1098/2006), without including the countless partial reforms. In general, just from the year 2000 to 2006 more than 50 criminal laws were issued, including the international conventions and protocols related to the subject [1, 10, 11].

2.2 CRISP-DM Method

The CRISP-DM Method (Cross-Industry Standard Process for Data Mining) was used for developing the research. It is a free distribution method used for data mining projects, developed in 1999 by the consortium of European companies named Pete Chapman and Randy Curber (NCR - Denmark), AG (Germany), Julian Clinton, Thomas Khabaza and Colin Shearer (SPSS - England), OHRA (Holland), and Thomas Reinartz and Rüdiger Wirth (DaimlerChrysler). This method consists of six phases [12, 13]:

PHASE 01: UNDERSTANDING THE BUSINESS Objectives and requirements from a non-technical view.

- Setting up the business objectives (objectives, needs, and success criteria)
- Assessment of the situation (requirements, assumptions, restrictions)
- Generation of the project plan (plan, tools, equipment, and techniques)

PHASE 02: UNDERSTANDING THE DATA Getting familiar with the data bearing in mind the business objectives

- Initial data collection
- Description of the data
- Exploring the data
- Verification of data quality

PHASE 03: ORGANIZING THE DATA: Obtaining the minable view or dataset

- Selection of data
- Data Cleaning
- Construction of Data
- Data Integration
- Data formatting

PHASE 04: MODELING Apply the data mining techniques to the dataset

- Selection of the modeling technique
- The evaluation design
- Construction of the model
- Evaluation of the model

PHASE 05: EVALUATION of the previous phase to determine if the dataset is useful for business needs

- Evaluation of Results
- Review of the Process
- Establishment of the next steps or actions

PHASE 06: IMPLEMENTATION Exploit usefulness of models, integrating them into the decision-making tasks in the organization.

- Implementation Plan
- Monitoring and Maintenance Planning
- Generation of final report
- Revision of the draft.

3 Materials and Methods

3.1 Database

From a total of 5,684 complaints a month, through sampling calculations, the result shows a total of 359 complaints a month in the year 2018, which will be used [14].

3.2 Methods

Regarding the development of the system, it was decided to opt for the OPEN UP Method because they are especially oriented to small projects, which is a tailor-made solution for that environment. For the development of the research, the CRISP-DM method was applied as a free distribution tool for data mining projects [12].

3.3 Indicators

According to a previous work [15], the indicators used in the early warning system are shown in Table 1.

Table 1. Early warning indicators of the system used for the study.

Variable	Dimension	Indicator	Description
Early Warning System based on data mining	Access to the Early Warning System	security Index of access to the system	Indicates the percentage of people in the Lambayeque Police Region who have access to the system according to the levels of security
	Characteristics of the Early Warning System	Index of satisfaction for the use of the Early Warning System	Indicates the percentage of people in the Lambayeque Police Region who feel satisfied with the use of the Early Warning System
		Usability of the Early Warning System	Indicates whether the Early Warning System is easy to be learnt and used by people
		Query Response Time	Delay time of the system to give a response to a user request
		Number of reports by persons denounced	Indicates the number of reports issued by the system about persons who have been denounced
		Time of dedication in the registration of the complaint	Time spent on the process for registering a complaint
		Time in information searches by criteria	Indicates the time it takes the system to determine a person with a criminal profile
Process of detection of criminal patterns	Denounce	Delay time in the issue of resolution of denounces	Delay time in response to a denounce
		Index of errors in data matches of the one involved by case file	Indicates the percentage of denounces registered with errors (duplication and inconsistency of data)
	Desktop Materials	Cost per desktop materials	Indicates the amount of material used in the process for registering denounces

4 Results

The scope of the Early Warning System includes the registration process and storage of complaints, determination of groups with similar characteristics that were involved in a complaint and determining criminal profiles. It also allows the maintenance of intervention unit, crime, type of crime, the aggrieved, involved, and attorney. It also presents the maintenance of user (police), consultations, the generation of reports, and finally allows to assess the possible suspects in a committed crime.

4.1 Selection of Data

The fields that were considered for the study were chosen considering the following factors:

- Variables that each field or attribute can take.
- Data quality.
- Importance of the attribute for the study according to the objectives of the project, delivery information and its significance.

The fields that are considered for the present study are the following [14]:

- Age: Numeric field that indicates the age of the suspect. This attribute takes values from 0 years up to the age of 100 years.
- Sex: Numeric field that indicates the sex of the detainee.
- Marital status: Describes the marital status of the involved at the time of committing a crime.
- Psychic condition: Psychic condition of the involved at the time of committing a crime.
- Level of instruction: A variable that describes the level of education of the involved.
- Recidivism: Represents the type of recurrence of the suspect in relation to a crime committed.
- Incidence Time Zone: This variable indicates the time of the offense, which was specified by ranges of 6 h each group.
- Type of weapon: Referred to the type of weapon used by the involved (delinquent) at the time of committing a crime.
- Id of the crime: Identification variable of the crime that was committed, which has already been established by the NPP.
- Id of the unit of intervention: Indicates the commissioner that registered the complaint.
- Aggression: Type of aggression caused to the victim at the time of a crime.
- Crime circumstances: Referred to the additional crimes committed according to the circumstance of the development of the main fact.
- Stolen amount: In the case of offenses against property or other type of crime incurred in the robbery of goods, a range of the value stolen was specified.

4.2 Selection of the Modeling Technique

At this stage, the techniques and algorithms used in the development of data mining were selected. Initially, non-supervised techniques were used, and it was necessary to make a data grouping or clustering to ease the discovery of hidden information. The algorithm used was K-Means, one of the most commonly used methods and the most popular of the “partition” grouping methods which is also the most widely used algorithm in the studies consulted for the research [13].

The next step was the validation of groups or clusters with specialists in the criminal field. Such validation enabled the classification technique. Finally, neural networks were applied as part of the supervised classification technique considering the previously defined cluster to achieve the identification of decision rules that help explain the composition of each group.

4.3 Clustering

The K-means algorithm was used in order to complete the data grouping, resulting in 12 criminal profile groups through their most significant features.

Firstly, before having analyzed the cluster, it was found that most of the clusters show a greater proportion of men who have committed a crime (sex = 2), which was the most frequent, according to records obtained from the regional police and with a smaller percentage for women, see Table 2.

Table 2. Distribution of recidivism of the involved.

Recidivism	
Description	%
Male	75.8
Feminine	23.8
No record	3.4
Total	100%

Therefore, males are expected to obtain more frequency in all the runs of clustering within formed groups since it was the most predominant sex in the records of complaints.

Figure 1 shows the scatter chart for the cluster analysis.

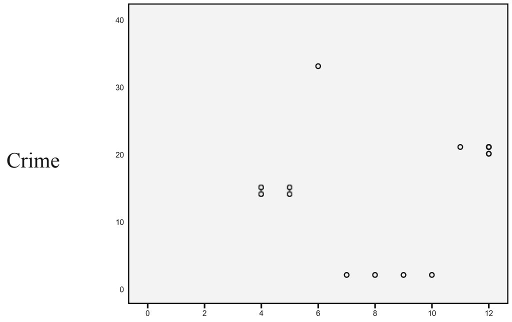

Fig. 1. Distribution of the cluster according to Crime.

The x-axis was represented by the 12 clusters, whereas the ordered ones were represented by the offenses. Since the 2 represented the offense of injury, it was observed that the clusters 1, 7, 8, 9 and 10 were the ones which formed this group of offenses, although cluster 9 obtained a greater proportion (55.6%) of the total number of crimes by injury. The x-axis was represented by the 12 clusters, whereas the ordered ones were represented by the offenses. Since the 2 represented the offense of injury, it was observed that the clusters 1, 7, 8, 9 and 10 were the ones which formed this group of offenses, although cluster 9 obtained a greater proportion (55.6%) of the total number of crimes by injury.

On the other hand, the 14 represented the offense of simple and aggravated robbery, and was contained in clusters 3, 4 and 5, cluster 4 obtained the highest proportion (43.8%) and cluster 3 obtained the lowest proportion with 3% of the total number of crimes by simple and aggravated theft.

The 15 represented the offense of simple and aggravated robbery, and was contained in clusters 3, 4 and 5, with the highest proportion in cluster 4 (60.2%) and the lowest proportion in cluster 3 with 6.1% of the total number of crimes by simple and aggravated robbery.

The 20 represented the offense of simple and aggravated damage. It was contained by cluster 12, which contained such offenses (100).

The 21 represented the crime of Usurpation/Extortion and was contained by clusters 11 and 12 with the highest proportion in cluster 12 (76.2%) and the lowest proportion in cluster 11 with 23.8% of the total number of crimes.

The 33 represented the crime of domestic violence, and was contained by the clusters 2 and 6, with the highest proportion in cluster 6 (94.3%) and the lowest proportion in cluster 2 with 5.7% of the total number of crimes by family violence.

4.4 Interpretation of the Clusters

Based on the information obtained from both the analysis of graphs prior research and collaboration of specialists, a first interpretation of each formed cluster was obtained. It is convenient to highlight, as discussed with the specialists, that for performing a deeper and accurate analysis, it would be necessary to have access to each suspect individual history like knowing the family and cultural origin of that person. However, a very good approximation was obtained by the present research. For reasons of space and privacy, just analysis of cluster 1 is shown:

- Cluster size = 4 (0.8% of the total sample).
- Average age = 21–30 (50% of the cluster) 31–40 years (25% of the cluster) and 41–100 years (25% of the cluster).
- Sex = Male (100% of the cluster).
- Marital Status = Single (25% of the cluster), married (25% of the cluster), Divorced (25% of the cluster) and Cohabitant (25% of the cluster).
- Level of education = Secondary incomplete (5% of the cluster) and technical (75% of the cluster).
- Recidivism = Recidivist (75% of the cluster) and multi-recidivist (25% of the cluster).

- Psychic Condition = Sick/altered (50% of the cluster), drugged (25% of the cluster) and Fair (25% of the cluster). ϖ Incidence Time Zone = 8:01 a.m.–2:00 p.m. (25% of the cluster) 2:01–8:00 p.m. (50% of the cluster) and 2:01 a.m.–7:59 a.m. (25% of the cluster).
- Intervention Unit = El Porvenir Police Station (75% of the cluster) and La Victoria Police Station (25% of the cluster).
- Type of Weapon = None (50% of the cluster) and knife (50% of the cluster).
- Aggression = Both types of aggression, namely physical and psychological integrity (100% of the cluster). Crime = Crimes against Life, Body and Health (100% of injuries).
- Crime circumstances = Crime Against Freedom (25% of the cluster of Violation to Personal Freedom offense, and 50% of the 95 cluster of Home Violation offense) and simple and aggravated robbery (25% of the cluster).
- Stolen amount = There was no robbery (75% of the cluster), robbery valued at more than 100 and less than 500 soles (25% of the cluster).

4.5 Classification

After obtaining the groups, the supervised classification was performed using Neural Networks with MATLAB R2010a [16–18]. The layers of hidden inputs and outputs were defined using 13 entries corresponding to the variables defined in the clustering stage. To know the optimum number of hidden layers [9], the number of inputs were contrasted with the outputs, resulting in a hidden network in order to optimize the work and preventing the overfitting within the training. The layers of output are 12 referring to groups that are already established. 150 iterations were used, avoiding to give a small number as the network cannot achieve the purpose of training. In addition to the process of neural network, three phases were considered: Training phase in which the 359 valid complaints were used to determine the parameters that define the neural network model. Later, the validation phase was applied to avoid overfitting, so this stage allowed to control the learning process.

5 Conclusions

The identification of criminal patterns was achieved with the support of the Early Warning System based on Data Mining developed for the Police Region of Bogota, whereby 12 criminal groups were defined, with different characteristics and behaviors, allowing the validation of pre-existing knowledge. In addition, it was possible to characterize those involved in a crime based on their most relevant attributes.

The use of the Early Warning System demonstrated its efficiency as it was contrasted with the information obtained prior the use in the Police Region, resulting in a considerable reduction of time were the system spent 4.9 min in the process of registering complaints and more of 1 h and a half for the processes of criminal profile searches. The system allowed maintaining the information ordered and updated in such a way that information can be accessed quickly, managing to reduce an average of 11 min in manual searches.

References

1. Green, W.J.: A History of Political Murder in Latin America: Killing the Messengers of Change. State University of New York Press, Albany (2015)
2. Jaramillo, J., Meisel, A., Ramírez, M.T.: More than 100 years of improvements in living standards: the case of Colombia. Cliometrica, October 2018. Online First
3. Karl, R.A.: Forgotten Peace: Reform, Violence, and the Making of Contemporary Colombia. University of California Press, Oakland (2017)
4. Roskin, M.G.: Crime and politics in Colombia: considerations for US Involvement. Parameters: US Army War Coll. Q. **34**(1), 126–134 (2001)
5. Santos Calderón, E.: El país que me tocó (Memorias). Penguin Random House Grupo Editorial (2018)
6. Calderón, M., Marconi, S.: Santuarios de la memoria: historias para la no repetición. Relatos de actos humanitarios en la vereda Beltrán y el municipio de Marsella, Risaralda (Tesis de Pregrado). Universidad Santo Tomás, Bogotá (2017)
7. Martínez, L.: Violencia y desplazamiento: Hacía una interpretación de carácter regional y local. El caso de Risaralda y su capital Pereira, en Revista Estudios Fronterizos, vol. 7, no. 14, julio–diciembre 2006
8. Martínez, S.: Núcleos urbanos y de frontera en el Centro Occidente Colombiano. Un proyecto de institucionalización del Estado Nación en el siglo XIX, en Americanía. Revista de Estudios Latinoamericanos. Nueva Época, no. 3, enero–junio 2016
9. Núñez, M.: Contexto de violencia y conflicto armado, Monografía Político Electoral del Departamento de Risaralda, 1997–2007. Misión de Observación Electoral, Corporación Nuevo Arcoiris, CERAC, Universidad de los Andes, Bogotá (2017)
10. Ovalle, L.: Memoria y codificación del dolor. Muertes violentas y desapariciones forzosas asociadas al narcotráfico en Baja California, en Revista de Estudios Fronterizos (2010)
11. Palacios, M.: Violencia púbica en Colombia 1958–2010. Fondo de Cultura Económica, Bogotá (2012)
12. Huber, S., Seiger, R., Kuhnert, A., Theodorou, V., Schlegel, T.: Goal-based semantic queries for dynamic processes in the Internet of Things. Int. J. Semant. Comput. **10**(2), 269 (2016)
13. Wirth, R., Hipp, J.: CRISP-DM: towards a standard process model for data mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, pp. 29–39 (2000)
14. Merchan-Rojas, L.: Conducta Criminal: una Perspectiva Psicológica Basada en la Evidencia. Acta Colombiana De PsicologíA **22**(1), 296–299 (2019)
15. Azevedo, A.I.R.L., Santos, M.F.: KDD, SEMMA CRISP-DM: a parallel overview. IADS-DM (2008)
16. Vásquez, C., et al.: Cluster of the Latin American universities top100 according to webometrics 2017. In: Tan, Y., Shi, Y., Tang, Q. (eds.) DMBD 2018. LNCS, vol. 10943, pp. 276–283. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93803-5_26
17. Viloria, A., Mercedes, G.-A.: Statistical adjustment module advanced optimizer planner and sap generated the case of a food production company. Indian J. Sci. Technol. **9**(47), 1–5 (2016)
18. Varela, I.N., Cabrera, H.R., Lopez, C.G., Viloria, A., Gaitán, A.M., Henry, M.A.: Methodology for the reduction and integration of data in the performance measurement of industries cement plants. In: Tan, Y., Shi, Y., Tang, Q. (eds.) DMBD 2018. LNCS, vol. 10943, pp. 33–42. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93803-5_4