July 2021

Artificial Intelligence to Prevent the Assassination of Human Rights Defenders

THE GENEVA CHALLENGE 2O21

Kevin Steven Mojica Muñoz - Paula Alejandra Algarra Saavedra - Julian Peña Duarte

Colombia

The images and information provided on the cover page are taken from the Front Line Defenders website: https://www.frontlinedefenders.org/en/cases/violation/killing
We send special thanks to all the human rights defenders who daily inspire us to achieve a better world.

Abstract

Since 2015, at least 1,323 human rights defenders have been assassinated worldwide, with a growing tendency in recent years. Most of these killings occur in contexts of drug trafficking and armed conflict, concentrated in some Latin American and Asian countries, and shaping a crisis with devastating social effects. The situation has become critical since countries have difficulty targeting and assigning protective measures to prevent the murder of the defenders. In the context of this problem, this paper proposes the use of algorithms based on artificial intelligence as a useful tool to streamline the processes of targeting and prioritization of cases so that the allocation of protection is more efficient and accurate. To demonstrate that this is possible, the text lays out an application to the Colombian case, detailing the technical and data requirements that must be overcome to implement this scheme. This application project seeks to be a model to be used, not only in the Colombian case but in all possible countries with similar characteristics.

Kevin Steven Mojica Muñoz.
Born in Colombia, Kevin studied Government and Economics undergraduate programs at Universidad de los Andes and is currently finishing his master's degree program in both Economics and Public Policy. His main fields of work are public policy analysis, impact evaluation, and machine learning. He is currently a teaching assistant at Universidad de los Andes and a consultant on public policy issues. Kevin has a clear purpose in life, to improve the living standards of vulnerable populations through public policy solutions.

Paula Alejandra Algarra Saavedra.
Born in Colombia, Paula is convinced of the importance of love to transform lives. She is currently pursuing her master's in public policy at Universidad de los Andes, and she works at the Colombia's National Planning Department. Before her master's, Paula completed her bachelor's degree in economics at Universidad del Rosario, and she participated in projects of the UNDP and the Cellular Agriculture Society. She has experience in public investment, social development, and innovation. Paula hopes to make her country a better place to live.

Julian Peña Duarte.
Born in Colombia, Julian completed his bachelor's degree in economics at Universidad del Rosario in 2020. Currently, he is pursuing a master's degree in economics, and he works as Research Assistant at Universidad del Rosario. He has experience in areas of education, health, impact evaluation, and behavioral economics. Julian firmly believes that education is the key to achieve a better society; for this reason, he wants to contribute to the community by developing and promoting social programs that allow everyone accesses to quality education.

" (Aggressors) are often intended to intimidate, silence and stop human rights defenders from carrying out their work. There is no more direct attack on civil society space than the killing of human rights defenders"

(Lawlor, M., Declaration at DW of The U.N. Special Rapporteur on the situation of human rights defenders, 2021)

"The level of danger facing activists worldwide has reached crisis point. Every day ordinary people are threatened, tortured, imprisoned and killed for what they fight for or simply for who they are. Now is the time to act and tackle the global surge in repression of human rights defenders"

(Naidoo, K., Declaration at the Human Rights Defenders World Summit of The Secretary General of Amnesty International, 2018)

"*Alarming numbers of human rights defenders and journalists continue to be intimidated, attacked and killed, particularly those dedicated to protecting the environment and land rights. (...) I encourage decisive investigations and prosecutions of perpetrators*"

(Bachelet, M., Statement at the 45th session of the Human Rights Council of The UN High Commissioner for Human Rights, 2020)

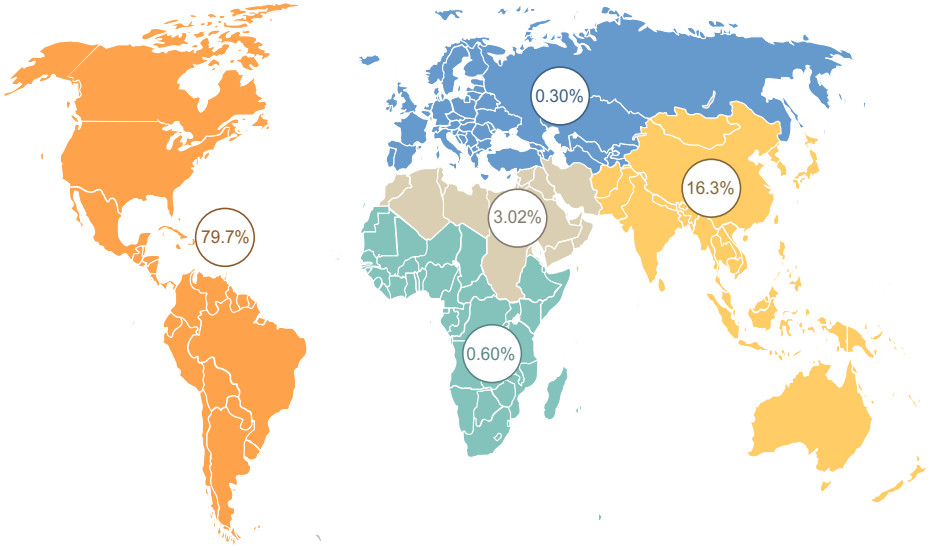
1. Problem identification: why is the killing of human rights defenders a social crisis?

According to The Armed Conflict Location & Event Data Project (2021), since 2016, at least 62,819 events of violence have occurred in which armed groups deliberately target unarmed civilians, generating at least one fatality. This is equivalent to 1.37 violent events per hour from 2016 until records are available for the year 2021. Most of these acts of violence occur in the context of armed conflicts or concerning illegal markets, such as drugs, human trafficking, arms trafficking, among others. Of all the types of violence against civilians, the assassination of human rights defenders is one of those with the most negative effects on society. According to The United Nations Human Rights Council Special Rapporteur (2021), at least 1,323 human rights defenders from 64 countries have been assassinated since 2015, with a growing tendency in recent years. In general, data collectors agree that underreporting is a common problem and that assassinations are fueled by "widespread impunity".

A human rights defender is defined as a person that assumes peaceful leadership in defense and surveillance of human rights (The UN General Assembly, 1998). It is important to highlight that this concerns environmental, ethnic, gender, sexuality, and even community representation issues. Despite the support that these agents of society have provided in their communities, human rights defenders around the world are targets of threats, assassinations, and other forms of violence. In Europe, Central Asia, Middle East, North Africa, and Asia-Pacific, human rights defenders are repressed by legal action; while in Africa and America, they are physically attacked, as reported by Front Line Defenders (2020). The main reason why human rights defenders are murdered in the world is because of their role in society. The human rights defenders are voices leading change in their communities, which makes them vulnerable to armed actors who want to secure their interests (Prem, Et.al., 2018).

Though the assassination of human rights defenders is a global problem, for some regions it is a crisis with remarkable social effects. Indeed, in 2020, 79.75% of the murders of human rights defenders occurred on the American continent, 16.31% in Asia and The Pacific, 3.02% in the Middle East and North Africa, 0.60% in Sub-Saharan Africa, and 0.30% in Europe and Central Asia (Front Line Defenders, 2020). Regarding the region most affected by this social problem, the Latin American and Caribbean region recorded the highest number of murdered defenders between 2015 and 2019, with 933 murders (UN Human Rights Council Special Rapporteur, 2021). Currently, countries such as Colombia, the Philippines, Honduras, Mexico, Afghanistan, Brazil, Guatemala, Iraq, Peru, and India rank as the ten countries with the highest number of murders of human rights defenders (Front Line Defenders, 2020).

Figure 1. The Killing of Human Rights Defenders Around the World by Region



Source: Constructed using data from Front Line Defenders (2021).

The killing of Human Rights Defenders has profound negative impacts on society and peace consolidation. According to USAID, CODHES, and CNC (2019), after the murder of a defender, there is a deterioration of the social fabric of the communities and the channels they have to demand their rights. When aggressions occur, people are afraid to participate, to organize themselves, their will diminishes, and the formation of new leadership becomes more difficult. This implies that the assassination of a human rights defender not only affects his or her own social agenda, but also the leadership initiatives of other leaders in their communities, which ends up silencing the entire community. This is the final purpose of armed

groups since they can boost their agendas without the interference of the defenders.

The causes of the assassination of human rights defenders depend on the country's context. In general, killed defenders try to defend their communities against injustices imposed by a system, a project (illegal or legal), armed groups, or a government group that has greater power. The killings of human rights defenders occur in contexts of structural violence and inequality (UN Human Rights Council Special Rapporteur, 2021). Indeed, most of the assassinated human rights defenders were in remote rural areas occupied by irregular armed groups (Lanteró, 2017), with little state presence (Prem et al., 2018; Uribe et al., 2020), and with an inefficient local judiciary (Prem et al., 2018). Furthermore, human rights defenders face travel restrictions and limited access to information, making them more vulnerable to any attack (Front Line Defenders, 2020). In this scenario, the governments face difficulties to offer real guarantees for the protection of the defenders. Developing countries, especially, do not have the capacity to generate effective responses for protection demands; largely because the budget is limited and represents an obstacle to ensuring an effective service (UN Human Rights Council Special Rapporteur, 2021).

Many governments around the world have implemented different schemes for the prevention of the killing of human rights defenders, with varying degrees of success. According to the UN Human Rights Council Special Rapporteur (2021), countries like Brazil, Burkina Faso, Colombia, Côte d'Ivoire, The Democratic Republic of the Congo, Guatemala, Honduras, Mali, Mexico, and Perú have developed different schemes and public policies to protect human rights defenders. However, some of these schemes and policies are often under-resourced, and their performance is criticized by defenders (UN Human Rights Council Special Rapporteur, 2021). Some countries have a reduced number of protection schemes to assign in the most critical cases; the problem is that sometimes these cases are not the ones that require an urgent response, and countries, like Colombia, present delays in the selection and risk evaluation process (Human Rights Watch, 2021).

This is exactly the part of the problem that this project intends to impact. If governments can easily identify the most at-risk human rights defenders, then the protection schemes will be assigned to the most critical cases and prevent the eventual murder of the defenders. In this scenario, Artificial Intelligence could play an important role. Some Machine Learning Algorithms have demonstrated their ability to accurately make predictions in other contexts such as crime, corruption, and murders (Aarvik, 2019; Alves et al., 2018; King et al., 2020; McClendon et al., 2015). To apply these methods to the prevention of the killing of human rights defenders, some ethical and technical concerns will probably need to be overcome, but it is an alternative that should be considered in the measures being taken to prevent this problem.

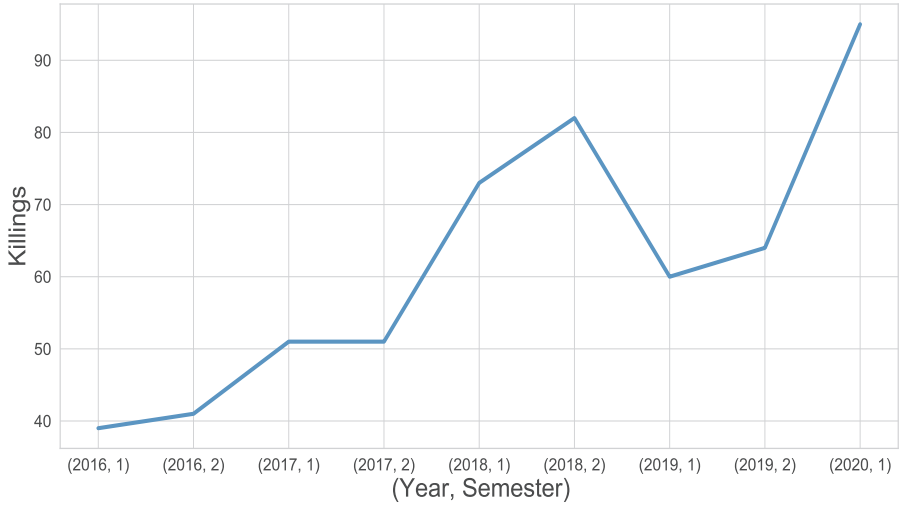
To illustrate how Artificial Intelligence can improve national protection schemes for human rights defenders, this project focuses on the Colombian case, one of the countries with the highest number of assassinated human rights defenders in the world (Human Rights Watch, 2021). The paper lays out the technical basis of an application of artificial intelligence for the prevention of the murder of human rights defenders, detailing the requirements in terms of data and the algorithms that could be used in each case. This application, although specific to Colombia, can be replicated in other developing countries with similar conditions.

2. The particular case of Colombia

Since 2016, more than 400 human rights defenders have been assassinated in Colombia, the highest number in Latin America according to the United Nations Office of the High Commissioner for Human Rights (2021). The increase in the killing of Human Rights defenders started in 2016 (See Figure 2). During this year, the government signed a peace agreement with the former FARC guerrilla after 50 years of an internal armed conflict. Although the Peace Agreement drastically reduced violence, the killing of human rights defenders has increased each year as other armed groups have stepped into the gap left by FARC guerrillas (Human Rights Watch, 2021). Armed groups oppress defenders to use them in the imposition of "rules" within communities. This increases the possibility of groups targeting them for actual or perceived non-compliance or alleged support of an opposing party (Human Rights Watch, 2021).

In addition, human rights defenders in Colombia face other kinds of abuse. According to the Human Rights Ombudsperson's Office (2021), 2,829 threats against human rights defenders were registered between January 2016 and June 2020, most of them

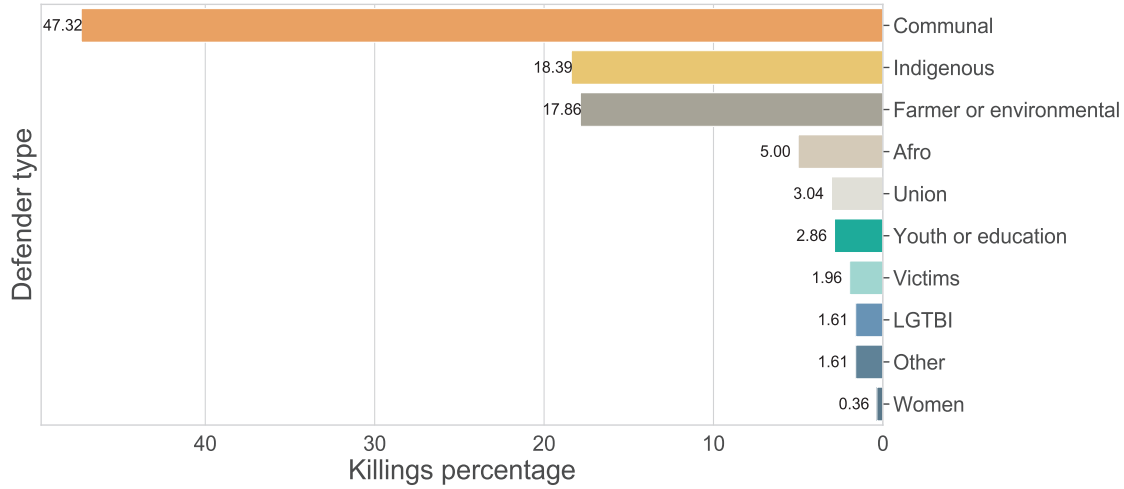
Figure 2. the Killing of Human Rights Defenders in Colombia (January 2016 - June 2020)



Source: Constructed using We are Defenders Reports (2016; 2017; 2018; 2019; 2020).

Source: Constructed using We are Defenders Reports (2016; 2017; 2018; 2019; 2020).

Figure 3. Categories of assassinated human rights defenders by percentage in Colombia (January 2016- June 2020)



were death threats (As cited in Human Rights Watch, 2021). As reported by the non-governmental organization We are Defenders, the main categories of assassinated defenders are communal, indigenous, and environmental defenders (See Figure 3).

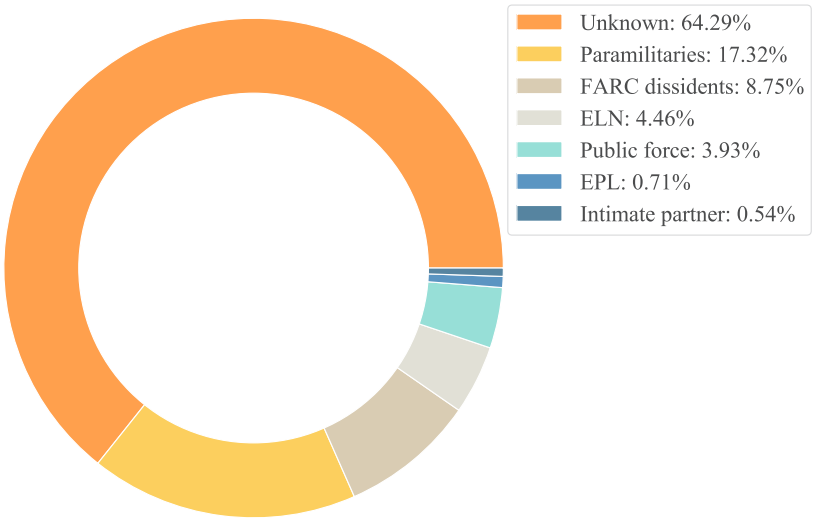
There are some common patterns shared by the areas in which the murders take place. According to Human Rights Watch (2021), of all the assassinations of human rights defenders occurring between 2016 and December 2020:

- 100 percent occurred in municipalities with poverty levels above the national average (measured based on a government "multidimensional poverty index").
- 98 percent occurred in municipalities where armed groups operate, which includes organized crime groups and parties to the armed conflicts.
- 97 percent occurred in municipalities with illegal economies, which includes drug trafficking and production, illegal mining, extortion, illegal logging, and smuggling.
- 92 percent occurred in municipalities with drug trafficking and production.
- 91 percent occurred in municipalities with murder rates of over 10 per 100,000 people (what the World Health Organization considers the threshold for "endemic violence").
- 70 percent occurred in rural areas (only 9 percent in Colombia's 13 "main cities").
- 57 percent occurred in municipalities considered by the government in 2017 as the "zones most affected by the armed conflict".
- 52 percent occurred in municipalities where the government has announced "Territorial Development Programs[1]" (PDET).
- 47 percent occurred in municipalities with illegal mining.

1. The Territorial Development Programs are a development initiative created by the peace agreement with the FARC for areas highly affected by the armed conflict, poverty, lack of state presence, and illegal economies.

These characteristics show that the murder of human rights defenders is concentrated in specific areas with low state presence and monopoly of force. This results in impunity for most of these crimes (Human Rights Watch, 2021), often even without information on the perpetrators. According to We are Defenders Reports (2016; 2017; 2018; 2019; 2020), from January 2016 to June 2020, in 64.29% of human rights assassinations, it was not possible to identify the perpetrator of the crime (See figure 4). The weakness of the state's capacity in these areas (Prem, Et al., 2018) to identify, investigate, capture, and effectively prosecute the perpetrators is probably one of the reasons why the phenomenon has not diminished in recent years.

Figure 4. The alleged responsible for the killings of human rights defenders by percentage in January 2016- June 2020



Source: Constructed using We are Defenders Reports (2016; 2017; 2018; 2019; 2020).

Other causes that could influence the murders of human rights defenders in Colombia are some policies stemming from the peace agreement with the FARC (Garzón-Vergara, 2015; Defensores et al., 2018; Uribe, Et al., 2020), such as the development of government programs for the substitution of illicit crops (Gutierrez, et al., 2020; Uribe, et al., 2020) and land restitution (Uribe, et al., 2020); as well as some struggles for territorial control by armed groups derived from the peace agreement (Kalyvas, 2006; Prem, Et al., 2018). It is also attributed to state abandonment (Prem, Et al., 2018; Uribe, Et al., 2020), poor definition of land property rights (Prem, Et al., 2018), illegal mining (Le Billon, 2020; Witness, 2020), poverty (Le Billon, 2020), active political participation (Lanteró, 2017), the presence of corruption (Albarracín et al., 2020), and the deterioration of the attention and visibility of the situation of human rights defenders in the public debate (Uribe, Et al., 2020).

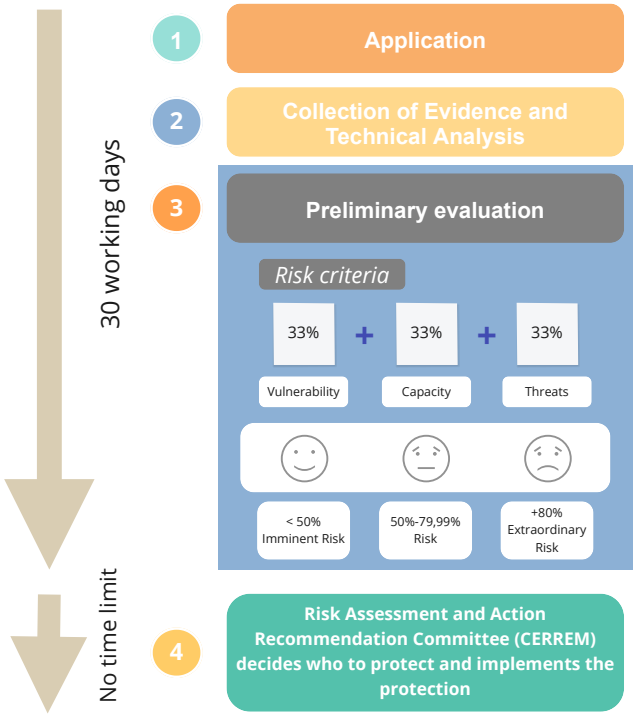
3. Protecting human rights defenders in Colombia: The Role of the National Protection Unit

To avoid the assassination of human rights defenders, some governments around the world have implemented mechanisms to guarantee their protection. In general, the mechanisms are laws, action policies, or protection programs, which are established coordinately with national and international human rights organizations under the regulatory framework of the UN Declaration on Human Rights Defenders (Quintana et al., 2011). In the Colombian case, the government created the National Protection Unit (UNP) by decree 4065 of 2011 to "organize, coordinate and implement the provision of opportune, efficient, and appropriate protection services to such persons as the National Government considers require them as a result of their activities, condition, or situation, which face extraordinary or extreme risk to life, integrity, liberty, and personal security" (Ministry of Interior, 2015).

To guarantee the fulfillment of the institution's mission, some mechanisms were established for the allocation of protection. Currently, the protection assignment process in Colombia has the following stages: (1) application; (2) collection of evidence and technical analysis; (3) preliminary risk evaluation through capacity, vulnerability, and threats criteria; (4) decision regarding protection and protection implementation (Lanteró, 2017). The first stage, the application, is made at the request of the interested party, who is the human rights defender in this case. There are three types of requests that a person can make: request for collective protection, request for individual protection, and requests about the implemented measures -corresponding to adjustments of the protection scheme-. In all situations, applicants must fill out a form and provide some additional documents that must be delivered in person at a National Protection Unit office (National Protection Unit, n.d.).

It is important to point out that the number of requests made is considerable; in 2020, the unit received over 31,000 requests for such schemes, at least 11,000 from individuals the unit considers to be human rights defenders (Human Rights Watch, 2021). The unit implemented approximately 1,600 such measures, although it is unclear how many people were able to benefit from them. In 2019 it was a similar scenario, 1,900 human rights defenders received protection schemes out of 13,000 who requested them; in both cases, the remaining requests were denied (Human Rights Watch, 2021).

Figure 4. The Protection assignment process in Colombia



Source: Constructed using information from National Protection Unit

The second stage of the process -The collection of evidence and technical analysis- oversees the National Protection Unit and the National Police of Colombia, who together collect on-site evidence so that it can be used as input in the third stage, -preliminary risk evaluation- (Lanteró, 2017). In the third stage, a preliminary risk assessment is made based on the following criteria: 1) The threats[2] faced by an individual or group of individuals (Threats); 2) The vulnerability[3] to these threats (Vulnerability); and 3) the resources available for a person to reduce the vulnerability (Capacity) (Lanteró, 2017). These are evaluated from the information provided in the application forms by a group of experts[4] and are assigned equal importance (33.33%) (Lanteró, 2017).

The level of risk is defined by the result of the total sum of the percentages obtained in each of the criteria (threats, vulnerability, and capacity). If the total score is between 50% and 79.99%, the person is classified in "Risk". On the other hand, if the total score is between 80% and 100%, the person is in "Extraordinary Risk", and he/she becomes the object of protection measures (Camilo Torres, personal communication, May 16, 2017). Finally, if the person does not classify in any of the previous categories, the person is cataloged as "Imminent risk" (Lanteró, 2017).

Later, in the fourth stage -decision regarding protection and protection implementation-, a risk assessment report is sent to The Risk Assessment and Action Recommendation Committee (CERREM), which decides to whom and what type of protection is given (Lanteró, 2017). However, this last stage does not have an established time limit, so the protection assignment processes can take months. In addition, despite the 30 days rule to generate the risk assessment report, in December 2020, the National Protection Unit told Human Rights Watch that, in May 2020, such risk analysis was carried out on average 160 days after the legal deadline, and in December, on average 101 days after the legal deadline (Human Rights Watch, 2021). Regarding these delays, some National Protection Unit officials have pointed out that "the analysis team works in a context where the number of requests is very high, and the staff is extremely limited, even with the assistance of extra personnel made up of 33 police officers and several private contractors to assist the National Protection Unit" (Camilo Torres, personal communication, May 16, 2017).

Another factor that contributes to the deficient protection of human rights defenders through the governmental mechanism is the budgetary constraints that the National Protection Unit faces. In 2019, its budget amounted to COP 688.747.241.558 (roughly US $209 million) (National Protection Unit, 2020) of which roughly half was used to protect government

2 Threat represents the possibility to physically, materially, or morally harm someone or their property by an intentional, and usually violent, action (Lanteró, 2017).

3 Vulnerability refers to the extent to which an individual is affected by the threats directed against him or her (Lanteró, 2017).

4 The experts take into account: i) Whether the defender performs fieldwork in a conflict or isolated area with minimal state presence; ii) Access to materials and elements like safe transport, the installation of security cameras, and communications; and iii) The subjective perception of risk: this includes the way a person handles fear and how threats psychologically affect him/her at work.

authorities. The unit spent COP 200 billion (roughly US $61 million) more than its original budget for that year, using funds assigned to its 2020 budget (Human Rights Watch, 2021). A similar situation was faced in the 2020 budget (National Protection Unit, 2020).

In conclusion, despite the meticulous definition of regulatory parameters, activities, and institutions, the National Protection Unit faces inefficiencies in the prioritization and allocation of effective protection measures for human rights defenders. These inefficiencies are mostly explained by two factors: the first one, there is a budget constraint that limits the application of protection measures to only a small proportion of all human rights defenders at risk. And the second one, the process to identify the level of risk, is highly inefficient which is evidenced by delays in the assignment process. Moreover, the risk level is defined using a basic percentage sum methodology.

4. Artificial intelligence to prevent the murder of human rights defenders

In order to reduce the assassination of human rights defenders, the Colombian government should focus on the development of better systems for the allocation and focusing of protective measures. This diagnosis coincides with the one done by the National Protection Unit (UNP) Director, who mentioned the next affirmations in a recent interview:

"Today the UNP is an entity that must be more agile, we need to meet the goal set by the President of the Republic, which is to reduce the time of the protection route, and in this effort, we are reengineering the entity to prioritize the application of schemes to those who really need it."
(Palacios, D., Director of UNP interview for Semana, 2020).

"We also need a faster debugging system. Today the UNP receives more than 4 thousand requests for protection per month, and there we need to have the ability to differentiate who is not subject to protection and who is not at risk faster to give priority to those who do need it urgently"
(Palacios, D., Director of UNP interview for Semana, 2020).

This technical problem is where artificial intelligence could contribute the most, not only in the Colombian case but for all other countries with similar conditions. In this scenario, there is a mechanical task to determine the risk of human rights defenders. The risk is defined as the probability that the defender will face any damage to their physical integrity due to the activities related to the protection of human rights. Therefore, the technical problem to be solved by Artificial Intelligence is to identify whether a defender will suffer any physical aggression such as homicides in the future, based on the available contextual variables.

The diagnosis of the problem shows that the assassinated defenders share some observable common characteristics. This result suggests the existence of common underlying patterns that influence the risk of a potential murder; these patterns can be inferred through machine learning algorithms and generate a new risk estimate with a lower error. The next equation represents the relation between the unobservable real risk and estimations:

$$\underset{\text{Real Risk}}{Y_i} = \underset{\text{Estimate}}{f(X_i)} + \underset{\text{Error}}{\varepsilon_i}$$

In this equation, the real risk of assassination is equal to the sum of the risk estimate and the calculation error. The risk estimate is a function that takes the set of contextual variables and converts them into an approximate measure of the probability of murder. As this function is able to decipher with greater plausibility the underlying patterns in the data that influence the risk of murder, then it is possible to have measures with a lower error that serve to generate better guidance for the allocation of protection measures. The final purpose is to obtain the optimal function that minimizes the error by using machine learning so that the estimate is as close as possible to the real value of risk.

Two general approaches are considered to achieve this purpose, supervised machine learning and unsupervised machine learning. In supervised machine learning, the ultimate objective is the direct prediction of an eventual murder. In this case, the algorithm seeks to minimize the prediction error of a target variable (a dummy variable) that establishes whether the human rights defender is murdered or not from a set of contextual variables. The algorithm is trained based on information from past cases, previously assassinated leaders about whom information is available, and is applied to predict the probability of assassination of defenders today.

The second one is unsupervised machine learning, an approach that studies the distribution of contextual variables to identify patterns of interest. Unlike the first method, this one does not require a target variable, but it needs a set of meticulously selected contextual variables. In this case, the value of the algorithms lies in their ability to identify these underlying patterns in the data that may indicate a higher risk of murder. Many approaches could be considered to achieve this objective, but, for this particular exercise, we propose a multivariate clustering methodology to partition the observations into the characteristics that are deterministic of the murder of human rights defenders.

This paper aims to provide a general guide on how these different approaches could be applied to the Colombian case so that their application can be useful to improve the targeting of protection measures given the budget constraints faced. Nevertheless, it is important to note that this application can be extrapolated to other countries with similar conditions.

4.1 Data requirements

Before going into the methodological details of the two approaches presented above, it is useful to consider the data requirements to perform the task. In practice, the availability and quality of data are some of the main challenges in developing these types of methods for risk estimation. For supervised machine learning, this requires having a target variable to predict and a set of contextual variables from which to infer that prediction. If the contextual variables are not deterministic of the target variable, or if the target variable does not adequately express the researcher's interest, then the exercise would not be successful. For unsupervised learning, contextual variables should inform relationships or patterns of interest to the researcher. In this case, if the variables are not relevant or have significant measurement errors, then the outcome of the exercise will be unsatisfactory too.

In some countries, the idea of carrying out censuses of human defenders has been promoted to record their characteristics and generate better risk measures for the allocation of protection; nevertheless, in Colombia, this is not yet possible, so the only available records are the data collected by the National Protection Unit in the first and second stage of the protection assignment process.

The variables are collected through the protection application form for individual protection and some complementary documents. The registration form for the individual protection program of the National Protection Unit includes information that allows the individual to be characterized at the time in which he/she denounces the risk faced (National Protection Unit, 2019), these variables are: the person's name, birth location (country, department, and city), the home location (country, department, city, township, sidewalk, neighborhood, address), a dummy variable that identifies if the person lives in a rural or urban area, some contact information of the person like his/her telephone number and e-mail, information about the biological sex, gender identification, sexual orientation and age group.

It also identifies if the person requesting protection belongs to an ethnic group, has disabilities, performs care-taking tasks in their home, belongs to an organization that defends human rights or social representation; it distinguishes the type of subject that requires protection[5], the person's subjective perception of risk, the threat and who did it, the role of the person in the organization, and the organization name (if applicable). In addition, the National Protection Unit registers to which of the protective measures was assigned and of what type, in other words, if the person was granted armored vans, weapons, personnel trained to provide protection, instruction, etc[6].

The above data is linked to the *identification number* provided by the person in the complementary documents of the application. Given this, the information collected by the National Protection Unit can be combined with data available in other public entities if these datasets contain a variable with the person's *identification number* that allows linking the information from the different sources. Similarly, the data on protection requests from the National Protection Unit can also be complemented with municipal characteristics due to the location information since a variable for municipality code can be included to allow linking of the information with other databases.

The observation unit will be the human rights defen-

5 Some of the subjects that are distinguished are: Person of political opposition, armed conflict victim, Union, guild, ethnic, medic, witness, journalist, land claimant, public workers, persons related with peace policies and agreements, teachers or demobilized from armed groups.

6 For the complete list of variables collected by the UNP, consult the data dictionary available at: https://drive.google.com/file/d/1r6TOHHKPg8YUlXBd9-iAwvX0SRHSdWcS/view?usp=sharing

der on the day of the year in which they made their request for protection. It should be noted that the number of protection applications available in the UNP databases (see Section 3) provides a considerable number of observations that make the development of the proposal feasible.

4.2 Supervised Machine Learning

4.2.1 Definition of the target variable and predictors

Because the actual risk of murder is a variable that can only be identified in the future, training of supervised machine learning algorithms can only be done from past application data; consequently, the assassinated defenders that had requested protection before the murder are the main source of information for training algorithms. However, since there is no record of assassinations for individuals who made a request for protection and were not granted protection measures by the National Protection Unit, it is necessary to construct this target variable. That is why the data on protection requests for human rights defenders from the National Protection Unit and the data from the National Institute of Legal Medicine and Forensic Sciences for people assassinated in Colombia will be used.

Both data sets have a common variable that allows the identification of the person (*Identification number*). Therefore, a conditional function will be used to create a new variable that will distinguish which of the human rights defenders who made a protection request were killed. Thus, when the identification number *n* is found in the database of the National Protection Unit and the dataset for assassinations of the National Institute of Legal Medicine and Forensic Sciences, the new variable *Dummy for assassination* will take value one, and otherwise it will take value zero.

That target variable (*Dummy for assassination*) will allow us to predict the risk of murder in the future based on a set of predictor variables. Therefore, the possible predictors are as follows[7]: the human rights defender home location, rural/urban area of living, sex, gender identification, sexual orientation, age group, ethnicity, presence of disabilities, if he/she does care-taking work at home, human rights organization to which he/she belongs, his/her role in the organization, the activity he/she was engaged in when he was threatened, possible aggressors, type of threat, who made the threat, and the person's subjective perception of risk.

Additionally, relevant available municipal information can be included, which in light of the literature explains the level of risk faced by a human rights defender, such as the following: the number of hectares of coca, the presence of armed groups, the homicide rate, the level of poverty, the administrative capacity, the presence of illegal economies, the distance to the capital of the department and the rural rate of the municipality.

4.2.2 Methodology

In supervised Machine Learning, the final objective is to minimize a chosen loss function, which in this case would be the prediction error of a dummy variable that indicates whether a defender was assassinated based on the contextual variables available. A typical exercise of this type involves the training, evaluation, and application of machine learning methods that seek to infer underlying patterns in contextual data that indicate an outcome in the target variable. The following is an indicative outline of how the process should be carried out for the specific case:

Stage 1. Preparation: This phase includes the processes of i) data collection, ii) selection of predictor variables[8], iii) treatment of missing values[9], and iv) scaling of variables if necessary[10]. In this phase it is usual to divide the database into training and test subsets; normally, this division is 70% training, 30% test, but it is always necessary to check whether the training and test sub-samples are comparable since randomization may not always work.

7 It is also possible to create composite indices of the mentioned variables to test if the performance of models improves.

8 This step can be performed through different variable selection methods that can be found by default in packages such as *Fselector*, *boruta*, *VSURF*, etc. However, it can also be done manually using cross-validation (integrating with the training algorithms to be used). This second option can be much more complex to develop.

9 This step can be done by deleting observations with missing data, although this implies that these observations do not share any specific characteristic. A better option is to consider imputation methods such as *Missing Forest*, this method developed by Stekhoven & Buhlmann (2011) uses the Random Forest machine learning algorithm to fill in the missing values in the data.

10 This depends on the algorithm to be used, some of them require the predictor variables to be scaled to have a favorable performance, such as k-nearest neighbors algorithm, regression, etc.

Additionally, it is necessary to verify that the target variable, if dichotomous, is balanced. In the case in which it is not balanced, there are some resampling Bootstrap methodologies that can be applied. In this concrete case, it is likely that the target variable (Dummy for assassination) is not balanced since the negative cases (defender not murdered) are probably much higher than the positive cases.

Stage 2. Training: The purpose of this phase is to train an algorithm to generate the prediction of the target variable from the contextual variables. Training involves identifying under which hyper-parameters and which algorithms the exercise achieves a better prediction performance. This process is usually done through cross-validation[11] in an iterative process in which it identifies which combination of hyper-parameters[12] from a previously defined grid for each algorithm achieves a better result.

Normally, the selection of algorithms depends on the database and is usually a trial-and-error process; however, some of the algorithms that have demonstrated better performance on structured data in recent years are Gradient Boosting Machine[13] and Random Forest[14] (Olson et al., 2017; Wainer, 2016; Lee et al., 2018). These algorithms share the characteristic that their operation is based on a defined compilation of simpler basic algorithms, such as decision trees or regression models so that over-fitting is avoided in the aggregate.

Stage 3. Evaluation: Once the optimal hyper-parameters have been defined for each algorithm with cross-validation in the training sub-sample, the algorithm is tested in the evaluation sub-sample. This sub-sample has been independent of the training process; therefore, it allows a reliable comparison between the predicted value of the probability and the real value of the risk of assassination. The final purpose of this process would be to identify whether the algorithms have a good predictive ability or not.

Normally, for dichotomous variables, an AUC metric[15] above 0.70 is a good prediction result; however, in this case, it may be more useful to focus on other evaluation measures that focus on the prediction accuracy of positive cases (murdered defenders). Such a measure could be the percentage of positive cases that the algorithm predicts correctly; nevertheless, the problem, in this case, would be for the algorithm to be optimized in such a way that it predicts all observations as positive to minimize the prediction error. This is a risk that could lead to a different measure being required.

Stage 4: Prediction: After developing all the previous phases successfully, the algorithm can now be applied to the prediction of the risk of murder with new data that does not contain the target variable. The output of the algorithms will be the predicted probability of assassination for each defender. This probability would be the direct measure of risk presented by the defender and delimit which defenders should be prioritized in the review of cases and assignment of protection.

In general, the method described in this section would transform the third stage of the protection assignation process -preliminary evaluation through risk, vulnerability, and threats criteria-, in which the level of risk faced by the human rights defender is determined, as it constitutes a faster alternative methodology for predicting the risk of assassination faced by the defender.

4.3 Unsupervised Machine Learning

4.3.1 Definition of variables

In contrast to supervised learning, unsupervised learning does not have a target variable to predict. Instead, its primary function is to identify underlying patterns in the data that offer associations of interest. In this case, the aim is to process the data in such a

11 Cross-validation is a technique used to evaluate how the results of statistical analysis generalize to an independent data set.

12 These are model parameters that can not be automatically optimized for any given data set. In these cases, the optimal values must be calibrated by the researcher from the data.

13 Gradient Boosting Machine can be interpreted as an additive function of simple models, where each additional model is adjusted by the data that the previous model had greater difficulty in predicting or classifying. This is achieved by creating a synthetic version of the database, in which observations that were not correctly classified or predicted have greater weight.

14 The Random Forest algorithm makes a set N of decision trees, determining for each, a set of variables available at random so that each tree is different from the other. The prediction is made for the N trees and the final result is averaged so that an aggregate result is obtained.

15 The AUC metric is the Area Under the Curve ROC. The ROC Curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots the True Positive Rate against the False Positive Rate at different classification thresholds.

way as to identify the subset of defenders that exhibit characteristics that may make them more likely to be assassinated. This is developed for this exercise through a clustering methodology that performs this task on a set of contextual variables; these variables are considered important in determining the probability of assassination of a human rights defender. The set of variables is the number and type of threats received, whether the defender works on issues related to land property rights, the individual's subjective perception of risk, and two complementary indices that include municipal factors that, according to the existing literature, explain the systematic murder of human rights defenders in the territory.

The first proposed index will be a violence exposure index that will compile municipal indicators such as: the presence of armed groups, hectares of coca, the presence of illegal economies, and the homicide rate of the municipality. The second, a marginality index that will include dimensions such as the administrative capacity of the municipality (National Planning Department Index), the poverty level of the municipality, the distance of the municipality from the departmental capital, and the municipal rural rate. On the other hand, some extra variables are suggested to be tested in the development of the exercise, such as a municipal corruption index[16] and whether one belongs to an ethnic group.

4.3.2 Methodology

The methodology proposed to carry out this exercise of segmentation and identification of the risk of assassination is defined in general terms below:

Stage 1. Preparation: This phase includes the processes of i) data collection, ii) selection of clustering variables, iii) treatment of missing values, and iv) scaling of variables. The selection of variables is a complex process that must be carried out in light of the specialized literature on the subject. Specifically, we should seek to incorporate those variables that have had a greater association with the murder of social leaders, i.e., those that are deterministic of the phenomenon. The above selection is a first proposal of possible variables that could be used, but the final selection may be adjusted as the algorithm is tested.

Stage 2. Clustering: This phase involves performing a multivariate clustering method on the selected variables. This methodology would seek to group observations based on the underlying patterns of the data, to obtain groups whose observations share certain common characteristics. There are different methods to perform clustering, the most common is to use the k-means algorithm[17] with Euclidean distance as the clustering metric. On the other hand, the optimal number of clusters is usually determined from the data, for which the Elbow method can be used (Thorndike, 1953).

Stage 3. Identification: from the analysis of the clustering results, the objective is to identify those subgroups that present characteristics that could be more associated with the risk of being murdered. Therefore, in this phase, a detailed analysis of the cases is performed so that different associations between variables can be related to different levels of risk of being murdered. This is done through a review of the centroid value in the variables used (i.e., the average of the variables in the subgroup) and defines which associations may be related to a higher risk of murder. This classification can be checked with training data and identify whether the cluster classification coincides with the defenders who were murdered in the past; otherwise, other variables and cluster classification should be considered.

The identifying of high-risk clusters provides a list of criteria to be considered for prioritizing the allocation of protection. For the Colombian case, in addition to improving the preliminary risk assessment, this could expedite the Risk Assessment and Action Recommendation Committee (CERREM) discussions in the fourth stage of the National Protection Unit process, as the list of criteria provided by the exercise should facilitate a consensus about protection decisions among the parties.

4.4 How will the use of artificial intelligence make a difference?

In the Colombian case, the use of machine learning algorithms in the allocation of protection schemes for human rights defenders can lead to a gain in efficiency and accuracy, ultimately resulting in a better allocation of measures to prevent the assassination of defenders. This is because the algorithms can

16 In this case, it could be useful to include the municipal corruption risk estimation developed by Mojica (2021).

17 The k-means clustering seeks to divide the observations into k clusters in which each observation belongs to the cluster with the closest mean in the variables (cluster centroid). The final result is a partition of the data space into Voronoi cells.

handle the analysis of the risk of assassination faced by several human rights defenders in a matter of minutes, based solely on the defenders' contextual variables, speeding up the identification procedure that previously could take months. In addition, this new risk estimation can potentially be much more accurate than the result of the sum-of-percentages methodology since it considers a more flexible estimation process that starts from a set of variables that can be deterministic of the phenomenon.

5. Ethical Concerns

The use of artificial intelligence for the targeting of protective measures may raise some relevant ethical issues, as its use would indirectly lead the algorithm to decide who is most likely to live. This is a problem that needs to be given special attention and consideration because it may partly condition how these tools can be applied in practice. Specifically, it is not possible to assign complete responsibility to algorithms for the protection allocation decision, because this could lead to situations where a person's life may be affected by the decision of a machine that may not understand the overall context of the situation.

Therefore, the assignment decision should not be based solely on the algorithm's result, but rather be a complement to an interdisciplinary group that decides who receives the protection. In this way, the algorithm would function as the first filter or first stage of reviewing the applications and would establish an estimated risk. Subsequently, the interdisciplinary group would review the applications with the highest priority based on the level of risk of each person.

From this situation, the ethical problem can be nuanced if the algorithm performs better than humans in estimating the risk of assassination and prioritizing cases, i.e., if it makes fewer errors overall and performs with better results in terms of lives saved. This would result in a positive contribution to society, as it would reduce on the net the number of defenders killed. However, these algorithms can be biased and not perform perfectly, so they require constant monitoring and continuous adjustment to ensure the greatest transparency in the allocation of protection.

6. Policy Recommendations

The developed project shows the potential of machine learning tools to prevent the assassination of human rights defenders in the world. This application can make an important difference when there are problems in estimating the risk of assassination presented by the defender, which makes it difficult to target and assign protective measures.

The case of Colombia is illustrative of how these tools can be implemented in practice. The country presents a dramatic situation, in which there is no administrative capacity to adequately target and assign protective measures; at this point machine learning can be decisive to achieve a better prioritization of cases. The approach presented here is the first step towards a semi-automated model for assigning protection measures; however, the effectiveness of the algorithms needs to be tested in practice. This could not be done for this report due to access restrictions for this type of data, but it is a task that should be pursued in the future with the support of the National Protection Unit. Some public policy recommendations emerge from the analysis carried out:

- It is necessary that the countries with the highest incidence of the problem seek to collect and manage quality information on human rights defenders who could be at risk of assassination; data is the first step in developing schemes for targeting and allocation of protection measures that can generate an amelioration of the problem.
- The targeting and assignment of protection measures should be an expeditious process once a defender requests protection. Machine learning tools can be useful in estimating the risk of murder and streamlining the process of prioritizing critical cases. Two approaches can be considered for this: supervised learning and unsupervised learning.

- Implementing a targeting scheme supported by supervised learning algorithms requires a process of training and evaluation of the algorithms on historical data of assassinated and non-assassinated human rights defenders. The result would be an estimated probability of assassination that would be calculated from the contextual information of the defenders. The degree of success will depend on the ability of the contextual variables to predict the target variable; therefore, it is a process that must be verified to validate if it has a satisfactory result.
- Implementing a targeting scheme supported by unsupervised learning algorithms can be much simpler, but at the same time more difficult to validate; it requires a process of clustering the variables that are most deterministic of the problem and determining the subgroups with the highest relative risk. This can be sensitive to the selection of variables, so it must be validated before its practical application.
- In any case, it is necessary to ensure that the final decision is confirmed by experts in the field, to avoid placing all the responsibility for assignment on the machine. In this sense, the estimated risk cannot be the defining factor of the assignment, but rather a complement that supports humans and speeds up the process.

8. References

Aarvik, P. (2019). *Artificial Intelligence–a promising anti-corruption tool in development settings?*. Report núm. 1. Bergen, Noruega: U4 Anti-Corruption Resource Centre, Chr. Michelsen Institute.

Albarracín, J., Milanese, J. P., Wolf, J., Valencia, I., & Navarro de Arco, M. (2020). *La Lógica Política De Los Asesinatos De Líderes Sociales: Autoritarismo Competitivo Local Y Violencia En El Posacuerdo.* Retrieved June 28, 2021, from http://library.fes.de/pdf-files/bueros/la-seguridad/16811.pdf

Alves, L. G., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, 505, 435-443. https://doi.org/10.1016/j.physa.2018.03.084

Bachelet, M. (2020). *Statement by Michelle Bachelet at The 45th session of the Human Rights Council*. Retrieved June 28, 2021, from https://reliefweb.int/report/afghanistan/her-global-human-rights-update-bachelet-calls-urgent-action-heighten-resilience

ColombiaReports. (2018). 'They are killing us': Colombia's social leaders cry for help. Retrieved June 28, 2021, from https://colombiareports.com/they-are-killing-us-colombias-social-leaders-cry-for-help/

Defensores, P. S., del Catatumbo ASCAMCAT, A. C., Abierta, V., Antioquia, M. R. V., & por la Paz, P. (2018). *¿ Cuáles son los patrones? Asesinatos de Líderes Sociales en el Post Acuerdo.* Retrieved June 28, 2021 from https://www.cinep.org.co/publicaciones/es/producto/cuales-son-los-patrones-asesinatos-de-lideres-sociales-en-el-post-acuerdo/

Front Line Defenders. (2021). *Global Analysis 2020.* Retrieved June 28, 2021, from https://www.frontlinedefenders.org/sites/default/files/fld_global_analysis_2020.pdf

Front Line Defenders. (n.d.). *killing cases.* Retrieved June 28, 2021, from https://www.frontlinedefenders.org/en/cases/violation/killing

Garzón-Vergara, J. C. (2015). Avoiding the perfect storm: Criminal economies, spoilers, and the post-conflict phase in Colombia. *Stability: International Journal of Security and Development,* 4(1), 36. http://doi.org/10.5334/sta.fx

Gutiérrez, F., Marín, M., Machuca, D., Parada, M., & Rojas, H. (2020). Paz sin garantías: el asesinato de líderes de restitución y sustitución de cultivos de uso ilícito en Colombia. *Estudios Socio-Jurídicos*, 22(2).

Human Rights Watch (2021). *Left Undefended: Killings of Rights Defenders in Colombia's Remote Communities*. Retrieved June 28, 2021, from: https://www.hrw.org/report/2021/02/10/left-undefended/killings-rights-defenders-colombias-remote-communities#_ftn11

Kalyvas, S. N. (2006). The logic of violence in civil war. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511818462

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and engineering ethics.* 26(1), 89-120. https://doi.org/10.1007/s11948-018-00081-0

Lanteró Senan, D. (2017). Who killed Mario Andrés?: Aggressions against human rights defenders in Colombia: is the protection framework working? (Thesis dissertation). https://doi.org/20.500.11825/525

Lawlor, M. (2021). *Killings of human rights defenders 'go largely unnoticed by outside world.* Interview for DW News. Retrieved June 30, 2021, from https://www.dw.com/en/killings-of-human-rights-defenders-go-largely-unnoticed-by-outside-world/a-56757472

Le Billon, P., & Lujala, P. (2020). Environmental and land defenders: Global patterns and determinants of repression. *Global Environmental Change*, 65. https://doi.org/10.1016/j.gloenvcha.2020.102163

Lee, H. C., Yoon, S. B., Yang, S. M., Kim, W. H., Ryu, H. G., Jung, C. W., Suh, K. S., & Lee, K. H. (2018). Prediction of acute kidney injury after liver transplantation: machine learning approaches vs. logistic regression model. *Journal of clinical medicine*, 7(11), 428.

McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. M*achine Learning and Applications: An International Journal* (MLAIJ), 2(1), 1-12. 10.5121/mlaij.2015.2101

Ministry of Interior (2015). *Decree 1066 of 2015*. Retrieved from: https://www.mininterior.gov.co/sites/default/files/decreto_1066_de_2015_unico_reglamentario_del_sector_administrativo_del_interior.pdf

Mojica, K. (2021). *Inteligencia Artificial para la Detección de Corrupción en la Administración Municipal de Colombia.* Universidad de los Andes, Working Paper.

National Protection Unit (2019). *Registration form for the individual protection program.* Retrieved from: http://visor.suit.gov.co/VisorSUIT/index.jsf?FI=17622

National Protection Unit (2020). *Resolution 001 of 2020.* Retrieved from: https://www.unp.gov.co/wp-content/uploads/2020/02/1resolucion-001-del-02-de-enero-de-2020-desagregacion-presupuesto.pdf

National Protection Unit (n.d.). *Administrative procedures and processes.* Retrieved June 28, 2021, from: https://www.unp.gov.co/atencion-y-servicios-a-la-ciudadania/tramites-otros-procedimientos-administrativos/

Naidoo, K. (2018). World summit brings surge of new commitments to protect human rights defenders. Declarations cited by Amnesty International. Retrieved June 30, 2021, from https://www.amnesty.or.th/en/latest/news/9896/

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H. (2017). Data driven advice for applying machine learning to bioinformatics problems. *arXiv preprint arXiv*:1708.05070.

Prem, M., Rivera, A. F., Romero, D. A., & Vargas, J. F. (2018). *Killing social leaders for territorial control: the unintended consequences of peace.* Documento de Trabajo Universidad del Rosario No. 016385.

Palacios, D. (2020, May 26). *'69% of security schemes are for social leaders': UNP Daniel Palacios. Interview for Semana*. Retrieved from: https://www.semana.com/nacion/articulo/seguridad-lideres-sociales-entrevista-director-encargado-unp/673884/

Quintana, M. M., & Fernandéz, E. E. (2011). *Protection of Human Rights Defenders: Best Practices and Lessons Learnt: Part I: Legislation, National Policies and Defenders Units: Part II: Operational Aspects of Defenders' Protection*. Protection International.

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1), 112-118. https://doi.org/10.1093/bioinformatics/btr597

The Armed Conflict Location & Event Data Project (2021). *ACLED-Armed Conflict Location and Event Database.* Retrieved April 04, 2021, from https://acleddata.com/data-export-tool/

Thorndike, R. L. (1953). Who belongs in the family?. *Psychometrika*, 18(4), 267-276.

The United Nations General Assembly (1998). *Declaration on the right and responsibility of individuals, groups and organs of society to promote and protect universally recognized human rights and fundamental freedoms.* United Nations Declaration. A/RES/53/144.

The United Nations Human Rights Council Special Rapporteur (2021). *Final warning: death threats and killings of human rights defenders: Report of the Special Rapporteur on the situation of human rights defenders, Mary Lawlor*. United Nations Document. A/HRC/46/35.

Uribe, L., Posada, F., Castro, B., & ACLED, R. K. (2020). *Understanding the Killing of Social Leaders in Colombia During COVID-19.* Retrieved April 26, 2021, from https://blogs.lse.ac.uk/latamcaribbean/2020/10/06/understanding-the-killing-of-social-leaders-in-colombia-during-covid-19/

USAID, CODHES and CNC (2019). *Lidera la Vida: ¿Qué estamos perdiendo? Impacto de los homicidios*

de líderes y lideresas sociales en Colombia. Piloto para la interpretación de su impacto en las comunidades y organizaciones sociales. Bogotá, Colombia. ISBN: 978-958-8881-29-4.

Wainer, J. (2016). Comparison of 14 different families of classification algorithms on 115 binary datasets. *arXiv preprint arXiv*:1606.00930.

We Are Defenders Program (2017). *Contra las cuerdas.* Annual report 2016. Sistema de Información sobre Agresiones contra Defensores de DDHH en Colombia SIADDHH. Retrieved May 6, 2021, from is.gd/kyomPm

We Are Defenders Program (2018). P*iedra en el zapato.* Annual report 2017. Sistema de Información sobre Agresiones contra Defensores de DDHH en Colombia SIADDHH. Retrieved May 6, 2021, from is.gd/2aNk1f

We Are Defenders Program (2019). *La Naranja Mecánica.* Annual report 2018. Sistema de Información sobre Agresiones contra Defensores de DDHH en Colombia SIADDHH. Retrieved May 6, 2021, from is.gd/A3FR40

We Are Defenders Program (2020). *La Ceguera.* Annual report 2019. Sistema de Información sobre Agresiones contra Defensores de DDHH en Colombia SIADDHH. Retrieved May 6, 2021, from t.ly/XbaH

We Are Defenders Program (2020). *El Virus de la Violencia.* Semi-annual report January-June 2020. Sistema de Información sobre Agresiones contra Defensores de DDHH en Colombia SIADDHH. Retrieved May 6, 2021, from t.ly/nw54

Witness, G. (2020). *Defending tomorrow: The climate crisis and threats against land and environmental defenders.* Retrieved June 28, 2021, from https://www.globalwitness.org/en/campaigns/environmental-activists/defending-tomorrow

Fuente: ColombiaReports (2018)