Identifying markers of exposure to the Colombian armed conflict: a machine-learning approach

Maria Isabel Cano Achuri

Tesis de maestría para optar al título de Magíster en Ingeniería

Asesores

José David López Hincapié, PhD.

Claudia Victoria Isaza Narváez, PhD.

Universidad de Antioquia

Facultad de Ingeniería

Maestría en Ingeniería

Medellín

2023

Identifying markers of exposure to the Colombian armed conflict: a machine-learning approach

MARIA ISABEL CANO ACHURI

Universidad de Antioquia
Facultad de Ingeniería
Grupo de Investigación en Sistemas Embebidos e Inteligencia Computacional – SISTEMIC
Medellín, 2023

Acknowledgements

I would like to express my gratitude to my advisor, Professor José David López, for providing me with all the necessary tools to undertake this project, for his constant support and confidence in my abilities, for guiding me throughout the process and for his patience. To my coadvisor Claudia Victoria Isaza, who motivated me to work consistently with her discipline, who kindly shared her knowledge and support, and without her help this work would not have been possible.

I am very grateful to all my family for the continuous support they have always given me. Especially, to my boyfriend Julián Gómez, who accompanied me unconditionally throughout the process, motivating and supporting me at all times. My father Freddy Cano, who has always encouraged and supported me in my academic career. To my aunt Dora Cano, who supported me in making difficult decisions and whose words of encouragement were always with me.

To all the colleagues I have met along the way, for their support and solidarity in moments of progress and setbacks, complaints and joys. I extend my heartfelt wishes for their continued success in all their future projects, just as they have supported me in mine.

To all of them, my most sincere thanks for being part of this very important project in my life.

Maria Isabel Cano Achuri

This work was supported by MinCiencias grant 111584467273.

Abstract

The Colombian armed conflict has affected, to some degree, its entire population. Health authorities require markers to determine the consequences of this exposure and provide appropriate mental health interventions. In this thesis, we propose a novel methodology to automatically find the features that best relate to the level of exposure to the armed conflict and related risks (drug use, college desertion, among others) using the information provided by unsupervised techniques. We use clustering techniques that do not use predefined labels to cluster the data and obtain relevant information from cluster centers. This methodology was tested on two databases with more than 500 mixed response variables (dichotomous, categorical, Likert scale, etc.), the first with 346 subjects with a direct measure of their level of exposure in the context of the Colombian armed conflict, and the second one with 9467 subjects, but without a direct measure of their level of exposure. For the latter, the missing data problem was addressed by finding the appropriate parameters for eliminating and imputing missing values. As a result, 60 features related to exposure were identified as markers, which divided the subjects into three groups, in which characteristics related to high levels of exposure were highlighted. We created an artificial neural network (ANN) based model to confirm the features found as markers of exposure to violence. The model was able to estimate the level of exposure with an accuracy of 99% in training and 76% in validation using the selected features as input.

Keywords

Armed conflict, Health, Feature selection, Unsupervised learning, Clustering.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

ARMED conflicts bring direct economic, humanitarian, and social consequences to the civilian population. These affectations commonly hinder mental health outcomes such as depression, anxiety, post-traumatic stress, suicide risk, and undesired changes in cognition [1, 2]. These outcomes also differ with the length and intensity of the conflict. Colombia, for example, has been involved in a long-lasting, low-intensity armed conflict (over 60 years, the oldest active in America) with millions of civilians directly and indirectly affected, especially those in rural areas. These conditions make this population more vulnerable to suffering mental health outcomes [3].

Moreover, its impact seems to go further than mental health issues since there is growing evidence that physical health could also be affected by long-duration conflicts like the Colombian one [4, 5], although it has been poorly studied. Authors have also found that the civilian population has been affected by the armed conflict similarly to ex-combatants and direct victims [6, 7, 8]. In these studies, higher levels of aggression and incidence of mental health disorders have been observed in populations with high exposure to conflict-related extreme experiences, even in subjects who do not perceive themselves as victims. This exposure could lead to long-term outcomes in mental health if not timely and adequately accounted for. Therefore, a better characterization of the population is required to help the health authorities and professionals.

Previous studies have found effects of exposure to conflict such as changes in executive processes [2], cognitive processes related to aggressive behavior measured with social skills questionnaires [9], emotional processing [10], behavioral and executive functions [11], empathy [12], among others. These populations were evaluated using questionnaires, scales, and tasks, sometimes synchronized with electroencephalography (EEG). However, most of these results were obtained primarily in small groups with methods that cannot be easily scaled to large populations (e.g., through EEG recordings). Therefore, the inclusion of other techniques, such as machine learning, could help to study the effects of exposure to conflict in a broader population.

Machine learning techniques have their foundations in mathematics and statistics, but they are more focused on prediction than traditional statistical inference methods. These techniques follow two basic approaches: i) supervised learning, which uses labeled data to train the machine, and ii) unsupervised techniques, which naturally find associations among data with sim-

ilar features without having the data labeled (see [13] for more information). Nevertheless, as sometimes civilians do not consider themselves as victims, the scope of unsupervised techniques could be better fitted to study this population because the labels may be misleading. Other authors who have studied violence and armed conflict using machine learning techniques [14, 15] have used regression methods, random forests, and deep neural networks to make dimensionality reductions looking for the most relevant features in the classification of violence. However, these studies use supervised techniques to select the variables (commonly veteran and victim), leaving aside the civilian population who has been affected by the conflict but misclassified as controls, as stated by [6]. Hence, it is desirable to find markers that relate exposure levels with the possible consequences on physical and mental health by implementing unsupervised approaches.

In unsupervised techniques, the algorithms learn from the data without needing previously defined labels or classes. Unsupervised learning focuses on grouping or segmentation tasks, called clustering, where the goal is to find groups with similar data in a dataset [16]. Clustering techniques can be based on hierarchical or partitioning methods; however, when dealing with large datasets, it is preferable to opt for partitioning techniques [17]. The most popular partitioning technique is K-means, which uses numerical data to group n observations into k groups in which each observation belongs to the group whose mean value is closest (this is explained broadly in Chapter 3). Other derived approaches have the same operating principle but with different minimization methods or types of data (e.g., categorical, see [18, 19] for details). The advantage of K-means is that it provides quantitative information about the characteristics of the groupings, which are later related to research questions.

In this thesis, we aim to find markers that indirectly relate subjects to the consequences of the armed conflict (i.e., without labels like control or ex-combatant). We do this because many civilians either are victims but have reasons for not being officially accounted for, or they do not perceive themselves as victims, despite being exposed to extreme experiences related to the conflict. Therefore, in Chapter 3, we propose an unsupervised-based methodology to accomplish this aim. We use clustering techniques to group subjects and assess whether these groupings are related to violence. If these relationships are found, information extracted from the clusters determines which features are most relevant for creating the groups. We propose this features as markers.

To test our approach, we used a database with a sample of civilian population with different levels of exposure to the armed conflict. Volunteers were evaluated through more than 500 economic, psychological, academic, and nutritional features, among other aspects. The evaluated features have mixed-type responses (numerical, categorical, and dichotomous, among others). For this study, there are records of more than 9,000 volunteers from rural and urban regions of Antioquia (Colombia). The database has missing data, a common issue with a high volume of subjects and variables. Although this topic has been accounted for, it still needs a proper guideline to follow [20, 21, 22]. Additionally, around 300 subjects have a direct measure of the level

of exposure to the conflict, which makes it an ill-conditioned sample due to the large number of variables in the database. The database is presented in Chapter 2, and our approach to account for missing values is presented in Chapter 4.

We expect that the unsupervised-based methodology proposed in Chapter 3 and tested in Chapter 5 will contribute to the social field by finding markers from features extracted from mixed-type response databases with missing data. Moreover, by applying our methodology to the databases mentioned above, we expect to provide additional information about the consequences of the Colombian armed conflict in the civilian population, with the main achievement of using solely easy-to-obtain surveyed information (e.g., the same questionnaires are surveyed to every new student of the University that recorded the larger database). With this information, experts in public health areas will have additional tools to develop social-cognitive training to prevent physical and mental health outcomes in the civilian population.

In summary, in this research, we seek to find markers of effects on the physical and mental health of people exposed to different levels of exposure. Two technical challenges with the databases were resolved before applying the machine learning approaches: a) The complexity of working with mixed response types of data and b) the missing data problem. Then, we propose a methodology to provide health markers of exposure to violence using unsupervised learning. The selected markers with the proposed methodology were validated through supervised machine learning techniques in Chapter 5.

1.1 Objectives

1.1.1 General

Propose a machine-learning-based methodology to identify markers of health outcomes due to exposure to the Colombian armed conflict in civilian population using databases with mixed- and missing data.

1.1.2 Specific

O1. Define a technique for treating large amounts of missing data in mixed-response type databases.

O2. Propose an unsupervised methodology for selecting variables from an ill-dimensioned mixed-response type database which relate with a direct measure of exposure to conflict.

O3. Propose an unsupervised methodology for selecting variables from a large mixed-response type database with missing data which relate with indirect exposure to conflict.

O4. Validate if the selected features are adequate health markers by creating a predictor of exposure to armed conflict.

1.2 Outline

This work is organized as follows:

Chapter 2 presents the database used in this work, describing each of the aspects, tests, and criteria for participation in this process. Additionally, we show how the data was divided for each analysis and our reasoning for the partitioning.

Chapter 3 presents the proposed methodology to identify the most relevant features in the database and the results obtained with a population with direct measurements of exposure levels.

In Chapter 4, the database with missing data is confronted, and an analysis of the possible techniques that can be used with this type of database is carried out. Additionally, the proposed methodology is used to determine the technique to be implemented in the next chapter.

In Chapter 5, the proposed methodology is tested with the complete database (without missing values), where exposure to conflict is measured indirectly. We present a comparison between the results obtained in chapter Chapter 3 and the ones obtained here, analyzed from both the mathematical and the application points of view, with input from experts in psychology. Additionally, the markers found above are tested with a supervised technique (Artificial Neural Network) to assess their ability to relate to the level of conflict exposure.

Finally, in Chapter 6 the conclusions and future work are presented.

1.3 Contribution of the research work

The main contributions of this work are:

- A methodology to identify relevant mental and physical health markers based on unsupervised techniques, which was tested in databases with a largely different number of subjects, showing robustness in the selection technique.
- A set of markers of possible health outcomes due to exposure to conflict which were validated by assembling a predictor of exposure levels with the help of supervised machine-learning techniques. These markers result from applying the proposed methodology to a specific problem.

This work was presented at the Second Colombian Conference on Applied and Industrial Mathematics (MAPI 2022). It was recognized with the "Best paper presented by a professional" award.

Finally, a product derived from this research was published in a Conference proceedings:

M. I. Cano, C. Isaza, A. Sucerquia, N. Trujillo, and J. D. López, "Markers of exposure to the colombian armed conflict: A machine learning approach," in *Advances in Artificial Intelligence–IBERAMIA 2022: 17th Ibero-American Conference on AI, Cartagena de Indias, Colombia, November 23–25, 2022, Proceedings*, pp. 185–195, Springer, 2023. https://doi.org/10.1007/978-3-031-22419-5_16

Chapter 2

Materials

THIS chapter describes the database used in the current work. We first describe the origin of the database, highlighting the importance of the scenario used to assess the health effects of the armed conflict. Then, each aspect in which the features are classified is described. Finally, the data distribution to achieve the objectives of this work is discussed, and the additional tests available to measure exposure levels are explained.

2.1 Database

The Wellbeing Unit of the University of Antioquia (www.udea.edu.co) provided us with a database from a characterization program with which the students can access different programs of accompaniment of university life. Aspects of family, academic, emotional, physical, and socioeconomic life are evaluated. Based on the information provided by the students, recommendations are made regarding self-care habits, practices, and a healthy lifestyle. The Unit accompanies them in the comprehensive training process [24].

The university is located in Antioquia (Colombia), one of the departments historically most affected by different types of violence, including the Colombian internal armed conflict, especially in rural areas [25]. The university offers the public service of higher education in the department's capital, Medellín, and nine rural regions. With this database, we aim to have a representative sample to analyze the physical and mental health of a civilian population with different levels of exposure to the Colombian armed conflict.

This database has 33,561 records of undergraduate and graduate students from different programs surveyed between 2016 and 2020. The students belonged to all socioeconomic statuses and were distributed along the 12 campuses. In total, there are 526 features divided into the eleven aspects listed below:

1. **Basic information:** In this aspect, the personal information of each individual is considered: identification, faculty, undergraduate program, region of their campus, sex, socioeconomic level, address, age, and if they are victims or internal displaced person (IDP) by violence.

2. **health and social security:** this aspect is focused on knowing if the student has a social security health system, if they have any disabilities, and how their health is in general terms.

3. **Academic:** this one inquires about whether the student has other studies (technical, technological, or professional), the reason why they chose the current academic program at the University of Antioquia, study habits and techniques, if they consider that the current career is what they have always wanted, among others.

4. **Psychological:** in this aspect, the consumption of psychoactive substances such as marijuana, alcohol, and cocaine, among others, and the frequency of consumption are evaluated. On the other hand, it evaluates if the student considers that activities such as chatting, surfing the internet, and gambling, among others, have affected their life in any aspect, such as family relationships, money, or health.

   Finally, the self-report questionnaire (SRQ) is surveyed in this aspect. The SRQ is a 22-question test that assesses the possibility of affective-type mental illness (depressive disorders or anxiety disorders) and psychotic disorders. Based on student responses, alerts for possible clinical indicators can be generated.

5. **Socioeconomic:** the housing conditions, the source of economic income, and the economic responsibilities in the family are evaluated. If they have any incentive inside or outside the university, such as being a teacher assistant or a researcher, or if they have government or private scholarships. Additionally, the family's monthly income and expenses are evaluated.

6. **Socio-familiar:** here, information about relatives is collected. Family dynamics are also evaluated, that is, if they consider them as support in decision-making-related aspects, if they speak, and if they are satisfied with the help received.

7. **Work:** in this aspect, it is inquired whether the student currently works, if this work is related to what they study, and what facilities they have to study and work.

8. **Sexuality and affectivity:** in this category, aspects related to sexuality are evaluated: if they know of contraceptive methods, sexually transmitted diseases, and which are their feelings related to affectivity and interpersonal relationships.

9. **Nutrition:** here is inquired about eating habits, such as how much food they eat per day, what kind of food they eat more or less frequently, how they feel about their body weight, and if they have used any techniques to lose weight.

10. **Sport, recreation and culture:** in this area, the artistic, physical, and cultural activities they practice are evaluated, and if they carry out group activities or which activities they usually carry out more frequently in their free time.

11. **Physical evaluation:** it consists of an evaluation of the body composition of the volunteers and some physical abilities. For this, measurements such as weight, height, blood pressure, and the measurement of abdominal folds, biceps, and thighs are taken. Additionally, a physical test is carried out that consists of taking the time it takes the volunteers to run 1.5 miles and later taking measurements of heart rate, oxygen consumption, and physical capacity. Finally, some questions are asked about whether the person is a smoker, suffers from diabetes, and how the frame of mind has been in recent days. Only 9,467 volunteers completed this evaluation.

The first ten aspects were obtained through a virtual survey via the university platform. The last aspect (the physical evaluation) required professionals to take the records in person and perform the necessary tests to collect said information. Table 2.1 shows the number of features by aspect.

Table 2.1: *Number of features by aspect*

Aspect	Number of features
Basic information	18
Health and social security	10
Academic	61
Psychological	82
Socioeconomic	36
Socio-familiar	61
Work	8
Sexuality and affectivity	30
Nutrition	40
Sport, recreation and culture	57
Physical evaluation	32
Risks	91

The 526 features (divided in the 11 categories) have different response types: i) numerically, such as anthropometric measurements, number of family members, and times of consumption of some substance. ii) Categorical, such as places of residence, academic program, housing types, employment, income ranges, economic expenses, and alerts from the SRQ questionnaire. iii) Dichotomous (yes/no) questions such as those about activities that affect daily life, inter-family relationships, study habits, or the SRQ questionnaire. iv) On the Likert scale, questions about the frequency of performing certain activities, frequency in the consumption of certain foods, and evaluation of the frame of mind, among others. Finally, v) Open responses where the subject can specify data or personal information.

2.2 Sampled databases

Below, we describe how the students were distributed to form two databases: the first one with a large number of subjects but without direct measurement of the level of exposure to

violence, and the second one with a considerably smaller number of subjects but with additional tests which directly measure the level of exposure.

2.2.1 Database 1

This database consists of the students who took both the virtual and the physical assessments. In total, the sample consisted of 9,467 subjects (5,239 females, 4,230 males) whose ages ranged from 15 to 63 years (mean=19.33, standard deviation(SD)=3.75). Of the subjects, 990 were considered victims of the armed conflict, and 1,012 were IDP due to the armed conflict. These were undergraduate students only. This database was used in Chapter 5. However, some variables had missing data, which was treated with and resolved in detail in Chapter 4, resulting in 515 features and the same number of initial subjects (9,467). This database was completely anonymized, removing personal information such as names, emails, phone numbers, and addresses.

2.2.2 Database 2

This second database is a subset of the Database 1. It consisted of 346 subjects who completed two additional tests and the evaluation of the first ten aspects of the Wellbeing Unit database. This thesis was supported by the research project *"Allostatic stress as a biomarker in predicting mental and cardiometabolic health outcomes in Colombian geographic regions affected by the armed conflict"* funded by the Colombian Ministry of Science (code 111584467273). As part of this project, these participants completed two additional psychological tests: The Extreme Experiences Scale (EX2) and the Interpersonal Reactivity Index (IRI).

- **Extreme Experiences Scale (EX2):** It is an 18-question instrument aimed to identify if a person has been highly exposed to situations of trauma, loss, or crisis due to the Colombian armed conflict [7]. This scale comprises two dimensions: direct (dEX2) and indirect (iEX2) extreme experiences, the first focused on personal physical situations and the second through third parties with whom the person has an emotional bond, such as relatives or friends. The response options are dichotomous (yes/no), and the result of this test is presented as the sum of affirmative responses per item. According to its validation [7], a person is considered to have a high level of exposure if they have a score greater than 2.5. This test was applied between 2019 and 2021 to students from all the university campuses. The result of this test provides a direct measure of the level of exposure to armed conflict.

- **Interpersonal Reactivity Index (IRI):** This test is used to assess empathy through 28 questions, allowing to measure the ability to understand others based on what is observed, verbal information, or perspective taking. The 28 questions of this scale are equally distributed into four categories: Perspective Taking (PT), Fantasy (FS), Empathic Concern (EC), and Personal Discomfort (PD). The first two categories evaluate cognitive processes, while the last two measure emotional reactions [26, 27]. The way to measure this scale is

based on Likert-type responses with five response options that range from 0 to 4 according to the degree of affirmation that corresponds (0 represents "it does not describe me well," and 4 represents "it describes me very well") .

Due to the anonymization of the databases, it was not possible to know if any of the 346 subjects had completed the physical evaluation, therefore, this aspect was not included for this subset. In this database, there was one master's student in biology; the rest were undergraduates. There were 173 males and 173 females aged between 16 and 53 years (mean=21.87, SD= 4.65). Of the subjects, 237 were considered victims of the armed conflict, and 227 were IDP. This database was used in Chapter 3.

2.3 Summary

This chapter presented the database provided by the University of Antioquia. Two samples of the database were obtained to develop the methodology proposed in Chapter 3, to solve the problems of missing data in Chapter 4, and to analyze the markers obtained with the proposed methodology in both databases in Chapter 5.

Chapter 3

Methodology

THIS chapter presents the proposed methodology to find the markers that relate levels of exposure to the Colombian armed conflict with aspects of daily life in the civilian population. The proposed methodology starts with a data preprocessing stage where we treat the mixed data and convert it to numerical values. Then, we present the proposed methodology, which uses an unsupervised technique that makes it possible to indirectly find groups that are related to violence. It also provides relevant information to detect markers of health consequences from the armed conflict. Finally, we present the results of testing our approach with Database 2 (presented in Chapter 2). Results are analyzed using existing conflict exposure labels but which were not used when applying the unsupervised technique. These results were subsequently validated in Chapter 5.

3.1 Data preprocessing

The database used in this work includes qualitative entries that must be converted to standard numerical ones to avoid biases in the classifier. The cleaning of data starts by removing those open response entries that may contain personal information (email, address, etc.). This step is required to protect the identity of the volunteers. Next, the yes/no responses were numerically binarized with 1 or 0, respectively. Subsequently, the features on the categorical Likert scale (for example, "it does not describe me well," "it describes me very well,") were changed to a numeric Likert scale between 1 and 5 respectively. Additionally, new variables were created for each categorical response, becoming dichotomous responses with values of 0 or 1 to indicate belonging or not to a specific category. Finally, a feature scaling was performed to normalize all the features in a range between 0 and 1:

$$x'_{i,f} = \frac{x_{i,f} - x_{\min,f}}{x_{\max,f} - x_{\min,f}}, \tag{3.1}$$

where $x'_{i,f}$ is the normalized value for subject i in the feature f, $x_{i,f}$ is the value to be normalized, $x_{\min,f}$ is the minimum value in the range of the feature and $x_{\max,f}$ is the maximum value in the same range.

Additionally, the 91 risk features, the victim and displaced features, and the alerts generated by the SRQ questionnaire were removed before using the quantified and normalized database. The questions and results of the EX2 and IRI tests were also not considered. These features would later be used to analyze the results obtained with the unsupervised technique. In total, 303 numerical variables were obtained to feed the algorithm of the methodology proposed below.

3.2 Proposed methodology

The proposed methodology consists of the flowchart presented in Figure 3.1. The process is iterative and begins by automatically grouping the subjects in nine partitions (from two to ten clusters) via k-means. Then, the adequate number of clusters is selected by applying an internal cluster index (Silhouette index[28]) to each partition. This index selects the partition with the most (within) compacted and (between) separated clusters. The next step involves finding the cluster center values (for each feature). Based on the measured center distances, the most relevant features are selected, and a new iteration begins with this new reduced set.

Starting from the second iteration and once an adequate number of clusters is determined, their correlation with conflict-related and risk features is obtained. If this correlation improves compared to the previous iteration, the algorithm continues by pruning non-relevant features and starting a new iteration with the reduced set. Otherwise, the process stops, and the set of features from the previous iteration is kept. These steps are explained below.

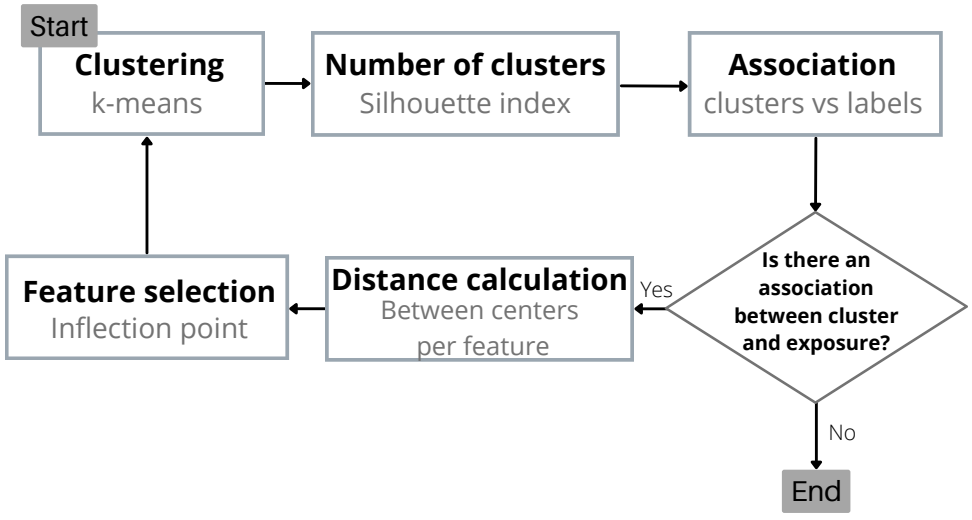

Figure 3.1: *Proposed methodology to find markers relating a general characterization of civilian population with their level of exposure to the armed conflict.*

Clustering

The first step consists of grouping the N_c subjects in clusters using the k-means algorithm [17]. Let define the objects $X = \{x_1, x_2, \ldots, x_{N_c}\} \in \mathbb{R}^{N_f}$, where each object is formed with the N_f features (or attributes) of a subject; and the clusters $C = \{c_1, c_2, \ldots, c_{N_k}\}$, where the number of clusters N_k is provided by the user. Our aim is to minimize the squared distance $\epsilon_{\text{k-means}}$ between the empirical mean of each k-th cluster (represented by its center $\mu_k \in \mathbb{R}^{N_f}$) and the dimensions of its own objects $X^{(k)}$:

$$\epsilon_{\text{k-means}} = \sum_{k=1}^{N_k} \sum_{i=1}^{N_c^{(k)}} \|x_i^{(k)} - \mu_k\|^2 \tag{3.2}$$

where $x_i^{(k)}$ is the i-th object belonging to the k-th cluster, which contains $N_c^{(k)}$ objects. The algorithm consists of the following steps:

1. Select the parameter N_k corresponding to the number of clusters.
2. Randomly create a starting position for the centers of each cluster.
3. Compute the Euclidean distance between each object and the centers.
4. Assign each object to the closest center.
5. To minimize the error of (3.2), μ_k is relocated from the calculation of the average of the objects that belong to the cluster c_k.
6. Repeat the steps 3 to 5 until it stabilizes (that means, that the centers practically do not move) or until reaching a maximum number of iterations.

To avoid local minima, several runs with different seeds should be made (we ran 500). The algorithm records the final value of $\epsilon_{\text{k-means}}$ per run, and returns the minimum one.

We repeated this procedure for nine partitions, varying the number of clusters in step 1 from $N_k = \{2, \ldots, 10\}$.

Finding the adequate number of clusters

Once all partitions are created, the one with the number of clusters that better groups the subjects (in terms of intra-group cohesion and inter-group separation) is validated through the silhouette index [28]. For each partition, the average dissimilarity $a \in \mathbb{R}^{N_c}$ among all objects from cluster c_k is calculated:

$$a(i) = \frac{1}{N_c^{(k)} - 1} \sum_{j, j \neq i}^{N_c^{(k)} - 1} d\left(x_i^{(k)}, x_j^{(k)}\right) \tag{3.3}$$

where $d\left(x_i^{(k)}, x_j^{(k)}\right)$ is the Euclidean distance between $x_i^{(k)}$ and $x_j^{(k)}$. Then, the average dissimilarity $b \in \mathbb{R}^{N_c}$ between the objects of c_k and the ones from the nearest cluster $c_{\widehat{k}}$ ($\widehat{k} \neq k$) is calculated:

$$b(i) = \frac{1}{N_c^{(\widehat{k})}} \sum_{j=1}^{N_c^{(\widehat{k})}} d\left(x_i^{(k)}, x_j^{(\widehat{k})}\right) \tag{3.4}$$

With a and b, silhouette values for each subject $s \in \mathbb{R}^{N_c}$ are computed:

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \tag{3.5}$$

each giving a score between -1 and 1. Finally, an average silhouette width per partition is computed: $S(j) = \text{mean}(s)$, with $N_k^{(j)} = 2, \ldots, 10$; and the chosen partition corresponds to $N_k = \arg\max(S)$.

Distance calculation

With the proper number of clusters N_k found, the maximum distances among all cluster centers per feature $\delta \in \mathbb{R}^{N_f}$ are computed:

$$\delta(f) = \max\left(\mu_f^{(k)} - \mu_f^{(\widehat{k})}\right);\ \forall k, \widehat{k} = 1, \ldots, N_k;\ k \neq \widehat{k} \tag{3.6}$$

with $f = 1, \ldots, N_f$. A graphic example is presented in Figure 3.2, where the largest distance among centers for Feature X is $\delta(\text{X}) = 0.7$ (between $\mu_{\text{x}}^{(1)}$ and $\mu_{\text{x}}^{(3)}$), while for Z is $\delta(\text{Z}) = 0.15$. All distances are computed from the largest center position, so all values are positive. This distance provides information about the relevance of the feature to the differences between the clusters. For example, in Figure 3.2, feature X is more relevant than feature Z.

Finding the most relevant features

In this step, the distances δ are arranged from largest to smallest and plotted as in Figure 3.3 (left). Then, the slope between the minimum and maximum values is calculated (see Figure 3.3 (mid)) and the graph is rotated until the slope is zero (Figure 3.3 (right)). Finally, the new minimum value is the inflection point (red dot in Figure 3.3 (right)), features found before that value are considered as the most relevant for the clusters (toy features X, V, and Q in this example).

Associating clusters with levels of exposure

Starting in the second iteration and once the partition with the adequate number of clusters is found, the correlation between the set of clusters and the features related to conflict and risks (labels) is computed. If the correlation is higher than in the previous iteration (i.e., if the reduced set of features behaves better), there is a chance that further reducing the number of features

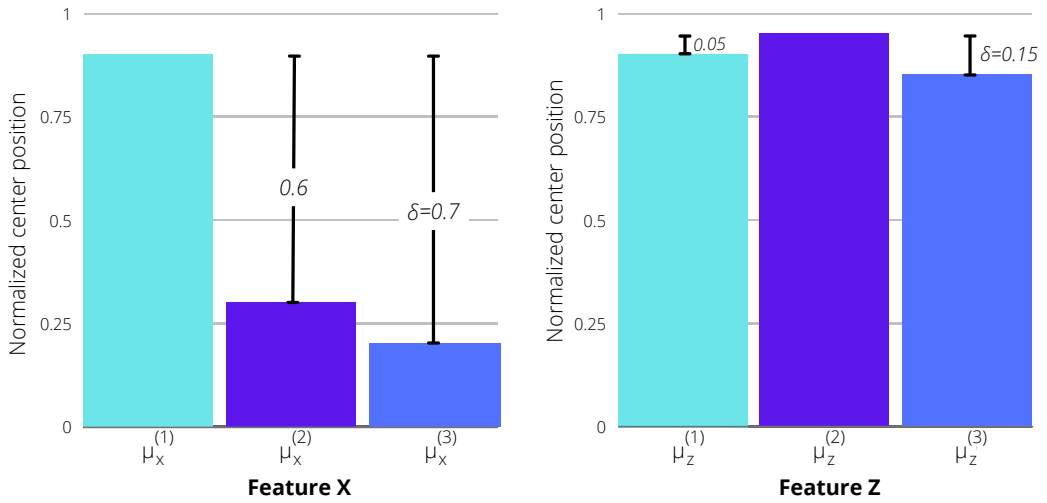

Figure 3.2: *Examples of distances among cluster centers for each feature. The y-axis represents the position of the centroid in each feature dimension. Distances for Feature X are significantly larger than for Feature Z, indicating that X allows better differentiation of clusters than Z.*

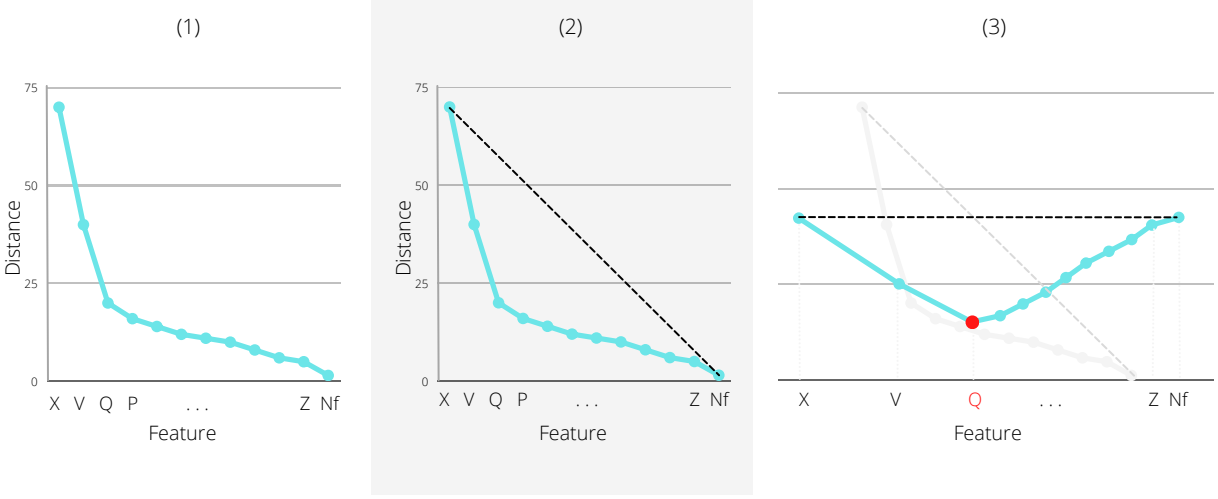

Figure 3.3: *Procedure to select the features that contribute the most to the clustering. The y-axis represents the largest distance δ found for each feature. Toy names X, V, Q, Z and Nf represent the features ordered in descending order. Left: Distances calculated with (3.6). Middle: The slope between the maximum and minimum distances is calculated. Right: The plot is rotated, and a new minimum is found. In this example, the inflection point occurs in feature Q; therefore, features X, V, and Q are the more relevant for the clustering.*

would improve the results. Otherwise, the set of features from the previous iteration is kept as the one that better differentiates the clusters in terms of levels of exposure to conflict and risks.

3.3 Results using Database 2

Applying the proposed methodology with the Database 2 led to find 62 features as relevant markers (a reduction of 79% from the 303 initial ones), grouping the subjects in three clusters after just three iterations. They are listed in Table 3.1. It shows which features were more relevant per aspect, numbering them in relevance order (1 being the most discriminant).

Table 3.1: *Relevant features defined as markers with the proposed methodology using Database 2. They are separated in the aspects to which each feature belongs, and each feature includes its relevance within the set.*

	Description
psychological	
1	SRQ: Do you feel sad?
2	SRQ: Do you feel bored?
3	SRQ: Do you feel tired all the time?
4	SRQ: Have you lost interest in things?
5	SRQ: Do you feel nervous or tense?
7	SRQ: Do you sleep poorly?
8	SRQ: Is it difficult for you to do your job? / Has your job been affected?
11	SRQ: Is it difficult for you to enjoy your daily activities?
12	SRQ: Do you have difficulty making decisions?
14	SRQ: Have you had the idea of ending your life?
16	SRQ: Do you have frequent headaches?
17	SRQ: Have you noticed interference or something strange in your thinking?
18	SRQ: Do you have a bad appetite?
19	Did you use this substance in the last year? Marijuana
22	SRQ: Are you unable to think clearly?
25	SRQ: Do you cry very often?
30	Do you consider that working has affected any aspect of your life?
32	SRQ: Do you suffer from tremor in your hands?
33	SRQ: Do you get scared easily?
34	SRQ: Are you unable to play a useful role in your life?
35	SRQ: Do you suffer from poor digestion?
39	Do you consider that the Internet has affected any aspect of your life?
40	Do you consider that chatting has affected any aspect of your life?
48	Do you consider that online games have affected any aspect of your life?
49	SRQ: Do you feel that someone has tried to hurt you?
52	Do you consider that social networks have affected any aspect of your life?
58	Do you consider that sex has affected any aspect of your life?
academic	
9	Study habits: Do you have a chair to study?
10	Study habits: Do you have a desk to study?
26	Study habits: is the place where you study illuminated?
37	Have you thought about quitting school?
41	Study habits: Do you have time to study?
42	Study habits: Do you have a smartphone to study?
academic	
43	Study habits: Do you have a quiet space to study?
46	Why did you choose the academic program? possibility of exchange to another program
47	Study habits: Do you have a desk to study?
59	Why did you choose the academic program? professional vocation
61	Why did you choose the academic program? low cost
62	Study habits: Do you have a solitary space to study?
socio-economic	
20	Type of home you live in: Room
29	Type of housing: Own
31	Type of housing: Leased
44	Who do you depend on financially? Parents
45	Who do you depend on financially? Yourself
54	What type of employment relationship do you have? independent
55	Type of housing you live in: House
socio-family	
13	Family dynamics: Are you satisfied with the time you and your family spend together?
24	Family dynamics: Do you discuss with each other the problems you have at home?
27	Family dynamics: Are you satisfied with the help you receive from your family?
28	Family dynamics: Are important decisions made together at home?
50	Do you feel that your family loves you?
51	Have you suffered pressure from your family in making decisions?
nutrition	
6	On most days of the week, do you have a snack between lunch and dinner?
15	On most days of the week, do you have set times for your meals?
21	On most days of the week, do you eat snacks between breakfast and lunch?
23	Are you currently using any strategy to keep from gaining or losing weight?
36	On most days of the week, how much fruit do you consume?
38	In a typical week, how many days do you eat fish?
57	In a typical week, how many days of the week do you consume legumes?
*others**	
53	In general terms, how do you think your health is?
56	do you have social security?
60	sex: female or male

*Others refers to health, social safety, and basic information

The features that contributed the most to the separation of the groups were: psychological aspects (specially the SRQ scale), followed by academic, and socio-economic features. From the 22 SRQ features, the 20 selected (see Table 3.1) are focused on depression and anxiety disorders. To a lesser extent, relevant aspects related to socio-family, nutrition, health and social safety, and basic information (see Table 3.2 for details). In past studies, demographic factors, age, sex, and education have been used for traditional analyses in populations related to the armed conflict [8, 29, 9]. However, our methodology uses the data to find natural clusters that in turn may be related to exposure to conflict, so that other useful aspects not considered before can be included.

Fig. 3.4 (left panel) shows the distances of the final set of features. This information was used to define the relevance list of Table 3.1 and allows the users to quantify this relevance, i.e., these distances provide traceability in post-hoc studies, where psychologist might orient their intervention by focus in cognitive and conduct trainings in specific aspects.

The clustering identified representative groups associated with exposure to conflict and risks.

Table 3.2: *Number of features by aspect found with the proposed methodology using the database 2*

Aspect	Number of features
Psychological	27
Academic	12
Socioeconomic	7
Socio-familiar	6
Nutrition	7
Health and social security	2
Basic information	1

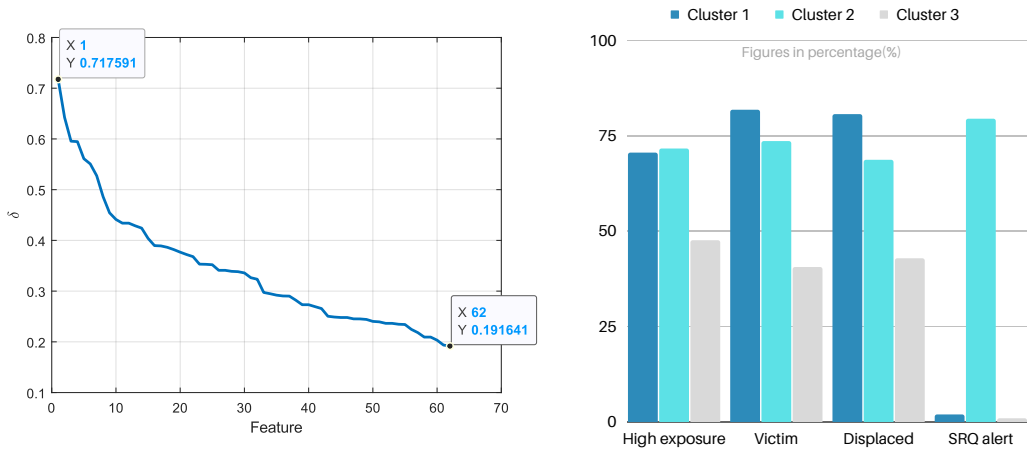

Figure 3.4: *Left: Maximum distance δ of each of the final 62 selected features. This chart provides the relative relevance among features. Right: Percentage of subjects in each cluster that present high exposure, consider themselves as victims of the conflict, are IDP, or presented high SQR alert. These charts provides the profiles later used by psychologists to define new alerts and provide cognitive and conduct trainings.*

Fig. 3.4 (right panel) shows that Clusters 1 and 2 included a higher percentage of subjects with those features more directly related to the conflict (e.g. "high exposure" from the EX^2) and mental health risks (e.g. SRQ results). The results of the IRI test did not show any differentiation between the groups, the scores in each cluster were similar for the four categories (TP, FS, EC and PD) and for the three clusters. Consequently, they were not conclusive for the analysis of the levels of exposure to the conflict. Therefore, it was not used in the current and future analyses. Note how the SQR based alert was highly associated to Cluster 2. This coincides with the top features shown in Table 3.1.

3.4 Conclusions

We proposed a methodology for finding relevant features that better relate to conflict-related and risk variables extracted from a characterization database with mixed personal information

(demographic, socioeconomic, psychological, among others). For this, we used k-means, which provides information of interest in quantitative values of the features in the feature space. It provided relevant information on the weight of each variable in the grouping (a future work could include other clustering techniques such as k-prototypes or hierarchical clustering). The selected features were chosen with a novel pipeline based on distance from the cluster centers, which resulted in a set of 62 markers from the initial set of 303 features, with a high relevance of the mental health SRQ scale. This latter relation has been used in the literature but not tested before in the Colombian armed conflict context.

3.5 Summary

In this chapter, We proposed a clustering-based methodology to find markers related to exposure to conflict. It was tested in Database 2. In the next chapter, this methodology will be used to deal with the problem of missing data in Database 1. Then, in Chapter 5, these results will be compared with those obtained with each of the databases to verify if the results obtained agree, are stable, and conclusive.

Chapter 4

Missing data

MISSING data is common in databases, especially those of high volume [30]. Database 1 presented in Chapter 2 has some missing values (MVs) that will make it challenging to apply the methodology proposed in Chapter 3. Therefore, filling in the empty data or removing subjects/variables from the database is necessary. There are three reasons for having MVs:

- Unwillingness of volunteers in certain measures (such as running 1.5 miles before to measure heart rate, oxygen consumption, and physical capacity).
- Transcription errors in the physical test, generating inconsistent values (TYPOs).
- Loss of information during the database saving process.

Missing data can be categorized depending on mechanisms, the first being the Missing Completely At Random (MCAR) mechanism, where the MVs are entirely unrelated to the data itself. That is, it is independent of the observed or missing data. In this case, the probability of an MV in a feature is random. The second is the Missing At Random (MAR) mechanism, where the probability of a MV item depends on the observed data; i.e., the probability that a data sample is missing from a feature may depend on values in another feature in the dataset. Finally, in the Missing Not At Random (MNAR) mechanism, the missing can depend on both observed features and missing data [21]. Missing values in this database were determined to be random. This suggests that applying imputation techniques is better to avoid biases [31].

Missing data can be treated by discarding or imputation. Discarding consists of eliminating the records that contain MVs, while imputation seeks to estimate the value of the MV using the information from the neighboring recordings or the information present in other features of the data set [32, 33]. As for discarding, there are three main ways to do it:

- Complete-case analysis (Listwise Deletion): This method is the simplest and eliminates all the data records with at least one MV. The disadvantage of this method is that it will also delete existing data, which could lead to significant data loss.

- Available-Case Analysis (Pairwise deletion): Unlike the previous one, this method only eliminates the box with the MV, keeping the information that is known. Its disadvantage is that each feature will have different dimensions, which complicates the data processing and makes the results difficult to interpret or not comparable.

- Dropping variables: If more than a percentage of MVs is missing for a feature, the entire feature is dropped. In this case, no metric specifically indicates from what value a feature should be eliminated or not. It depends on the application, the expert handling the data, or a specific condition. Thresholds between 10% and 50% of MVs are generally used [31].

There is the option of imputing the missing values to avoid a significant data loss, that is, generating the unknown data from the available one. In this field, many techniques could apply, but choosing one still will depend on the analyst's expertise, the types of data, etc.

In a 2020 review [31], the authors examined 111 journal papers published between 2006 and 2017, analyzing different techniques in various databases. Two large groups of techniques were distinguished, statistical techniques such as mean/mode, expectation maximization, least squares, Markov chain Monte Carlo, among others; and on the other hand, machine learning techniques such as decision trees, k-nearest neighbors, clustering, or random forest (to mention the most popular). More recent works [34, 35, 36, 37] evidence that there is still no single guideline for making imputations.

Based on these studies, we selected seven of the most commonly used techniques to determine the appropriate amount of data to impute in Database 1 and which technique provides better imputation results at the established threshold. It is worth noting that most studies analyzing missing data used small databases, except for a few that used sensor data and could achieve a high number of samples but a low number of features. It differs from Database 1 of this work, as we have hundreds of features of different aspects.

4.1 Materials and methods

4.1.1 Database

Database 1 was presented in Chapter 2. It has a sample of 9,467 subjects and 526 features. Of those features, 43 present MVs, where 21 are categorical, and 22 are continuous. Separated by aspects, 20 correspond to the physical evaluation, 17 to socioeconomic, 2 to sexuality, and one to academic, risk, family, and social security aspects. Additionally, the amount of MV is different in all features. Figure 4.1 presents a stacked bar chart by aspect, where each color represents a feature with MVs, and the horizontal axis represents the number of MVs. These vary between 7,861 MVs (corresponding to 83.04% of MVs for a feature) and 23 (0.24%).

In case of missing data per row, i.e., per subject, the number of missing features per subject is shown in Figure 4.2. Only 650 subjects had the complete information, and the remaining

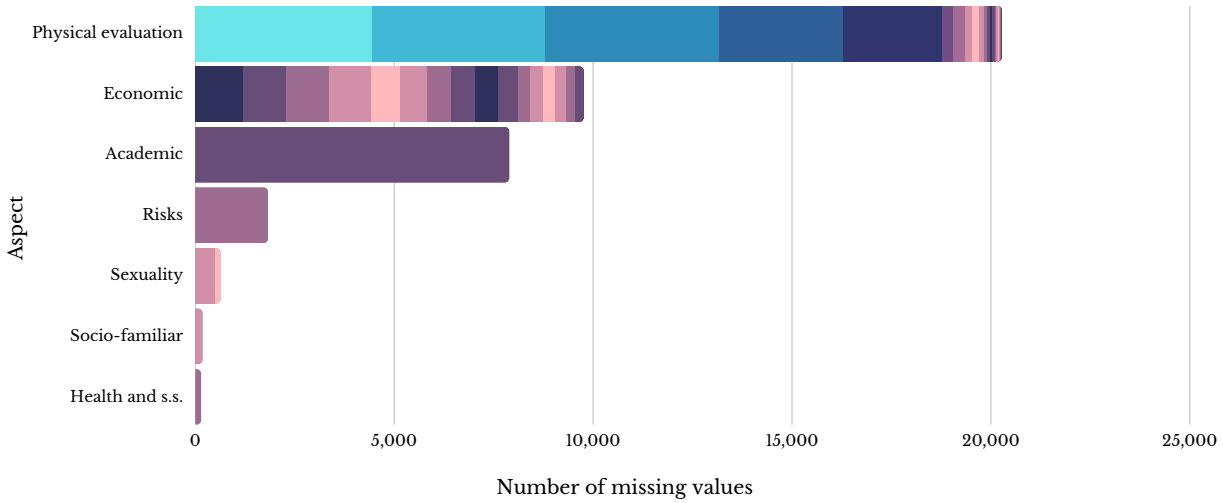

Figure 4.1: *Amount of missing data by feature and by aspect, where each color represents a feature with missing values.*

subjects had between 1 and 24 MVs. In this case, it was not necessary to eliminate by list since the subjects had less than 5% of missing data for each case.

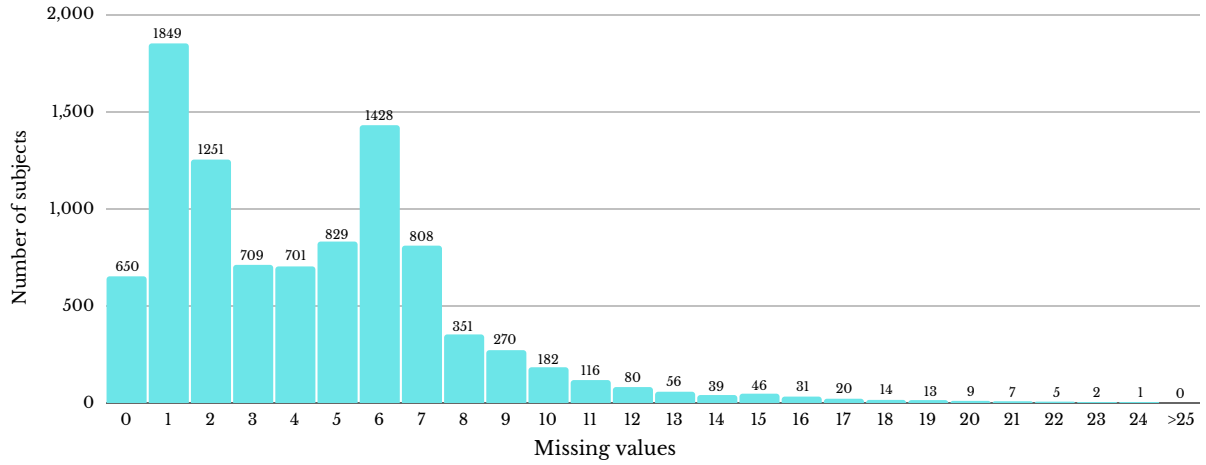

Figure 4.2: *Amount of missing data per subject.*

4.1.2 Methodology

We aim to prepare Database 1 to apply the methodology proposed in Chapter 3. The proposed solution to the problem of MVs is worked in a mixed way, that is, making imputations and elimination. Removals are made based on the amount of missing data a feature has. To do this, we must determine how much missing data to accept per feature. This is done by evaluat-

ing different thresholds to determine when to eliminate a feature. The missing data is imputed if the feature does not exceed the threshold.

To evaluate the results of imputation when a certain threshold is applied, databases are simulated with the subjects who have the complete information (pseudo database). Then, MVs are generated using an utterly random mechanism in proportion to the number of MVs in each feature in Database 1. For example, if a feature has no MVs, no MV is generated for that feature. If another feature originally had 30% of the missing data, MVs will be randomly generated at 30% in the pseudo database. By applying different missing data acceptance thresholds, features will be eliminated. In turn, the number of subjects with complete information will increase to create the subsequent pseudo-database and its evaluation.

Six thresholds are set: No threshold, 50%, 40%, 30%, 20%, and 10%. These thresholds are interpreted as "if a feature has this percentage of missing data; then it will be removed and not considered for imputation". At the first threshold (no threshold), all features are accepted no matter how much missing data it has. Thus, when creating the first pseudo-database (DB1), all subjects with some missing data are removed, leaving only those with complete information (i.e., without removing features, 650 subjects in all database have complete information). Next, the next threshold (50%) is applied, eliminating all features with more than 50% missing data. Once these features are eliminated, the second database (DB2) is created with subjects with complete information (i.e., after remove features with more than 50% MV, a total of 2225 subjects provided complete information). This process is repeated for the other thresholds. The information from each pseudo database is shown in Table 4.1.

Table 4.1: *Pseudo databases to apply imputation techniques.*

Database	Number of subjects	Number of features
DB1 (No threshold)	650	526
DB2 (50% threshold)	2225	525
DB3 (40% threshold)	2914	522
DB4 (30% threshold)	3294	521
DB5 (20% threshold)	4332	520
DB6 (10% threshold)	6350	515

With the generated pseudo databases, we simulated the MV amount for the corresponding features using the completely at-random mechanism (MCAR). Then, we applied the different imputation methods for each one, repeating this process three times with enough information for further analysis. The imputation results obtained are shown as the error percentage when comparing the pseudo and imputed databases.

Then, the imputation of MVs is carried out using seven methods selected from the literature, considering both statistical and machine learning techniques. These imputation techniques were applied using Python 3.8.3 and its libraries:

1. **Mean and mode:** simplest scenario, using the mean obtained with all the available data

for each numeric feature with MVs. For categorical features, the most frequent value for each feature is used, i.e., the mode excluding incomplete cases.

2. **Median and mode:** the middle number in a set of numbers ordered by value size. For categorical values, the mode is used again.

3. **Expectation maximization (EM):** The EM algorithm is a general-purpose iterative algorithm for finding maximum likelihood estimates in parametric models for incomplete data. This algorithm consists of two steps: expectation (E) and maximization (M). The steps to execute this algorithm are as follows (for more technical information, see [38]):

   (a) Choose a seed from the available data.

   (b) Expectation step (E): compute the objective function, which is, in the case of the missing data problem, equal to the expected value of the log-likelihood of the observed data, given the observed data and the current parameters.

   (c) Maximization step (M): determine the parameter vector maximizing the log-likelihood of the imputed data (or the imputed log-likelihood).

   (d) Iterate steps 2 and 3 until convergence, i.e., parameter estimate does not change over a tolerance between iterations.

4. **Markov chain Monte Carlo (MCMC):** generates pseudorandom samples from probability distributions through Markov chains. MCMC employs repeated random sampling to exploit the law of large numbers. The samples are generated by running a Markov chain, which is created so that its stationary distribution follows the input function, for which a proposal distribution is used. A full review can be found at [39].

5. **Clustering:** To use this technique in imputing MVs, the data with all features are grouped into a set of objects with similar characteristics, identifying the center (or centroid) of each group, which is the average of the values of the objects that belong to the same group. For each subject that has an MV, the cluster to which it belongs is estimated without regard to the MV. Then, the missing value is imputed using the closest cluster information. (technical details can be found in [40]).

6. **k Nearest Neighbor (kNN):** with the kNN method, the missing values of an observation are imputed based on the number of instances (k) in a dataset with similar information. It uses a distance measure d, corresponding to the Euclidean distance between two instances x_i and x_j:

$$d\left(x_i, x_j\right) = \sqrt{\sum_{h=1}^{N_f} \left(x_{ih} - x_{jh}\right)^2} \tag{4.1}$$

   where N_f is the number of features. The mutually observed features are used to calculate the distance between observations [41].

The imputation technique can be described as follows:

- For each observation, the distance function d is applied to find the k nearest neighbor vectors within the training data set.
- The missing features are imputed by the average of the features corresponding to those nearest neighbors.

This technique has some issues regarding the multiple options that can be presented, such as the distance function (we used the Euclidean distance) and the choice of k. In this case, cross-validation was used to determine the best value of k for each case [42].

7. **Random forests (RF):** The basis of RF is decision trees, a non-parametric learning method for classification and regression that seeks to find the best partition to divide the data into data subsets. These divisions are made from questions (the decision nodes of the tree) that allow dividing the data until reaching a final decision. Random forest models comprise a set of individual decision trees, each trained on a slightly different data sample. This allows a prediction to be obtained from the predictions of all the individual trees (for more information, see [43]).

Finally, the imputation results are assessed by directly evaluating the difference between the original value in the pseudo-database and the estimated or predicted value in the simulated incomplete database. The Mean Absolute Percentage Error (MAPE) was used for the imputations of continuous values:

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{x_i - \widehat{x}_i}{x_i} \right| \tag{4.2}$$

while for categorical values, discrete attributes and their scoring are performed with the percentage of incorrect predictions obtained as follows:

$$\text{Percentage of incorrect predictions} = 100 \times \frac{\text{number of incorrect predictions}}{\text{total number of predictions}} \tag{4.3}$$

Finally, the imputation error will be taken as the average between the $MAPE$ and the *percentage of incorrect predictions*. Based on the results of each imputation method applied to the proposed thresholds, we will determine which is most appropriate for this database. This makes it possible to define the features to be eliminated and the imputation method that gives the best results. This procedure is summarized in Figure 4.3.

4.1.3 Statistical analysis

A two-way ANOVA is performed to determine if the feature acceptance threshold and the imputation method significantly affect the imputation error. Three observations are made for each case (for each method and each threshold). This procedure is performed using R software, and the effects are reported as significant at $p < 0.05$.

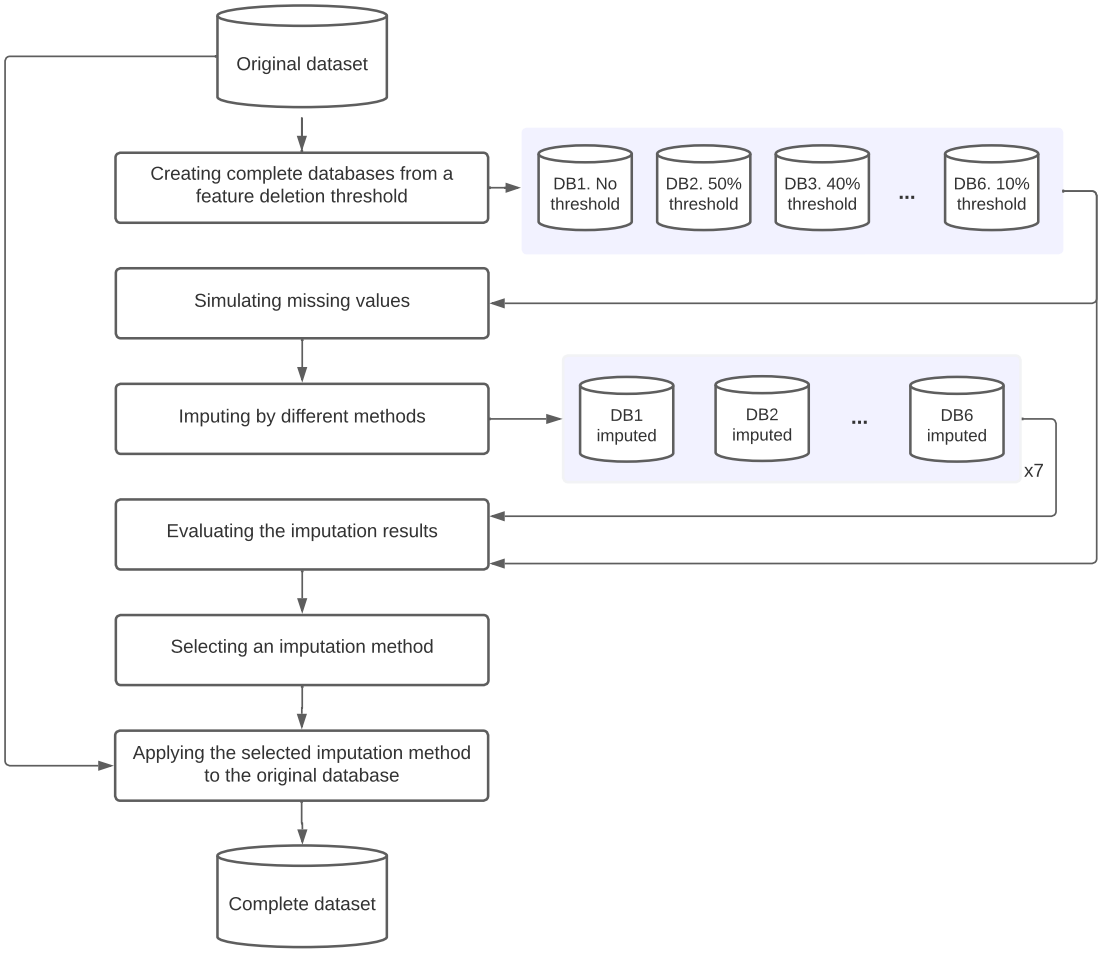

Figure 4.3: *Summary of the proposed methodology.*

4.2 Results

Table 4.2 shows the descriptive statistics of the results of the proposed methodology. Regarding execution times, almost all techniques produced results in a few seconds in a personal computer (all less than 2 seconds), except for the Random Forest technique, which took a few minutes depending on the size of the database, and Clustering, which took about 10 seconds.

These results evidence that the kNN technique should be discarded due to its low performance. Henceforth, this imputation technique was excluded from subsequent analyses. The results of the imputation method (excluding kNN) and the threshold vs. the error are shown in Figure 4.4. A two-way ANOVA was performed to determine if the choice of threshold and imputation technique resulted in differences in imputation errors. The results are presented in Table 4.3.

The results in Table 4.3 show that the threshold used affects the imputation error and that

Table 4.2: *Descriptive statistics for the imputation results in terms of the error.*

	mean/mode **M ± SD [%]**	**median/mode** **M ± SD [%]**	**kNN** **M ± SD [%]**	**clustering** **M ± SD [%]**	**EM** **M ± SD [%]**	**RF** **M ± SD [%]**	**MCMC** **M ± SD [%]**
No threshold	3.949 ± 0.000	4.499 ± 0.000	12.688 ± 2.980	4.585 ± 1.287	4.952 ± 1.225	4.692 ± 2.153	3.811 ± 2.206
50% threshold	4.529± 0.000	5.815 ± 0.000	13.355 ± 2.919	4.537 ± 2.049	4.703 ± 1.676	4.804 ± 1.908	4.942 ± 1.583
40% threshold	3.015 ± 0.000	3.268 ± 0.000	12.440 ± 1.956	4.198 ± 2.230	3.299 ± 2.123	0.872 ± 2.100	1.353 ± 2.053
30% threshold	0.981 ± 0.000	2.195 ± 0.000	12.441 ± 2.919	1.058 ± 2.061	1.307 ± 1.990	1.096 ± 2.163	2.390 ± 1.920
20% threshold	2.179 ± 0.000	1.015 ± 0.000	12.075 ± 1.755	1.069 ± 1.942	2.145 ± 2.494	0.524 ± 1.051	2.173 ± 1.517
10% threshold	1.015 ± 0.000	1.329 ± 0.000	13.155 ± 2.918	1.069 ± 1.942	2.145 ± 2.494	0.246 ± 2.106	2.133 ± 2.179

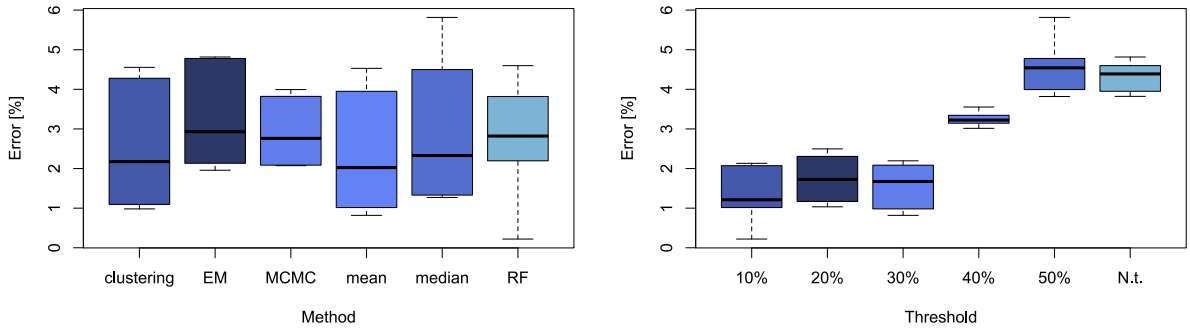

Figure 4.4: *Box plot for imputation method vs. error (left) and threshold vs. error (right).*

Table 4.3: *Two-way ANOVA results for imputation error.*

	Type III Sum of Squares	df	Mean Square	F	p
Intercept	945.744	1	945.744	483.455	<0.001
Method	3.978	5	0.796	0.407	0.843
Threshold	**221.798**	**5**	**44.360**	**22.676**	**<0.001**
Method*Threshold	47.528	25	1.901	0.972	0.513
Error	140.848	72	1.956		
Total	1359.896	108			

significant differences exist in the thresholds chosen. However, the imputation method does not have a statistically significant effect on the error.

Figure 4.4(right) shows that, regardless of the method, the 10% threshold is the one that gives the best results in terms of imputation error. When analyzing the results of each method at this threshold, even though RF presented the smallest errors, the execution time compared to the other methods ,excludes it as the best option. Although all the techniques presented low percentages of errors at this threshold, those that offer the smallest errors could be considered, which could be mean/mode and clustering for this database. As mentioned, the clustering technique took a longer execution time than the mean and mode. Additionally, it required the evaluation of an additional parameter to select the number of clusters that best grouped the subjects. Therefore, we decided to use a threshold of 10% and to impute using the mean/mode for

Database 1. Thus, eleven features were eliminated (described in Table 4.4), leaving 515 features available. The number of samples remained the same (9,467).

Table 4.4: *Dropped features.*

Feature	Aspect	Data type
Academic grade point average	academic	continuous
Heart rate	physical evaluation	continuous
Time them takes to go 1.5 miles	physical evaluation	continuous
Pulse	physical evaluation	continuous
Oxygen consumption calculated from the physical test	physical evaluation	continuous
Risk related to the level of physical activity	physical evaluation	categorical
Total monthly expenses	socioeconomic	categorical
College tuition Expenses	socioeconomic	categorical
Monthly education expenses	socioeconomic	categorical
Total economic income	socioeconomic	categorical
Vascular age calculated from indirect variables	physical evaluation	continuous

4.3 Discussion and conclusion

The information provided in Figure 4.1 shows that the highest missing data occurred in the physical assessment, where more than half of the features had missing data (between 0.24% and 46%). However, given that this information was not available in Database 2, it was relevant to use the maximum available data from this aspect to determine whether the consequences of exposure to extreme experiences were associated with physical health. The academic aspect presented a single feature with MVs, but 83% of the data was lost, so it had to be discarded. The MVs in the other features are lower or present in as many features.

For the amount of missing data per row (i.e., per subject) in Database 1, we found that more than 90% of the subjects had at least one MV, resulting in only 650 subjects having complete information. However, with a high number of performances (initially 526), the number of MVs per feature is relatively low (4.75% in the worst case). This value is superficial and allowed us to define that it is not necessary to eliminate any subject due to this problem, obtaining a final database with a sample of 9,467 as initially presented in Chapter 2.

The artificially generated databases made it possible to evaluate the performance of MVs at different acceptance levels. It was noteworthy that with a higher threshold, more subjects would have complete information, and the sample for the pseudo-DB would be higher. This allowed more information to be used in the imputations and a smaller amount of data to be imputed. The statistical differences were assessed with the ANOVA results in Table 4.3, where significant differences were found in the threshold applied but not in the imputation method to be applied.

K-Nearest Neighbors (kNN) had the worst performance at the time of imputation, where the technique proved unstable and had low predictive power for this particular database. According

to [42], kNN usually works better when the number of features is small, which was not the case for us.

This leads to two conclusions: i) accepting features with an MV percentage of less than 10% reduces the imputation error. Applying this to Database 1, if a feature has about 900 or more MVs, it could be dropped instead of trying to impute it; and ii) the imputation techniques provided excellent approximations for imputing missing data. The kNN technique was the least successful of all. The RF technique was computationally expensive. The clustering technique required evaluating the best number of clusters at each time, making its processing require more computational resources but still acceptable. Finally, applying the mean and mode technique to perform the imputations was the option chosen for the next chapter.

4.4 Summary

This chapter introduces the problem of missing data in Database 1. We implemented a methodology to estimate the most appropriate threshold to accept features with MVs. We found that the imputation shows homogeneous results with the different methodologies applied. This gives us confidence that the imputed Database 1 is stable and that the imputation does not affect the reliability of its use. In the next chapter, the imputed Database 1 will be used to apply the feature selection method proposed in Chapter 3. But this time with larger population size. The results will be compared with those obtained with Database 2.

Chapter 5

Case study

CHAPTER 3 presented a methodology that proposes identifying markers of the consequences of exposure to conflict using a database of 346 subjects (Database 2) with a direct measure of exposure (the EX2 test). Sixty-two markers were found to be associated with exposure to conflict, with a strong relationship between this measure of exposure and the results of the SRQ questionnaire. In Chapter 4, we dealt with the missing data problems in Database 1, eliminating variables with large amounts of missing data and imputing the other incomplete data. The database showed to be robust to MVs.

In this chapter, the methodology proposed in Chapter 3 is tested with Database 1 to identify markers from a larger sample. This database has an aspect that Database 2 misses: physical features, such as anthropometric measurements, and records of biological variables, such as pulse and pressure, among others. We finish this chapter by implementing a supervised technique using the EX2 test as a label to determine whether the features are relevant as markers of exposure to conflict and, in turn, to create a predictor of exposure levels.

5.1 Procedure

The methodology proposed in Chapter 3 is applied here, in this case using Database 1 after the imputation/elimination method defined in Chapter 4. Initially, the database had a total of 515 categorical and continuous variables. We first removed the open-response variables, then the categorical variables were quantified, and finally, the values were normalized between 0 and 1, as described in Section 3.1. A total of 324 numerical features resulted from the preprocessing stage. Next, the algorithm includes these features using the proposed methodology for marker identification. The results obtained here will be analyzed in contrast to those found in Chapter 3 (Table 3.1).

Finally, since the EX2 test is a direct measure of exposure to extreme experiences in conflict-related contexts, it would be desirable to determine the capabilities of the selected feature set to estimate high conflict exposure without including the EX2 test. For this purpose, we created an artificial neural network (ANN) using the found markers with the proposed methodology as the input layer, and the high or low exposure classification of the EX2 test as the output layer.

The ANN consists of a feed-forward network with four hidden layers, each with the same number of neurons (corresponding to the number of markers). A random 80/20 partition was used for training and testing. The activation function for the hidden layers was a symmetric sigmoid transfer function, and a descending gradient with an adaptive learning rate was used. A 5-fold cross-validation was performed to avoid generalization problems. The choice of these parameters depended on generalization tests to avoid overfitting in network training [14]. See [44] for a deeper insight into these algorithms.

The results were analyzed using the subjects from Database 2 since they were the ones who had taken the EX2 test. With this aim, we first balanced the classes, leaving 120 samples per class and a total population of 240 subjects. The procedure was implemented for two cases, first using the 62 markers found in Chapter 3 as input variables (Table 3.1). The second was implemented using the markers obtained using the Database 1 in this chapter.

5.2 Results

The proposed methodology using Database 1 showed 60 variables as relevant markers or indicators of health consequences due to exposure to conflict. The selected features are shown in Table 5.1, grouped by aspects and numbered in order of relevance, with 1 being the most discriminating. These results represent a reduction of 81.48% of the initial number of possible indicators. The number of markers per aspect is shown in Table 5.2.

Table 5.1: *Relevant features defined as markers with the proposed methodology using Database 1. They are separated in the aspects to which each feature belongs, and each feature includes its relevance within the set.*

	Description
psychological	
1	SRQ: Do you feel bored?
4	SRQ: Have you lost interest in things?
5	SRQ: Do you feel sad?
6	SRQ: Do you feel nervous or tense?
7	SRQ: Do you sleep poorly?
8	SRQ: Do you have difficulty making decisions?
9	SRQ: Do you feel tired all the time?
10	SRQ: Is it difficult for you to enjoy your daily activities?
12	SRQ: Have you noticed interference or something strange in your thinking?
13	Do you consider that the Internet has affected any aspect of your life?
14	Do you consider that social networks have affected any aspect of your life?
15	SRQ: Have you had the idea of ending your life?
17	SRQ: Do you have frequent headaches?
19	Do you consider that chatting has affected any aspect of your life?
21	SRQ: Do you have a bad appetite?
22	SRQ: Do you cry very often?
23	SRQ: Are you unable to think clearly?
25	SRQ: Is it diffcult for you to do your job? / Has your job been affected?
27	SRQ: Do you get scared easily?
29	SRQ: Do you feel that someone has tried to hurt you?
36	Have you suffered pressure in decisions related to love and sexuality?
37	SRQ: Do you suffer from poor digestion?
38	SRQ: Are you unable to play a useful role in your life?
46	Do you consider that using headphones has affected any aspect of your life?
52	SRQ: Do you suffer from tremor in your hands?
academic	
2	Study habits: Do you have a desk to study?
3	Study habits: Do you have a chair to study?
11	Study habits: Do you have a quiet space to study?
18	Study habits: is the place where you study illuminated?
28	Study habits: Do you have a habit of taking time to study?
30	Study habits: Do you have a smartphone to study?
41	Is the university degree you are doing the one you have always wanted?

	Description
academic	
42	Do you have internet to study?
50	Study habits: Do you usually study alone?
53	Study habits: Do you have a desktop computer?
54	Study habits: Do you usually read texts?
57	Study habits: do you have a laptop?
58	Study habits: Do you usually study at night?
nutrition	
16	On most days of the week, do you have set times for your meals?
20	On most days of the week, do you have a snack between lunch and dinner?
24	On most days of the week, do you eat snacks between breakfast and lunch?
32	How do you currently feel with your body weight? Normal
35	In a typical week, how often do you eliminate any of the three main meals (breakfast, lunch and dinner) of the day?
44	In a typical week, how many days of the week do you consume nuts and seeds?
47	On most days of the week (4 days or more), how many vegetable salads do you eat?
55	In a typical week, how many days of the week do you consume avocado or olive oil?
56	On most days of the week, how much fruit do you consume?
60	In a typical week, how many days of the week do you consume legumes?
Sports, recreation and culture	
33	Do you participate in groups or activities that are related to group sports?
40	Of the physical activity you do the most, how often do you do it weekly?
45	Do you go to the cinema in your spare time?
48	do you go to the gym?
socio-family	
26	Are you satisfied with the time you and your family spend together?
31	Are important decisions made together in the house?
59	Are you satisfied with the help you receive from your family when you have problems?
health and social security	
39	In general terms, how do you think your health is?
43	Do you have social security?
49	Have you visited the dentist in the last year?
socioeconomic	
34	monthly family income
51	Type of housing: Own

The clustering technique divided the subjects into three different sets. The psychological as-

Table 5.2: *Number of features by aspect.*

Aspect	Number of features
Psychological	25
Academic	13
Nutrition	10
Sport, recreation and culture	4
Socio-familiar	3
Health and social security	3
Socioeconomic	2

pect is the one that contributes most to the separation of the groups, with 25 features, of which 20 correspond to the SRQ test and the remaining five to aspects of daily life that have affected their life in the last six months (health, time, money, family, social, academic, or work relationships).

As for the groups, the 60 markers found are distributed as follows: The first group has 2,459 subjects, the second has 3,617, and the third has 3,391 subjects. These groups have a specific profile that is related to the selected variables. In the first group, which we will call Cluster 1, about 50% of the students have an alert in the SRQ questionnaire, while in the other two groups, less than 1% present an alert in this questionnaire. In addition, Cluster 1 also showed a higher percentage of members with affirmative answers to questions with negative aspects (for example, do you feel bored, sad, or tired?).

Additionally, we found that Cluster 1 has the highest number of warnings (i.e., risk variables) for psychological and academic risks, among others. In the next group (Cluster 2), the variables related to positive aspects, such as the consumption of certain foods, study habits, and sports, have a higher percentage of students whose answers are positive. Finally, the third group (Cluster 3) does not stand out in any of these profiles, being an intermediate point between the two previous profiles. An example is shown in Figure 5.1, where the number of subjects per group that responded positively to some of the questions listed in Table 5.1 is shown proportionally. It is important to emphasize that the proposed methodology found this information without any prior labels, using only the data available in the databases.

The next stage of our validation consisted of determining the ability of the markers obtained with both databases to predict individual exposure to conflict (using the EX2 score as a label). An ANN was implemented to this purpose. Using the 62 markers obtained with Database 2, the results showed a mean accuracy of 99.58% (SD=0.83) for the training stage and 75.36% (SD=0.98) for the test stage. With the 60 variables obtained from Database 2, the results obtained showed an accuracy of 99.33 (SD=0.37) in training and 76.08 (SD=3.08) in testing.

5.3 Discussion

The obtained results of applying the proposed methodology in Database 1 highlight the similarities and brief differences between the results obtained with a sample of 346 volunteers

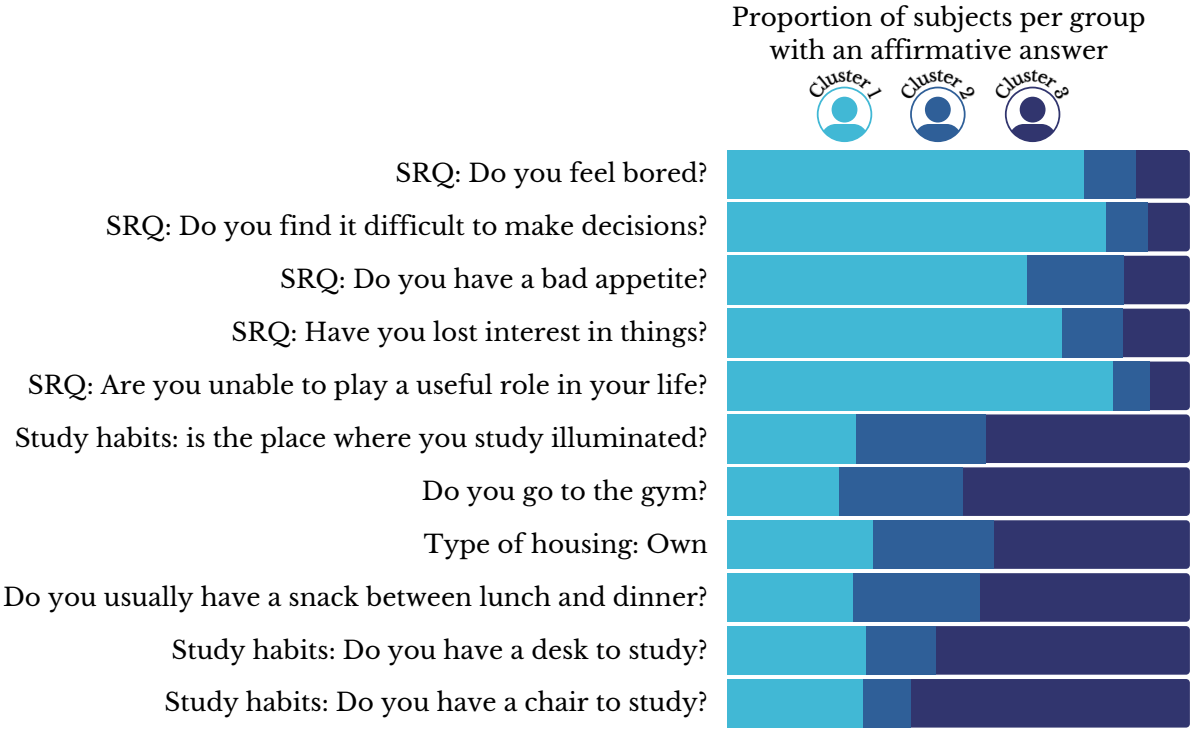

Figure 5.1: *Positive responses by group.*

(Database 2) and one of more than 9,000 (Database 1). This suggests that the sample obtained in Database 2 is a significant sample of Database 1. Additionally, the methodology proves to be robust for selecting the markers. Below, we expand on this analysis.

In the psychological aspects, compared with the SRQ markers obtained using Database 2 (Table 3.1), the new markers (Table 5.1) are the same, with slight variations in the order of relevance. There are differences in the presence of some aspects that impact life, such as online games or sex, and the use of marijuana, which is also absent this time. The academic aspects also show high similarities with Database 2, highlighting that study habits are the ones that most differentiate between the groups. In Database 1, the reasons for choosing the current university and career were not part of the markers.

The nutritional aspects increased their presence, focusing on those related to eating habits and the consumption of certain food groups (seeds, fruits, vegetables, and olive oil, among others). The socio-economic and socio-family aspects were no longer as relevant for forming groups (although they still contributed, to a lesser extent). Gender, a relevant variable in the selection of markers in Database 2 (although in the 60th position), is not relevant in this case in the selection of features. Finally, the initially absent aspects (sport, recreation, and culture) now contribute with four variables related to physical activities and spending free time.

Regarding the aspect related to the physical evaluation, none of these features functioned as marker for identifying the groups. In this case, the available physical variables did not dif-

fer between the groups, and no relationship was found with the other variables. This does not preclude the possibility that other physical or biological measures may provide relevant information on the health effects generated by direct or indirect exposure to situations of extreme experience related to the armed conflict.

Looking for individual variables of interest, in this case the features "victim" and "IDP" could not be distinguished in each group. Contrary to what is shown in Fig. 3.4(left) with Database 2, where two groups were directly related to these variables. However, the number of subjects who were victims or IDPs in Database 2 was proportionally higher (a ratio of 1:3 people considered victims or IDPs) than those processed in Database 1 (a ratio of 1:10 in this case). This does not mean that being a victim or an IDP does not have significant consequences, but rather that maybe the imbalance of the classes does not allow this to be reflected directly.

From the point of view of the application, these markers make it possible to find a profile of people exposed to armed conflict. They reveal aspects not usually accounted for on the field but opened the way to different analyses. For example, the least exposed students tend to have better eating habits than the other groups. This is an invitation to study these less common aspects and to develop strategies to improve the quality of life of people affected by exposure to conflict.

Finally, the validation with a supervised learning method confirms that the variables selected by the above procedure are discriminative. Therefore, in the absence of the EX2 test (the usual situation), the health authorities and the University Welbeing Unit can automatically generate a highly accurate alert on the students more exposed to the conflict and make necessary interventions. In addition, the correlation with other risks could help design training for this population. This can be complemented with the groups to which each subject would belong (clusters 1, 2, or 3) to identify a possible profile for each student. From a technical point of view, these results show that the proposed methodology is robust and the results are reliable. This work constitutes a tool for decision-making and monitoring of students for the University Wellbeing Unit of the University of Antioquia, which receives students from low socio-economic levels and who, due to the context of the country and the region, are more likely to be directly affected by the violence.

5.4 Conclusions

The most relevant aspects for identifying possible health consequences from exposure to the armed conflict were identified. Using a database with a smaller or larger number of variables did not show many differences in the aspects included in this selection. The students were divided into three groups with specific characteristics in both cases. As analyzed, the results found with the larger population showed a percentage-wise improvement in the ability to estimate the levels of exposure to the conflict. Therefore, 60 markers that allow such discrimination were determined (Table 5.1). In addition, these markers allow distinguishing a group whose activity is more related to negative aspects, another more related to positive aspects, and the last one

that behaves neutrally.

This predictor can determine the exposure level for a database with more information but without the EX2 test (for example, Database 1 or the full database of 33,561 records) if necessary.

5.5 Summary

This chapter presents a case study that finds 60 possible markers of exposure levels by applying a proposed methodology that uses unsupervised machine learning techniques. This methodology was initially implemented with a sample of 346 subjects, whose results are similar to those obtained by implementing a larger sample (9,346 subjects). Differences were obtained between three sets of subjects who shared similar characteristics regarding aspects that could be considered risky for early intervention or management of this problem. The supervised methodology made it possible to validate the selector markers and to define those that allow the higher differentiation of the groups.

Chapter 6

Conclusions and further research

THE aftermath of the armed conflict is a public health problem that has affected the Colombian population, both those directly affected by the conflict and the civilian population. The assessment methods traditionally used to study the consequences of exposure to conflict include techniques that are difficult to scale. For this reason, this research has proposed a novel methodology that uses an unsupervised technique to identify markers of exposure to conflict using clustering techniques. Here, we proposed to use the cluster centers to determine the relevance of the features in separating the subjects into different groups directly and indirectly related to the level of exposure to violence.

The proposed methodology was applied to two databases that differed mainly in the direct measure of the exposure level and the number of samples. Regarding the first aspect, only 346 students filled the EX2, a direct test that allowed evaluating exposure levels with a classification as "high" or "low." The methodology selected 62 markers related to the level of exposure. In addition, we found that the SRQ correlated with these exposure measures. On the other hand, we used a sample with a larger number of subjects that did not fill the EX2; however, based on the first result, the SRQ functioned as an indicator of the consequences of the armed conflict, and 60 markers with similar characteristics were found.

Before using the larger database, we faced the missing data problem. We developed a pipeline to determine the best percentage of missing data to impute or discard a variable with missing data. We found that applying a 10% threshold for accepting missing data in a given variable allowed better results to be obtained during the imputation stage. Furthermore, the imputation technique did not significantly influence this particular database. Despite working with mixed data, most of the techniques responded adequately.

From our proposal, psychological and academic aspects were the most influential in differentiating groups. Aspects usually associated with the aftermath of conflict, such as demographic and socioeconomic data, also contributed to separate the groups, but to a lesser extent. This methodology provides an overview of large amounts of variables not possible with traditional statistical methods. For example, we found that nutritional aspects were also related to the level of exposure to conflict. All of the above led to the division of the subjects into three groups (clusters), each with a characteristic behavior. In the first group, most were students with neg-

ative assessments, high school dropout risk, alcohol and drug use, and negative socioeconomic aspects. The second group presented positive responses in positive aspects, such as better nutrition, study habits, and activities that improve physical and mental health, such as playing sports and participating in cultural or recreational events. Finally, there was a third group in which neither positive nor negative aspects were emphasized.

Although some of these results could feel intuitive (e.g., a better socioeconomic situation with better nutrition), this methodology provides the evidence needed to consider them. A policymaker cannot make decisions just because some correlation looks feasible.

Implementing a supervised technique validated the effectiveness of the markers found with the different databases. It demonstrated that these markers effectively estimate exposure to extreme experiences in the context of the Colombian armed conflict with an accuracy of up to 80%. As we needed the EX2 as a label, the small number of samples compared to the number of variables could limit obtaining a higher precision. Nevertheless, the results showed consistency and an excellent approximation to study this population.

6.1 Future work

This research invites testing other types of clustering algorithms other than k-means, trying to avoid transforming the data into numerical values and proposing another way to identify markers in an unsupervised way. Additionally, developing a model that improves the accuracy of predicting exposure levels could allow working with the entire Well-Being database (which has more than 33,000 samples) to use the additional information available. Finally, complementing these results with more information from different biological measures would further highlight the effects on physical health.

Bibliography

[1] A. Campo-Arias, H. C. Oviedo, and E. Herazo, "Prevalencia de síntomas, posibles casos y trastornos mentales en víctimas del conflicto armado interno en situación de desplazamiento en colombia: una revisión sistemática," *Revista colombiana de psiquiatría*, vol. 43, pp. 177–185, 2014.

[2] S. Trujillo, N. Trujillo, S. Valencia, J. E. Ugarriza, and A. A. Mesas, "Executive and behavioral characterization of chronic exposure to armed conflict among war victims and veterans.," *Peace and Conflict: Journal of Peace Psychology*, vol. 25, p. 312, 2019.

[3] W. Tamayo-Agudelo and V. Bell, "Armed conflict and mental health in colombia," *BJPsych international*, vol. 16, pp. 40–42, 2019.

[4] M. Jawad, E. P. Vamos, M. Najim, B. Roberts, and C. Millett, "Impact of armed conflict on cardiovascular disease risk: a systematic review," *Heart*, vol. 105, pp. 1388–1394, 2019.

[5] M. A. Konstam and A. D. Konstam, "Gun violence and cardiovascular health: We need to know," *Circulation*, vol. 139, pp. 2499–2501, 2019.

[6] A. Quintero-Zea, L. M. Sepúlveda-Cano, M. R. Calvache, S. T. Orrego, N. T. Orrego, and J. D. López, "Characterization framework for ex-combatants based on EEG and behavioral features," in *Congreso Latinoamericano de Bioingeniería –CLAIB*, pp. 205–208, 2017.

[7] L. S. Giraldo, D. C. Aguirre-Acevedo, S. Trujillo, J. E. Ugarriza, and N. Trujillo, "Validation of the extreme experiences scale (ex2) for armed conflict contexts," *Psychiatric quarterly*, vol. 91, pp. 495–520, 2020.

[8] S. Trujillo, L. S. Giraldo, J. D. López, A. Acosta, and N. Trujillo, "Mental health outcomes in communities exposed to armed conflict experiences," *BMC psychology*, vol. 9, pp. 1–9, 2021.

[9] A. Quintero-Zea, J. D. López, K. Smith, N. Trujillo, M. A. Parra, and J. Escudero, "Phenotyping ex-combatants from eeg scalp connectivity," *IEEE Access*, vol. 6, pp. 55090–55098, 2018.

[10] S. P. Trujillo, S. Valencia, N. Trujillo, J. E. Ugarriza, M. V. Rodríguez, J. Rendón, D. A. Pineda, J. D. López, A. Ibañez, and M. A. Parra, "Atypical modulations of n170 component during emotional processing and their links to social behaviors in ex-combatants," *Frontiers in Human Neuroscience*, vol. 11, p. 244, 2017.

[11] D. A. Montoya, Ángela María Pareja, A. M. Valencia, C. M. Díaz, N. Trujillo, and D. A. Pineda, "Sistemas de activación y de inhibición de conducta y su relación con el funcionamiento ejecutivo en excombatientes irregulares del conflicto armado colombiano," *Medicina UPB*, vol. 39, pp. 2–10, 2020.

[12] C. Tobón, A. Ibañez, L. Velilla, J. Duque, J. Ochoa, N. Trujillo, J. Decety, and D. Pineda, "Emotional processing in colombian ex-combatants and its relationship with empathy and executive functions," *Social neuroscience*, vol. 10, pp. 153–165, 2015.

[13] F. C. Pereira and S. S. Borysov, *Machine Learning Fundamentals*, pp. 9–29. Elsevier, 2019.

[14] D. E. Goin, K. E. Rudolph, and J. Ahern, "Predictors of firearm violence in urban communities: a machine-learning approach," *Health & place*, vol. 51, pp. 61–67, 2018.

[15] H. Santamaría-García, S. Baez, D. M. Aponte-Canencio, G. O. Pasciarello, P. A. Donnelly-Kehoe, G. Maggiotti, D. Matallana, E. Hesse, A. Neely, and J. G. Zapata, "Uncovering social-contextual and individual mental health factors associated with violence via computational inference," *Patterns*, vol. 2, p. 100176, 2021.

[16] C. C. Aggarwal *et al.*, *Data mining: the textbook*, vol. 1. Springer, 2015.

[17] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, pp. 651–666, 2010.

[18] M. Khanum, T. Mahboob, W. Imtiaz, H. A. Ghafoor, and R. Sehar, "A survey on unsupervised machine learning algorithms for automation, classification and maintenance," *International Journal of Computer Applications*, vol. 119, no. 13, 2015.

[19] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," *Supervised and unsupervised learning for data science*, pp. 3–21, 2020.

[20] A. Ahmad and S. S. Khan, "Survey of state-of-the-art mixed data clustering algorithms," *Ieee Access*, vol. 7, pp. 31883–31902, 2019.

[21] M. S. Santos, R. C. Pereira, A. F. Costa, J. P. Soares, J. Santos, and P. H. Abreu, "Generating synthetic missing data: A review by missing mechanism," *IEEE Access*, vol. 7, pp. 11651–11667, 2019.

[22] I. Adjerid and K. Kelley, "Big data in psychology: A framework for research advancement.," *American Psychologist*, vol. 73, p. 899, 2018.

[23] M. I. Cano, C. Isaza, A. Sucerquia, N. Trujillo, and J. D. López, "Markers of exposure to the colombian armed conflict: A machine learning approach," in *Advances in Artificial Intelligence–IBERAMIA 2022: 17th Ibero-American Conference on AI, Cartagena de Indias, Colombia, November 23–25, 2022, Proceedings*, pp. 185–195, Springer, 2023.

[24] U. de Antioquia, "Caracterización de estudiantes."

[25] C. Echandía Castilla and L. Salas, *Dinámica Espacial de las Muertes Violentas en Colombia (1990-2005)*. 07 2008.

[26] V. Mestre Escrivá, M. D. Frías Navarro, and P. Samper García, "La medida de la empatía: análisis del interpersonal reactivity index," *Psicothema (Oviedo)*, pp. 255–260, 2004.

[27] M. A. Garcia-Barrera, J. E. Karr, N. Trujillo-Orrego, S. Trujillo-Orrego, and D. A. Pineda, "Evaluating empathy in colombian ex-combatants: Examination of the internal structure of the interpersonal reactivity index (iri) in spanish.," *Psychological assessment*, vol. 29, no. 1, p. 116, 2017.

[28] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[29] V. Bell, F. Méndez, C. Martínez, P. P. Palma, and M. Bosch, "Characteristics of the colombian armed conflict and the mental health of civilians living in active conflict zones," *Conflict and health*, vol. 6, pp. 1–8, 2012.

[30] T. D. Little, T. D. Jorgensen, K. M. Lang, and E. W. G. Moore, "On the joys of missing data," *Journal of pediatric psychology*, vol. 39, pp. 151–162, 2014.

[31] W.-C. Lin and C.-F. Tsai, "Missing value imputation: a review and analysis of the literature (2006–2017)," *Artificial Intelligence Review*, vol. 53, pp. 1487–1509, 2020.

[32] R. J. Little, D. B. Rubin, and S. Z. Zangeneh, "Conditions for ignoring the missing-data mechanism in likelihood inferences for parameter subsets," *Journal of the American Statistical Association*, vol. 112, pp. 314–320, 2017.

[33] S. V. Buuren, *Flexible imputation of missing data*. CRC press, 2018.

[34] M. D. Samad, S. Abrar, and N. Diawara, "Missing value estimation using clustering and deep learning within multiple imputation framework," *Knowledge-based systems*, vol. 249, p. 108968, 2022.

[35] M. Garcia-Peña, S. Arciniegas-Alarcón, and W. J. Krzanowski, "Missing value imputation using least squares techniques in contaminated matrices," *MethodsX*, vol. 9, p. 101683, 2022.

[36] B. Huang, Y. Zhu, M. Usman, and H. Chen, "Semi-supervised learning with missing values imputation," *arXiv preprint arXiv:2106.01708*, 2021.

[37] A. Yazdani and A. Yazdan, "Using statistical techniques and replication samples for imputation of metabolite missing values," *arXiv preprint arXiv:1905.04620*, 2019.

[38] G. Molenberghs and G. Verbeke, "Multiple imputation and the expectation-maximization algorithm," *Models for discrete longitudinal data*, pp. 511–529, 2005.

[39] C. Karras, A. Karras, M. Avlonitis, and S. Sioutas, "An overview of mcmc methods: from theory to applications," in *Artificial Intelligence Applications and Innovations. AIAI 2022 IFIP WG 12.5 International Workshops: MHDW 2022, 5G-PINE 2022, AIBMG 2022, ML@ HC 2022, and AIBEI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings*, pp. 319–332, Springer, 2022.

[40] S. Zhang, J. Zhang, X. Zhu, Y. Qin, and C. Zhang, "Missing value imputation based on data clustering," *Transactions on computational science I*, pp. 128–138, 2008.

[41] Y. Ding and A. Ross, "A comparison of imputation methods for handling missing scores in biometric fusion," *Pattern Recognition*, vol. 45, no. 3, pp. 919–933, 2012.

[42] S. Zhang, "Nearest neighbor selection for iteratively knn imputation," *Journal of Systems and Software*, vol. 85, no. 11, pp. 2541–2552, 2012.

[43] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," *Ensemble machine learning: Methods and applications*, pp. 157–175, 2012.

[44] V. K. Ojha, A. Abraham, and V. Snášel, "Metaheuristic design of feedforward neural networks: A review of two decades of research," *Engineering Applications of Artificial Intelligence*, vol. 60, pp. 97–116, 2017.