Date of publication SEP-30, 2024, date of current version AUG-23, 2024.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.23.3.XXXX

Using Big Data Analytics to Identify Trends and Group Crimes through Clustering

JORGE MARIN, GUSTAVO GUERREROS, DAVID CALDERON-VILCA
Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional Mayor de San Marcos, Lima, 15081, Perú

Corresponding authors: Jorge Marin (e-mail: jorge.marin4@unmsm.edu.pe), Gustavo Guerreros (e-mail: gustavo.guerreros@unmsm.edu.pe), David Calderon (e-mail: hcalderonv@unmsm.edu.pe).

ABSTRACT The incidence of crime in a city presents a challenge in the absence of trend analysis that impacts citizen security. The objective of this research was to analyze and visualize crime trends in the area, using the concepts and fundamentals of Big Data Analytics, Data Mining and Clustering, the problem is addressed with a quantitative approach, using the CRISP-DM process, Principal Component Analysis (PCA) and the K-Means algorithm for clustering. Validation is performed with the Elbow Score and the Average Silhouette method, ensuring the robustness of the data clustering. The results show that crimes against property, such as robbery and theft, are the most frequent. Four crime clusters are identified, each associated with a specific category, providing a detailed view of crime distribution. Comparison with previous studies highlights the effectiveness of Big Data technologies in reducing crime, providing a solid basis for more accurate security strategies.

KEYWORDS Big Data; crime; crime trends; clustering; crimes; data mining; security.

I. INTRODUCTION

CRIME has been a major concern in society, and in recent years it has been aggravated by a number of factors, including the pandemic and poor government management. In Latin America, especially in Peru, there has been a notable increase in crime rates and an increasingly palpable sense of insecurity in recent decades. In order to address this problem, an exhaustive search was conducted in the Web of Science and Scopus databases, covering articles and quantitative studies related to the topic of delinquency and Big Data Analytics.

Juvenile delinquency has contributed to the increase in crime rates worldwide [1]. It is crucial to address this problem and seek solutions to reduce its incidence in society. On the other hand, [2] points out that criminality has experienced an increase due to various social and family factors, representing a threat to global security and development. Furthermore, data compiled by [3] reveal a wide variation in the prevalence of crime in different regions of the world, with the highest levels of common crime and homicides observed in the Global South, while organized crime is concentrated in Africa, Latin America and Asia.

The American continent ranks third in crime levels worldwide, with high rates in the Central and South American sub-regions, due to the presence of the criminal market and lack of resilience, generating a negative impact on several aspects such as crime, the economy, forced displacement, human trafficking and smuggling [2]. Victims of crime in Latin America show lower trust in local police, reflecting the challenges in terms of trust in public institutions and the capacity of governments to address the consequences of crime. Therefore, policies that reduce victimization risks and restore trust in public institutions are needed to improve social welfare in the region [4].

In recent years, crime in Peru has increased significantly, generating a generalized sense of insecurity in the country [5]. This situation is reflected in crime statistics, which show a variety of crimes reported to police authorities. Crime in Peru is influenced by multiple factors, such as poverty, inequality, lack of access to education and employment, as well as the political and economic problems the country has faced [5]. The main groups of factors related to crime in Peru include reported crimes, reported misdemeanors, dismantled gangs, and family and sexual violence, each of which is influenced by specific environmental characteristics of each district [5]. During the period from October to December 2022, the most frequent complaints in the country were related to crimes against property, followed by complaints against public safety, life, body and health, and freedom [5]. Crimes against property, such as robbery, theft, fraud and fraud, are of particular concern due to their impact on society and the frequency with which

they are reported, which affects people's security and peace of mind [6].

Citizen insecurity in Lima-Peru has experienced a worrying increase due to a series of risk factors, such as economic and social inequality, poverty, urban design, alcohol and drug consumption, among others [6]. These same factors are present in Metropolitan Lima, where thousands of crime reports were registered during the last quarter of 2022, reflecting a worrisome situation in terms of security [6]. A 27.4% increase in the number of crime reports was reported in Metropolitan Lima in 2022 compared to the previous year. In addition, an increasing trend in the number of complaints was observed over the last two years, with districts such as Lima, San Juan de Lurigancho, San Martin de Porres and Comas leading the list [5]. These figures highlight the need to implement effective policies to combat crime and improve citizen security in Metropolitan Lima and its districts.

The following research highlights the effectiveness of advanced data analysis techniques in crime prediction and monitoring, providing solid evidence of their positive impact on crime fighting. For example [7] point out that crime analysis requires more advanced approaches, such as data mining, due to the complexity of the data and their intangible relationship. Furthermore, in [8] they highlight that data mining allows discovering new insights and gaining a deeper understanding of criminal phenomena. By using Big Data Analytics and data mining techniques, it is possible to identify crime patterns more accurately and efficiently, which facilitates crime prevention and management. Likewise, in [7] they indicate that traditional methods of crime analysis are insufficient due to the exponential growth of data and the complexity of their relationships, which requires further research in data mining applied to crime analysis. Finally, in [9] they highlight that data mining and predictive analytics are fundamental in the fight against crime, as they allow the classification of crimes and the prediction of future criminal incidents.

We face the challenge of addressing the lack of adequate tools and methodologies to analyze large amounts of crime data in the Lima region. The lack of these tools makes it difficult to make timely and effective decisions in the fight against crime.

The objective of the research is to analyze and visualize crime trends in the area, using Big Data Analytics, Data Mining and Clustering techniques. The research is justified because it shows the impact of crime for the authorities that would allow them to prevent and fight more efficiently. We clearly recognize that the results of this study have the potential to generate a significant impact on the fight against crime, providing valuable information that can guide more effective strategies.

II. STATE OF THE ART

A. DETAILED EXPLORATION OF CRIME PATTERNS IN A SPATIAL AND TEMPORAL CONTEXT

The spatial and temporal analysis of crime patterns, in the reviewed studies, addressed various perspectives and techniques to understand these patterns in specific regions. In [10] highlights the importance of kernel density analysis, spatial clustering and spatial modeling, identifying areas with high criminal activity and evaluating the relationships between variables. This analysis, based on primary and secondary data, revealed deficiencies in police coverage in high crime areas. In contrast to [9] focused on data collection from news sources, using classification and machine learning techniques to detect spatio-temporal patterns, they highlight a 75% accuracy in crime prediction and the identification of previously unrecognized factors influencing crime.

In the study [11] they conducted a comparison of methods for the detection of uncertain spatio-temporal crime patterns. Their study highlights that the probabilistic distance based method outperformed others, while approaches such as possible world and expected distance based approaches presented limitations and inconsistent results. In contrast with [12] they used statistical techniques, principal component analysis and hierarchical clustering in cell phone theft crime reports in Bogota. Their findings revealed demographic characteristics and temporal patterns, such as a higher frequency in the 25-30 age group and higher incidence in men, indicating the need for specific preventive strategies in areas identified with high crime incidence.

B. CRIME PREDICTION: ADVANCES IN PREDICTIVE METHODS

Recent research [13] highlighted the effectiveness of combining deep learning and exponential smoothing techniques to improve crime prediction in New York, highlighting the superiority of this combination over other methodologies. This approach uses advanced analytical techniques, such as clustering, to identify patterns in crime data [14], emphasizing the prevalence of supervised learning in crime prediction, noting remarkable performances of artificial intelligence techniques; this observation highlights the use of the K-Means algorithm.

On the other hand, work [15] on the generation of geographical profiles and prediction of crime locations using probability statistics methods highlights the importance of considering the mobility of the criminal, an aspect that could influence the spatial distribution of crimes. This comparison allows reflecting on the applicability of the approach. In research [16], they used machine learning techniques to predict crimes in Porto, linking demographic variables with crime rates, and analyzing sentiment on Twitter to assess the perception of safety. In contrast, the study [17] on crime prediction and classification using decision trees and Bayesian classification provides an alternative approach that can enrich our understanding of predictive techniques.

C. DATA MINING FOR THE PREVENTION AND UNDERSTANDING OF CRIME

In the field of data mining for crime prevention and understanding, several studies have applied varied approaches and techniques, [8] highlighted the effectiveness of advanced models such as Prophet and LSTM to identify trends and seasonality in crime data. On the other hand, in [18] they highlighted the importance of data mining techniques such as entity extraction and association rule mining to detect hidden patterns in structured and unstructured crime-related data, offering valuable insights into the use of clustering.

In [7] they proposed a systematic approach using SOM and MLP neural networks for clustering and classification of crime data, stressing the need to consider spatiotemporal data and behavioral variables of crime for a more complete understanding of its dynamics. In [19], they combined fuzzy logic and conventional social theories to map and analyze criminal cases, providing a unique perspective that could enrich data analysis.

III. METHODOLOGY

In our research, we used a quantitative, descriptive and predictive type of study, with the objective of describing crime patterns and behaviors in the Lima-Peru region. We applied the CRISP-DM methodology, a data mining process model that guided us through the stages of data exploration, variable selection, modeling and evaluation. We describe the data collection, processing and analysis of large amounts of crime data, through data analysis and the implementation of a neural network model to predict future trends. We used Big Data Analytics, Data Mining and Machine Learning techniques.

A. DATA COLLECTION

Our main source of crime information comes from the Peruvian government's open data site [20], as reported by the Ministerio Público Fiscalía de la Nación in 2022. This site provides access to public information on various topics, including citizen security and reported crimes. For our analysis, we collected and integrated data from Lima-Peru from 2016 to 2022, which allowed us to examine patterns and trends over time. Data collection was conducted through secondary sources, mainly from the National Police of Peru, as also indicated by the Ministerio Público Fiscalía de la Nación in 2022. The National Police provided us with detailed information on crimes committed over five years. These data were obtained from multiple crime reports made by citizens at police stations nationwide and included details on the type of crime, date and place where it occurred, among others.

The data collection flowchart shown in Figure 1 represents the process used by the Ministerio Público Fiscalía de la Nación to collect information from the reports.

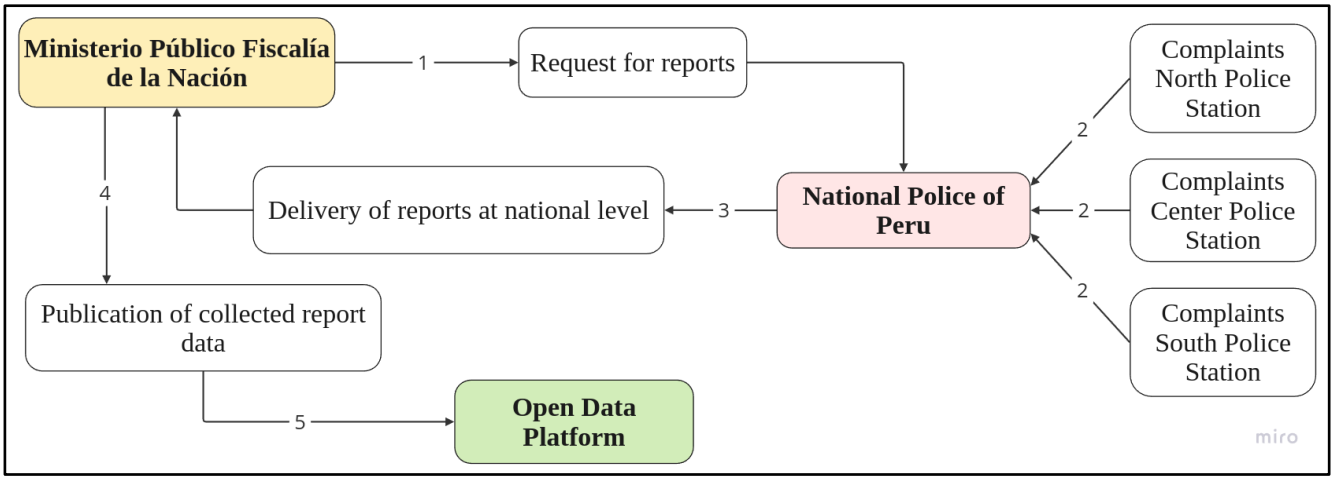

Figure 1. Data collection flow. Ministerio Público Fiscalía de la Nación

To ensure the accuracy and reliability of the data we collected in the past, we implemented additional measures. We cross-checked the data using additional sources, such as the Instituto Nacional de Estadística e Informática (INEI). In addition, we made sure to use only official and public data sources for our study, which provided us with accurate and reliable data.

The data library compiled from the Open Data of the Peruvian State was a valuable source of information in our study, providing a wide range of relevant variables on the recorded cases. This allowed us to better understand the reported events and their context. We used a detailed table to describe the variables present in the data, which facilitated the analysis and understanding of the characteristics of the data collected.

Table 1. Description of variables

ID	Variable	Data type	Description
1	*periodo_denuncia*	Text	Reporting period
2	*anio_denuncia*	Numerical	Year of the report
3	*fecha_descarga*	Text	Date of data download
4	*distrito_fiscal*	Text	Tax district of the report
5	*especialidad*	Text	Case specialty
6	*tipo_caso*	Text	Type of case
7	*generico*	Text	Generic category of the case
8	*subgenerico*	Text	Sub-generic case category
9	*articulo*	Text	Article related to the report
10	*des_articulo*	Text	Description of the article
11	*cantidad*	Numerical	Amount related to the report
12	*ubigeo_pjfs*	Numerical	Supraprovincial Prosecutor's Office Location
13	*dpto_pjfs*	Text	Department of Supraprovincial Prosecutor's Office
14	*prov_pjfs*	Text	Province of the Supra-provincial Public Prosecutor's Office
15	*dist_pjfs*	Text	District of the Supraprovincial Public Prosecutor's Office
16	*fecha_corte*	Text	Cut-off date of the report
17	*fecha_descarga*	Text	Date of data download

B. DATA PREPARATION

In our study, data preparation played a crucial role, as mentioned by [21]. According to the author, in predictive modeling projects, machine learning algorithms learn by mapping input variables to a target variable. However, these algorithms cannot work directly with raw data. Therefore, it was necessary to transform the data to meet the specific requirements of each algorithm.

Figure 2 provides a detailed overview of the crucial steps followed during the data preparation process prior to data analysis.

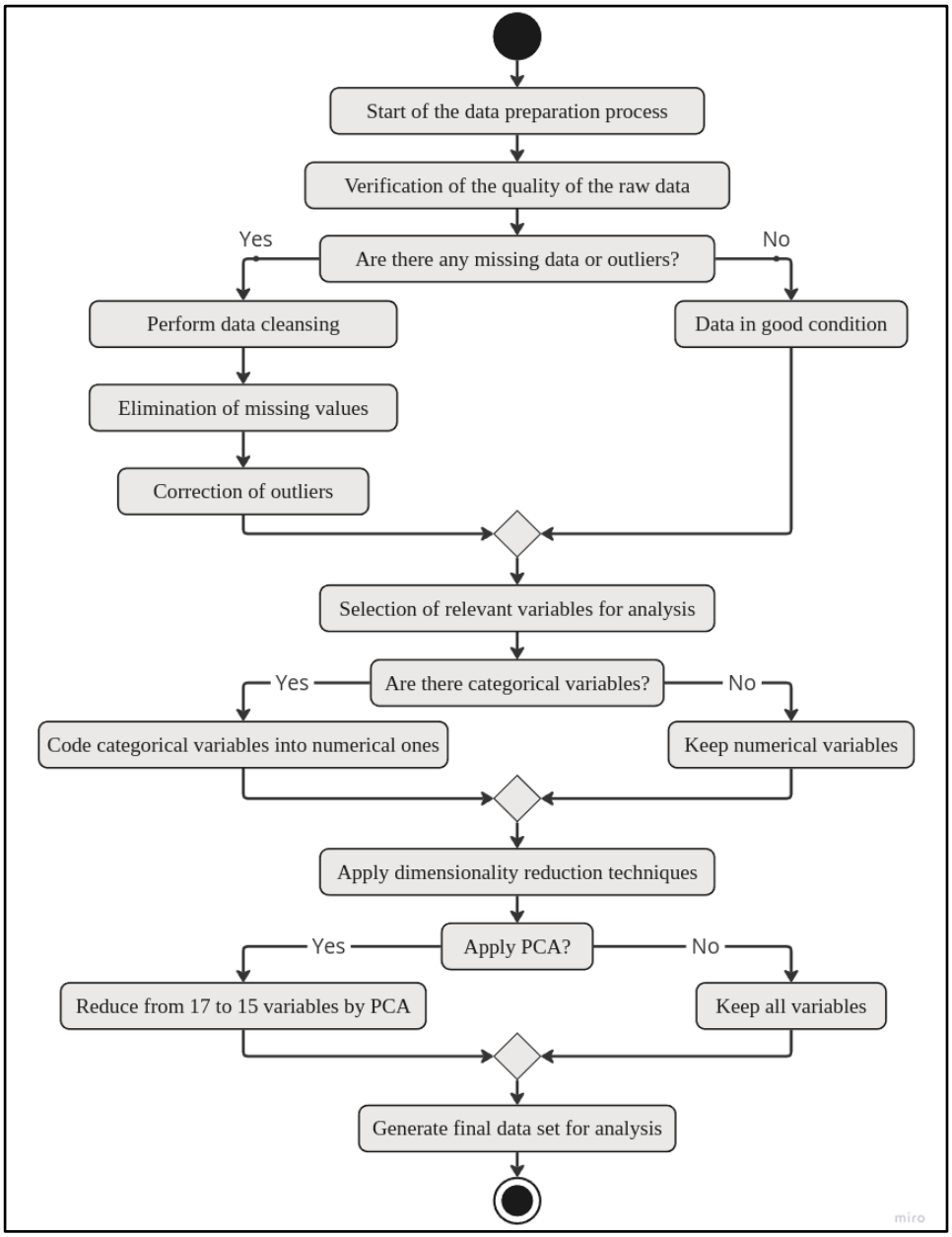

Figure 2. Data preparation

An exhaustive review of the columns that made up the data set in question was carried out, and those columns that lacked relevance in the context of the analysis undertaken were identified and subsequently eliminated; this action was carried out with the aim of purifying the data set and retaining only those variables that had a relevant informative value for the study.

Subsequently, we proceeded to evaluate the presence of missing values within the data set. In this process, we identified the cells that lacked information and designed a management strategy according to the amount of missing data and the relevance of these values in the analysis. In instances where the amount of missing data was minimal compared to the total volume of data, the corresponding rows were excluded. However, in cases where the absence of data was substantial, imputation techniques based on statistics or models were implemented to accurately and reliably fill in the missing values.

Once the missing values treatment stage was completed, a rigorous detection and correction of outliers within the data set was carried out. This process was based on the calculation of descriptive statistics, such as mean, median and standard deviation. In addition, visualization techniques, including histograms, box plots and scatter plots, were used to identify

possible outliers. In cases where outliers were identified, a thorough evaluation of their relevance in the context of the study was carried out, determining whether their exclusion or treatment was appropriate, always taking into account the specific context of the research.

Finally, an exhaustive verification of the consistency and validity of the data was carried out. This phase involved a meticulous examination of the columns to ensure that they complied with the predefined constraints and conditions. For example, the correct formation of dates, the validity of numerical values within established ranges and the appropriateness of the categories present in the columns were verified. In cases where inconsistent or invalid data were identified, corrections or deletions were applied as necessary.

C. DATA CLEANING

In our research work, we recognize the importance of data cleaning as a fundamental step in the process of preparing data for further analysis [21]. For this study, we focused on a data set consisting of 9363 rows and 17 columns.

Figure 3 provides a detailed overview of the fundamental process of data cleaning, a crucial step in data preprocessing. This essential step is responsible for cleaning and structuring the data, ensuring the quality and reliability of the information that will later be subjected to analysis. Through a graphical representation, the different phases and techniques employed in this process are illustrated, providing a clear visual guide on how to approach the rigorous preparation of data prior to subsequent analysis and modeling.

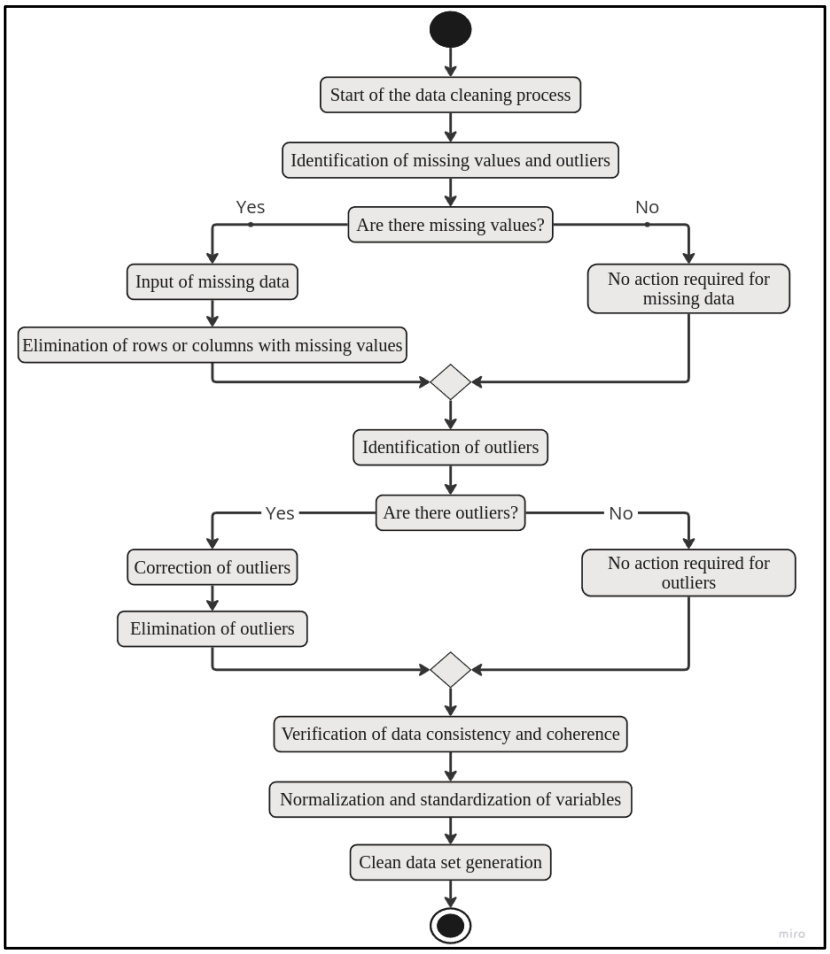

Figure 3. Data cleaning

During the cleaning and preprocessing of the database, we performed several operations to ensure the quality and consistency of the data, as a result, we obtained a final table organized into two categories: categorical variables and numerical variables.

In the case of the categorical variables, these contained descriptive information related to the reported cases, for these variables, we performed the following actions: we verified the existence of duplicate records and eliminated them to ensure the uniqueness of the data. In addition, we reviewed the categorical variables for spelling errors, inconsistencies or discrepancies, correcting them and normalizing the values to achieve consistency and uniformity in the data. We also evaluated the relevance of each categorical variable in terms of our research objectives, eliminating those that did not provide meaningful information. For example, we recorded the tax district where the report was made, the specialty associated with the case, the type and generic category and specific subcategory of the case, as well as the description of the article related to the reported case. In addition, we collected geographic information, such as the department, province and geographic district related to the reported case. These categorical variables allowed us to classify and organize the reported cases according to different relevant characteristics.

The numerical variables reflected quantitative aspects of the reported cases. We performed the following actions: we identified records with missing values and addressed them by eliminating the records or imputing values according to the

context and percentage of missing values; we examined the numerical values for outliers, eliminating or correcting them to avoid bias in the results; and we applied normalization techniques to ensure a comparable scale and facilitate interpretation. The numerical variables recorded included the number of reported cases, which measures the incidence of each type of case, and the article number related to the case, facilitating its reference and classification according to the corresponding legislation.

D. FEATURE SELECTION

In our methodology, we rely on the recommendations of [21] on the importance of feature selection in the development of predictive models. This process is essential and aims to reduce the number of input variables, thus improving model performance and computational efficiency. Statistical based methods are used to evaluate the relationship between the input variables and the target variable, selecting those that are most relevant.

We used two approaches in feature selection: a supervised one based on the target variable, and an unsupervised one independent of the target variable. Both approaches allowed us to identify the most significant variables and to discard those that had a limited impact on our predictions.

During the feature selection process, we employ statistical measures of correlation between input and output variables. The input variables are those used as input to our model, while the output variables are those that we attempt to predict. We took into account the type of output variable, whether it is a regression or classification problem, when applying the appropriate techniques.

Steps for variable selection with Principal Component Analysis (PCA) technique

Step 1: Data preparation

To calculate the covariance matrix between the specific numerical variables, each variable has been denoted as follows:

Table 2. Numeric variables

ID	Variable	Type	Denotation
2	*anio_denuncia*	Numerical	PART N° (X_1)
11	*cantidad*	Numerical	SQUARE (X_2)
12	*ubigeo_pjfs*	Numerical	SQUARE (X_3)

Step 2: Data normalization (Standardization)

In this step, the numerical variables were normalized before applying PCA, in order to ensure that all variables have the same scale. The standardization technique was used, which consists of subtracting the mean and dividing by the standard deviation of each variable.

The study worked with the following numerical variables: $anio\_denuncia$, which represents the year in which the report was made; $cantidad$, which indicates the frequency of reports; and $ubigeo\_pjfs$, the location code of the Supraprovincial Prosecutor's Office. These numerical values, having different ranges and magnitudes, needed to be standardized to facilitate the analysis and prevent the variables with higher ranges from disproportionately influencing the PCA.

Standardization was applied as follows: the numerical variables to be standardized were selected: $anio\_denuncia$, $cantidad$, and $ubigeo\_pjfs$. Then, for each variable, the mean and standard deviation were calculated. For example, for the variable quantity, the mean (μ) and standard deviation (σ) were determined. The z-score standardization formula was used to transform the values of each variable.

$$z = (x - \mu) / \sigma,$$

Where:

- z : is the standardized value.
- x : is the original value of the variable.
- μ : is the mean of the variable.
- σ : is the standard deviation of the variable.

This formula was applied to each observation in the variables $anio\_denuncia$, $cantidad$, and $ubigeo\_pjfs$, transforming these data to a common scale with a mean of 0 and a standard deviation of 1.

After standardization, the data were transformed to a uniform scale. This allowed PCA to identify the principal components that captured the most variance in the data without any dominant variable biasing the analysis. By applying PCA on the standardized data, the principal components reflected a balanced combination of the numerical variables, facilitating effective dimensionality reduction and better interpretation of patterns in the crime data.

Step 3: Calculation of the covariance matrix

In this step, the covariance matrix was calculated to capture the linear relationships between the variables. To calculate the covariance matrix, the standardized data of $anio\_denuncia$, $cantidad$, and $ubigeo\_pjfs$ were used.

The formula used to calculate the covariance between two variables X and Y was as follows:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)}{n - 1},$$

Where:

- X_i and Y_i : are the individual values of variables X and Y respectively.
- μ_X and μ_Y : are the means of variables X and Y respectively.
- n : is the total number of observations.

Using this formula, the covariances between all combinations of pairs of variables were calculated to obtain the covariance matrix. In this case, a covariance matrix of size 8x8 was obtained, where each element represented the covariance between two variables.

For example, if a high positive covariance is found between $cantidad$ and $anio\_denuncia$, his would indicate that an increase in the frequency of reports ($cantidad$) could be associated with an increase in the year in which the reports were made ($anio\_denuncia$). This information is crucial for understanding the relationships between the variables and facilitates the interpretation of the principal components.

Once the covariance matrix was calculated, it was decomposed to find the eigenvectors and eigenvalues, which define the directions in which the data show the greatest variability. The projection of the normalized data on these new axes allows the dimensionality of the data set to be reduced while retaining most of the variability.

E. DATA TRANSFORMATION

Within the framework of the methodology employed, multiple data transformation techniques were carried out in order to improve the quality and ease of analysis of the data set in question. One of the techniques applied consisted of coding categorical variables using the method known as "One-Hot Encoding", which allowed the conversion of the original variables into numerical variables. In this process, separate columns were generated for each category present in the categorical variables.

As an example, in the case of the variable called"$distrito\_fiscal$", which represented the tax district where the report was made, binary columns were created corresponding to each of the tax districts, indicating the presence or absence of each district in each case analyzed. This same approach was systematically applied to the variables "$especialidad$", "$tipo\_caso$", "$generico$", "$subgenerico$", "$des\_articulo$", "$departamento\_pjfs$", "$provincia\_pjfs$" and "$distrito\_pjfs$".

This transformation made it possible to manipulate numerical variables that representatively reflected the original categorical variables, simplifying their analysis and comparison in the context of the data set. The availability of these numerical variables allowed the application of analysis and modeling techniques that require data in numerical format, including machine learning algorithms and data visualization techniques.

Additionally, to carry out the normalization, we started with the identification and correction of outliers in the actual numerical data set. This was done following a systematic approach based on descriptive statistical analysis. Key descriptive statistics were calculated, including the first quartile (Q_1) and third quartile (Q_3) of the numerical variable of interest, represented as column X in this context. These quartiles provide insight into the distribution of the data and establish a basis for anomaly detection.

Next, the Interquartile Range (IQR) was calculated, which is defined as the difference between Q_3 and Q_1:

$$IQR = Q_3 - Q_1.$$

Using the IQR, limits were defined to identify outliers. Values less than the $Lower\ limit$ or greater than the $Upper\ limit$ were considered outliers. These limits were calculated using the following formulas:

$$Lower\ limit = Q_1 - (1{,}5 * IQR),$$

$$Upper\ limit = Q_3 + (1{,}5 * IQR).$$

The outliers detected in column X were those that fell below the Lower Limit or above the Upper Limit. To ensure that these outliers did not negatively affect the analysis, they were corrected. The correction technique consisted of replacing the outliers by the median of column X. The median, represented as M, was calculated for column X and was used to update any outliers:

$$i = M,$$

where i represents an outlier in column X that is less than the $Lower\ limit$ or greater than the $Upper\ limit$.

For example, for the quantity variable, a detailed process was followed to ensure its accuracy and representativeness. The first and third quartiles were calculated and the interquartile range was determined. With the IQR, limits were established to identify outliers. Values outside these limits were considered outliers and corrected by replacing them with the median amount. This ensured a more robust and less biased data distribution, allowing for a more accurate representation of the frequency of complaints.

Table 3 provides clear criteria for the exclusion of outliers within the data set. Each step described in the table corresponds to a systematic method for identifying and correcting for these values, thus ensuring the accuracy of the analysis in the context of the study.

Table 3. Exclusion criteria

ID	Variable	Outlayer
11	*cantidad*	Step 1 : Calculate the interquartile range (IQR).
		Step 2 : We define the limits to identify outliers.
		Step 3 : We identify the outliers in our data.
		Step 4 : To correct the outliers, we replace these outliers with the median of the data set.

Figure 4 illustrates the Variable Normalization Process, which systematically addresses data handling and transformation.

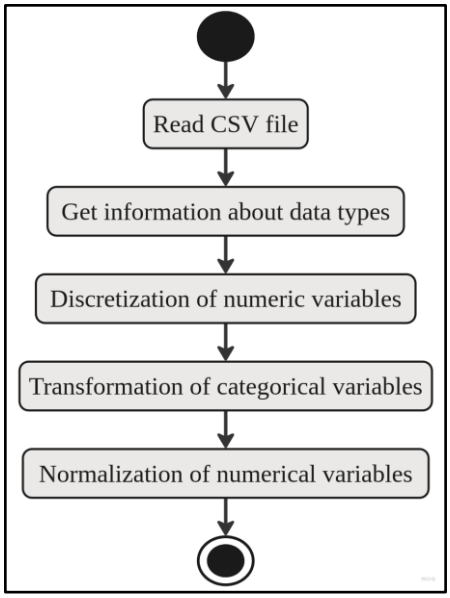

Figure 4. Normalization process for categorical variables

This process begins with reading the CSV file, followed by obtaining detailed information about the types of data. Next, the discretization of numerical variables is performed to group the data into defined intervals. Subsequently, categorical variables are transformed using coding techniques to facilitate their analysis. Finally, normalization is applied to the numerical variables, standardizing their values to ensure a uniform scale in the data set.

Table 4 provides a clear view of the original variables that have been converted into numerical format using coding techniques. Each entry in the table corresponds to a categorical variable that has been transformed to facilitate its analysis in the context of the study.

Table 4. Transformed categorical variables of the study

ID	TRANSFORMED VARIABLES
1	*periodo_denuncia_encoded*
4	*distrito_fiscal_encoded*
5	*especialidad_encoded*
6	*tipo_caso_encoded*
7	*generico_encoded*
8	*subgenerico_encoded*
9	*articulo_encoded*
10	*des_articulo_encoded*
13	*dpto_pjfs_encoded*
14	*prov_pjfs_encoded*
15	*dist_pjfs_encoded*

Table 4 shows how the study's categorical variables were transformed into numerical format using the coding method. Each original variable, such as $periodo\_denuncia$ and $distrito\_fiscal$, was converted to a numerical representation through the "One-Hot Encoding" technique. This transformation allowed variables such as $periodo\_denuncia\_encoded$ and $distrito\_fiscal\_encoded$ to be processed in analysis models and statistical algorithms.

F. CLUSTERING

To classify crimes, we employed the k-means algorithm, a widely recognized clustering technique. This algorithm groups crimes into k clusters, where k is a user-defined parameter. Each crime is assigned to the cluster whose centroid is the closest in terms of Euclidean distance.

APPLICATION OF EUCLIDEAN DISTANCE

To calculate the Euclidean distance between two points p and q, we use the transformed variables. For example, in a two-dimensional space using $periodo\_denuncia\_encoded$ and $distrito\_fiscal\_encoded$, points p and q represent the observations with these variables. The Euclidean distance is calculated as:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p - q)^2}\,.$$

This equation is generalized for a space with all variables transformed (n dimensions):

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (variable\_encoded_{i,p} - variable\_encoded_{i,q})^2}\,.$$

HANDLING OF CATEGORICAL VARIABLES

For categorical variables, we created a set of dummy variables. For example, $distrito\_fiscal$ was transformed into $distrito\_fiscal\_encoded$. We represent the original data in a matrix X, where each row i is an observation and each column j s a transformed categorical variable. The matrix X has dimensions (m, n), where m is the number of observations and n the number of transformed categorical variables.

SELECTION OF THE NUMBER OF CLUSTERS

For categorical variables, we use the Hamming distance, which measures the number of positions in which two categorical variables differ. We tested different values of k and used the Silhouette Score method to determine the quality of the clustering. We calculated the sum of Hamming distances within each cluster and selected the value of k that minimizes this sum.

EVALUATION METRIC

The Hamming distance between two partitions A and B is calculated as:

$$H(A,B) = \frac{1}{n}\sum_{i=1}^{n} I(A_i \neq B_i)\,,$$

where n is the total number of elements, m_i is the number of elements sharing the same group in both partitions, A_i and B_i are the sets representing the groups in partitions A and B, respectively.

EVALUATION METRIC

We use the silhouette coefficient to determine the optimal number of clusters. The silhouette value is calculated as:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))},$$

where $a(i)$ is the average distance between i and the other points in the same cluster, and $b(i)$ is the average distance between i and the points in the nearest cluster.

APPLICATION OF THE K-MODES ALGORITHM

We use the K-modes algorithm to find the centroids of the clusters and assign each observation to the nearest cluster in terms of the transformed categorical variables. The centroid matrix C has dimensions (K, n), where K is the number of clusters and n is the number of transformed variables. We perform cluster assignment using the Hamming distance between each observation x_i and each centroid c_k, and assign x_i to the cluster with the smallest Hamming distance. The

assignment matrix A has dimensions $(m, 1)$, where m is the number of observations.

To quantify the internal variance of a cluster $(W(C_k))$, we use the following metrics:

$$W(C_k) = \frac{1}{| C_k |} \sum_{i \in C_k} (x_i, c_k)\, d$$

where $d(x_i, c_k)$ is the Hamming distance between observation x_i and centroid c_k, and $| C_k |$ is the number of observations in cluster k.

III. RESULTS

In this study, we conducted an investigation on crime in Lima-Peru using Big Data Analytics, Data Mining and Clustering techniques. We show the identification of patterns and relationships in the crime data, facilitating the identification of homogeneous groups of crimes, then we present the findings and their relationship with the research objective, as well as their comparison with previous studies.

For the validation we considered the preprocessed and reduced data set, from 48,000 records to 500 records. Figure 5 shows the result of the application of Principal Component Analysis (PCA), where the dimensionality of 17 variables has been reduced to 15, which simplified the complexity of the data set.

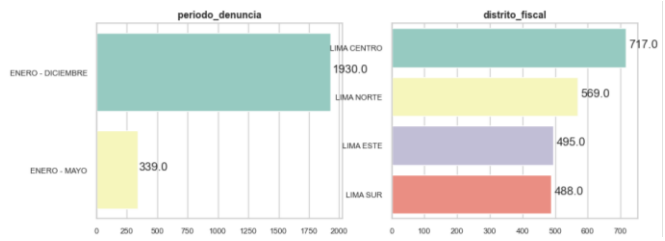

Figure 5. Results of applying PCA to the categorical variables.

DATA ANALYSIS

We applied the clustering technique using the K-Means algorithm. We used the Elbow Score method to determine the optimal number of clusters, finding that $k = 4$ was the ideal number. This clustering structure proved effective, allowing us a more accurate and meaningful understanding of the data distribution. Figure 6 shows the use of the Average Silhouette method to identify the optimal number of clusters.

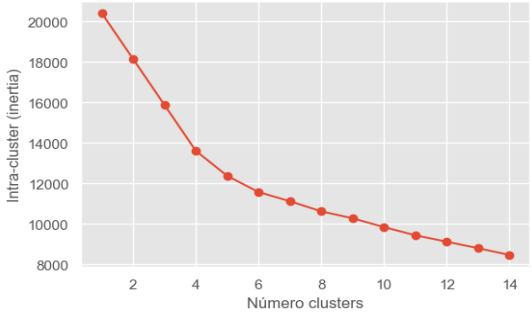

Figure 6. Average Silhouette method to identify the optimal number of clusters.

FINDINGS AND DISCOVERIES

In the analysis and discovery of crime clusters in Lima-Peru, shown in Figures 7, 8 and 9, cluster 0 – crimes against property, cluster 1 – crimes against life, body and health, cluster 2 – crimes against public safety and cluster 3 – crimes against public administration.

Cluster 0 – Crimes against property: This cluster represents crimes such as robberies and thefts, being the group with the highest incidence. This suggests a significant concern in the region regarding property security. It is crucial that the authorities and the community focus their efforts on prevention and security to address this issue. Figure 7 illustrates the number of crimes by type in this cluster.

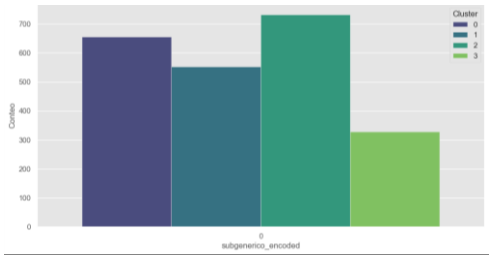

Figure 7. Number by type of crimes by cluster

Cluster 1 – Crimes against Life, Body and Health: This cluster includes homicides, injuries and assaults. Although less frequent than crimes against property, these crimes are of great importance due to their seriousness. Strategies should be implemented to reduce this type of crime and ensure the safety of citizens in Lima. Figure 8 presents a scatter plot of the classification of these crimes.

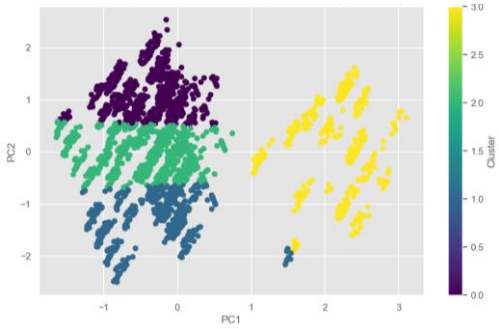

Figure 8. Crime classification scatter plot.

Cluster 2 – Crimes Against Public Safety: This cluster comprises crimes such as riots and public disorder. Although it is the least frequent cluster, its impact on overall community safety should not be underestimated. Addressing these crimes is essential to maintaining a safe and peaceful environment. Figure 9 shows a scatter plot of the classification of these crimes.

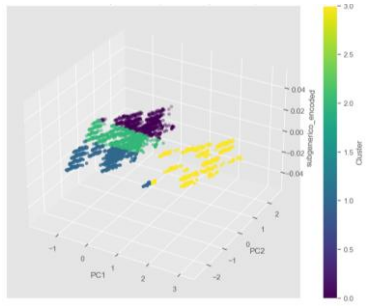

Figure 9. Scatter plot of crime classification.

Cluster 3 – Crimes Against Public Administration: This cluster includes crimes such as corruption and abuse of authority. Although less common than the other clusters, the presence of these crimes is a serious issue that requires attention. It highlights the importance of promoting integrity and transparency in the public sector and ensuring the accountability of authorities.

Our findings provide a clear picture of crime patterns, highlighting priority areas for security interventions and public policy. These results can serve as a basis for future research and crime prevention strategies in the region.

III. DISCUSSION

A. COMPARISON WITH PREVIOUS STUDIES

Compared to previous studies in the state of the art, the methodology employed in our research stands out for its comprehensive approach to address the problem of crime analysis in the Lima region. Unlike some studies that focus exclusively on standard Machine Learning algorithms, as observed in the research of [22], our research leverages CRISP-DM methodology and Big Data Analytics techniques to perform a comprehensive exploration of crime patterns.

Regarding the results obtained, our findings agree with the observations of [23] in terms of variability in crime types in urban contexts. However, unlike their purely descriptive approach, our research incorporates the application of clustering algorithms, such as K-means and K-modes, to identify specific clustering patterns in the data. This represents a significant advance, as it allows for a more precise and detailed categorization of crimes, overcoming the limitations of traditional descriptive methods.

Regarding data preparation, our methodology shares similarities with that proposed in [21], who advocated a rigorous approach to data cleaning and standardization to ensure the quality of the results. In the research they recognize the importance of data preparation as a critical phase in the analytical process. However, our research goes further by incorporating specific techniques, such as the use of PCA for feature selection, providing a more advanced methodology specific to our context.

In terms of contrasting strategies, the implementation of the CRISP-DM methodology in our research shares similarities with the approach proposed by [11], in the study they recognize the importance of a structured framework to guide the exploration, modeling and evaluation phases. However, our research differs by incorporating additional techniques, such as Hamming distance in the clustering process, allowing a more precise adaptation to the categorical nature of certain variables.

In terms of experimental design, our choice of a quantitative and predictive approach coincides with the recommendations of [24], who advocated the application of Machine Learning models to improve predictive capability. Despite this similarity, our research distinguishes itself by including a clustering strategy that reveals hidden patterns in the data, thus providing additional insight into crime dynamics.

B. LIMITATIONS AND AREAS FOR IMPROVEMENT

It is important to highlight the limitations and possible areas for improvement in our study, the test of our research focused on data from the Lima region of Peru, the results revealed reflect the current state of crime in that region, however, it is possible to apply the same procedures and techniques for other regions and/or other countries.

IV. CONCLUSIONS AND FUTURE WORK

The application of Big Data Analytics, Data Mining and Clustering in the analysis of crime data in Lima has provided deep insights into criminal dynamics. These approaches enabled the identification of patterns crucial to understanding the underlying factors influencing crime, significantly improving the understanding of specific crimes in the region. Addressing this issue is vital for security and quality of life, as a deeper understanding of crime trends allows for anticipating scenarios and adopting preventive measures, contributing to the reduction of crime incidence and a safer environment for citizens.

Data preprocessing and Principal Component Analysis (PCA) were applied to reduce dimensionality and improve the efficiency of the analysis. Finally, four clusters were identified, each one representing different types of crimes, which allowed a detailed and precise categorization of them.

As future work it is possible to replicate the same procedures to discover crimes in another country or another region, being the hypothesis that data mining and machine learning techniques will efficiently allow the discovery of knowledge in the security sector and welfare of citizens.

References

[1] P. R. Ventura Suclupe, & C. Etayo Pérez, “Informational treatment of crimes committed by minors,” *Studies on Journalistic Message*, vol. 23, issue 2, pp. 1005-1022, 2017. https://doi.org/10.5209/ESMP.58029.

[2] GI-TOC, The Global Organized Crime Index 2021: Taking the measure of crime, 2021, [Online]. Available at: https://globalinitiative.net/analysis/ocindex-2021/

[3] J. van Dijk, P. Nieuwbeerta, and J. Joudo Larsen, “Global crime patterns: An analysis of survey data from 166 countries around the world, 2006-2019,” *Journal of Quantitative Criminology*, vol. 38, no. 4, pp. 793-827, 2022. https://doi.org/10.1007/s10940-021-09501-0.

[4] A. Corbacho, J. Philipp, and M. Ruiz-Vega, “Crime and erosion of trust: Evidence for Latin America,” *World Development*, vol. 70, pp. 400-415, 2015. https://doi.org/10.1016/j.worlddev.2014.04.013.

[5] INEI, *Crime, Citizen Security, and Violence Statistics: A view from administrative records*, Technical Report, Lima, Peru: National Institute of Statistics and Informatics, April-June 2022.

[6] “Metropolitan Lima Regional Citizen Security Action Plan 2022,” *Metropolitan Lima Regional Citizen Security Committee*, 2022

[7] M. R. Keyvanpour, M. Javideh, and M. R. Ebrahimi, “Detecting and investigating crime by means of data mining: A general crime matching framework,” *Procedia Computer Science*, vol. 3, pp. 872-880, 2011. https://doi.org/10.1016/j.procs.2010.12.143.

[8] M. Feng, J. Zheng, J. Ren, A. Hussain, X. Li, Y. Xi, and Q. Liu, “Big data analytics and mining for effective visualization and trends forecasting of crime data,” *IEEE Access*, vol. 7, pp. 106111-106123, 2019. https://doi.org/10.1109/ACCESS.2019.2930410.

[9] P. R. Boppuru and K. Ramesha, “Spatio-temporal crime analysis using KDE and ARIMA models in the Indian context,” *International Journal of Digital Crime and Forensics*, vol. 12, no. 4, Art. 4, 2020. https://doi.org/10.4018/IJDCF.2020100101.

[10] T. O. Adewuyi, P. A. Eneji, A. S. Baduku, and E. A. Olofin, “Spatio-temporal analysis of urban crime pattern and its implication for Abuja Municipal Area Council, Nigeria,” *Indonesian Journal of Geography*, vol. 49, no. 2, Art. 2, 2017. https://doi.org/10.22146/ijg.15341.

[11] Y. Chen, J. Cai, and M. Deng, “Discovering spatio-temporal co-occurrence patterns of crimes with uncertain occurrence time,” *ISPRS International Journal of Geo-Information*, vol. 11, no. 8, Art. 454, 2022. https://doi.org/10.3390/ijgi11080454.

[12] E. J. Medina Hernandez and P. N. Ortiz Alvarado, “What characterizes cell phone theft in Bogota? Multidimensional analysis of complaints to the National Police in the period 2015-2018,” *Logos Ciencia & Tecnologia*, vol. 13, no. 1, pp. 19-35, 2021. https://doi.org/10.22335/rlct.v13i1.1225.

[13] U. M. Butt, S. Letchmunan, F. H. Hassan, and T. W. Koh, “Hybrid of deep learning and exponential smoothing for enhancing crime forecasting accuracy,” PLOS ONE, vol. 17, no. 9, 2022. https://doi.org/10.1371/journal.pone.0274172.

[14] F. Dakalbab, M. A. Talib, O. A. Waraga, A. B. Nassif, S. Abbas, and Q. Nasir, "Artificial intelligence & crime prediction: A systematic literature review," *Social Sciences & Humanities Open*, vol. 6, no. 1, Art. 100342, 2022. https://doi.org/10.1016/j.ssaho.2022.100342.

[15] Z. Ke and Z. Jin, "Research of crime prediction technology based on mathematical model," *Open Cybernetics and Systemics Journal*, vol. 8, no. 1, Art. 1, 2014. https://doi.org/10.2174/1874110X01408010860.

[16] M. Saraiva, I. Matijošaitienė, S. Mishra, and A. Amante, "Crime prediction and monitoring in Porto, Portugal, using machine learning, spatial and text analytics," *ISPRS International Journal of Geo-Information*, vol. 11, no. 7, Art. 7, 2022. https://doi.org/10.3390/ijgi11070400.

[17] S. Sathyadevan, M. S. Devan, and S. S. Gangadharan, "Crime analysis and prediction using data mining," *Proceedings of the 2014 Fourth International Conference on Communication Systems and Network Technologies*, 2014, pp. 406-412. https://doi.org/10.1109/CNSC.2014.6906719.

[18] H. Hassani, X. Huang, E. S. Silva, and M. Ghodsi, "A review of data mining applications in crime," *Statistical Analysis and Data Mining*, vol. 9, no. 3, pp. 139-154, 2016. https://doi.org/10.1002/sam.11312.

[19] S. Khalid, S. A. Khan, and S. Q. Ifzal, "A fuzzy logic-based framework for mapping crime data on established sociological hypothesis for societal disorder identification and prevention," *IEEE Access*, vol. 9, pp. 80197-80207, 2021. https://doi.org/10.1109/ACCESS.2021.3083542.

[20] Open Data Peruvian Government, MPFN Crimes: Public Ministry Office of the Prosecutor General., 2022, [Online]. Available at: https://www.datosabiertos.gob.pe/dataset/mpfn-delitos (in Spanish)

[21] Jason Brownlee, *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data*, Machine Learning Mastery, 2020.

[22] O. Llaha, "Crime analysis and prediction using machine learning," *Proceedings of the 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 2020, pp. 496-501. https://doi.org/10.23919/MIPRO48935.2020.9245120.

[23] A. Tymchyshyn, A. Semeniaka, S. Bondar, N. Akhtyrska, and O. Kostiuchenko, "The use of big data and data mining in the investigation of criminal offences," *Amazonia Investiga*, vol. 11, no. 56, pp. 278-290, 2022. https://doi.org/10.34069/AI/2022.56.08.27.

[24] Z. Wang and J. Wang, "Applications of machine learning in public security information and resource management," *Scientific Programming*, vol. 2021, Article ID 4734187, pp. 1-9, 2021. https://doi.org/10.1155/2021/4734187.

[25] T. Wang, C. Rudin, D. Wagner, and R. Sevieri, "Finding patterns with a rotten core: Data mining for crime series with cores," *Big Data*, vol. 3, no. 1, pp. 3-21, 2015. https://doi.org/10.1089/big.2014.0021.

[26] B. Arrigo and O. P. Shaw, "The de-realization of Black bodies in an era of mass digital surveillance: A techno-criminological critique," *Theoretical Criminology*, vol. 27, issue 2, pp. 265-282, 2023. https://doi.org/10.1177/13624806221082318.

[27] S. Changalasetty, W. Ghribi, A. Badawy, H. Bangali, A. Ahmed, L. Thota, R. Baireddy, and R. Pemula, "Using EM technique for juvenile crime zoning," *Proceedings of the 2021 IEEE International Conference on Soft Computing and Network Security (ISCON)*, 2021, pp. 1-6. https://doi.org/10.1109/ISCON52037.2021.9702353.

[28] A. Kumar, A. Kumar, A. K. Bashir, M. Rashid, V. D. Ambeth Kumar, and R. Kharel, "Distance based pattern driven mining for outlier detection in high dimensional big dataset," *ACM Transactions on Management Information Systems*, vol. 13, no. 1, Art. 1, 2022. https://doi.org/10.1145/3469891.

[29] E. P. Patulin and R. E. Talingting, "Crime prediction using autoregressive integrated moving average (ARIMA) algorithm," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 3, Art. 3, 2019. https://doi.org/10.30534/ijatcse/2019/59832019.

[30] G. Saltos and M. Cocea, "An exploration of crime prediction using data mining on open data," *International Journal of Information Technology and Decision Making*, vol. 16, no. 5, Art. 5, 2017. https://doi.org/10.1142/S0219622017500250.

[31] N. Tasnim, I. T. Imam, and M. M. A. Hashem, "A novel multi-module approach to predict crime based on multivariate spatio-temporal data using attention and sequential fusion model," *IEEE Access*, vol. 10, pp. 48009-48030, 2022. https://doi.org/10.1109/ACCESS.2022.3171843.

[32] S. Walczak, "Predicting crime and other uses of neural networks in police decision making," *Frontiers in Psychology*, vol. 12, 2021. https://doi.org/10.3389/fpsyg.2021.587943.

[33] P. Yerpude and V. Gudur, "Predictive modelling of crime dataset using data mining," *International Journal of Data Mining & Knowledge Management Process*, vol. 7, no. 4, Art. 4, 2017. https://doi.org/10.5121/ijdkp.2017.7404.

[34] J. Yin, "Crime prediction methods based on spatiotemporal data," *Discrete Dynamics in Nature and Society*, vol. 2018, Art. 1601542, 2018. doi: 10.1155/2018/1601542.

[35] A. Zeb, W. Rasheed, and A. Israr, "Spatiotemporal analysis of crime pattern and hotspot identification using GIS-based kernel density estimation," *Journal of Urban Management*, vol. 10, no. 4, pp. 34-51, 2021. https://doi.org/10.1016/j.jum.2020.12.001.

JORGE L. MARIN EVANGELISTA, Bachelor in Software Engineering from Universidad Nacional Mayor de San Marcos, Peru. His professional focus encompasses backend development, SQL and NoSQL databases, business intelligence, big data and Linux system administration. Committed to continuous learning, he is currently studying data analysis, data mining, advanced Linux administration and cloud computing.

GUSTAVO R. GUERREROS JACOBE Bachelor in Software Engineering from Universidad Nacional Mayor de San Marcos, Peru. His expertise is focused on data analysis, machine learning and web development. As a leader at Consigue Ventas, he manages systems teams and contributes to web projects, ensuring the improvement of the user experience and the achievement of business goals. Gustavo is committed to continuous learning, with recent courses in Data Analytics and Cybersecurity.

H. DAVID CALDERON VILCA, PhD in Computer Science, research professor of the "Artificial Intelligence" Group of the Universidad Nacional Mayor de San Marcos – Peru, advisor of undergraduate and graduate thesis projects related to Neural Networks, Machine Learning and Natural Language Processing. Professor of doctoral programs in other universities.