Inteligencia Artificial 24(67), 121-128
doi: 10.4114/intartif.vol24iss67pp121-128

INTELIGENCIA ARTIFICIAL

http://journal.iberamia.org/

COMPUTATIONAL MODEL TO SUPPORT THE DETECTION OF PROFILES OF MISSING PERSON IN COLOMBIA

Gerardo Ernesto Rolong Agudelo[1,A], Carlos Enrique Montenegro Marín[1,B], Paulo Alonso Gaona García[1,C]

[1] Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas

[A]gerolonga@correo.udistrital.edu.co,[B]cemontenegrom@udistrital.edu.co,[C]pagaonag@udistrital.edu.co

Abstract: In the world and some countries like Colombia, the number of missing persons is a worrying and growing phenomenon, every year, thousands of people are reported missing all over the world. The fact that this keeps happening might indicate that there are still analyses that have not been done and tools that have not been considered in order to find patterns in the information of missing persons. The present article presents a study of the way informatics and computational tools can be used to help find missing persons and what patterns can be found in missing person datasets using as a study case open data about missing persons in Colombia in 2017.
The goal of this study is to review how computational tools like data mining and image analysis can be used to help find missing persons and draw patterns in the available information about missing persons. For this, the state of art of image analysis in real world applications was reviewed in order to explore the possibilities when studying the photos of missing person, then a data mining process with data of missing person in Colombia was conducted to produce a set of decision rules that can explain the cause of the disappearance, as a result decision rules are generated which ones suggest links between socioeconomic stratification, age, gender and specific locations of Colombia and the missing person reports. Among the rules obtained some of them like the one that suggests Ibague as the municipality where enforced disappearance is occurring are consistent with studies that have been done in Colombia.
In conclusion, this work reviews what information about missing persons is available publicly and what analysis can be made with them, showing that data mining and face recognition can be useful tools to extract patterns and identify patterns in missing person data.

Keywords: Missing person, Data mining, Decision rules, Missing person patterns, Cost sensitive learning, Open data.

1 Introduction

Missing persons is a problem that all countries should embark on, “According to estimates from the Federal Government, 200,000 people disappear from their homes every year in Brazil”[1] and “Every year an estimated 200,000 people go missing in the UK”[2]**.** Examples like above keep repeating in several countries all over the world, in order to tackle it several organizations try to centralize and make available to the public pictures and information of missing ones; in the United States this data is mainly provided by the FBI and organizations like NameUs (National Missing and Unidentified Persons System) and The National Center for Missing & Exploited Children, in the worldwide context, it can be found the ICMP(International Commission on Missing Persons) and the LOST(Learning Opportunities, inStruments and Investigation Techniques to fight the growing phenomenon of missing person in Europe) projects. For the Colombian case sort of equivalent information can be found in the public consults site of El Instituto Nacional de Medicina Legal y Ciencias Forenses.

Although the majority of these projects share useful data, most of it is not in a format that it can be easily analyzed by everyone who is interested in doing it, you can visualize a few records at a time, but there is no easy

ISSN: 1137-3601 (print), 1988-3064 (on-line)
©IBERAMIA and the authors

way of getting a dataset with hundreds of records to be studied, also most of the initiatives present descriptive statistics of the information but no inferred conclusions are shown.

Considering that hundreds of missing persons are still reported every year and given that in Colombia in 2019 according to the public consults site of El Instituto Nacional de Medicina Legal y Ciencias Forenses 2528 males and 1427 females were reported missing a data mining analysis of a public data set of missing person in Colombia in 2017 using the Waikato Environment for Knowledge Analysis(WEKA)[3] was conducted.

The rest of this article is organized as follows. In section 2 we made a review about what studies and processing tools could be used with the published pictures and records of missing persons, in section 3 the methodology and steps made to data mine the dataset are presented, in section 4 it is presented the result of the data mining analysis and the work is concluded in section 5.

2 Literature review or research background

Using technology in the process of finding missing person can be a game changer factor, considering the availability of pictures of the missing person, it is interesting to consider what could be done in a real world situation where there are not hundreds of records per person, the before picture most likely is provided by the family while the after picture can come from police records or even a photo taken by someone with a phone in an unforeseen situation while trying to be not too evident, so it is logical that the picture can have bad illumination, most likely would not be a front picture of the face and could even be out of focus.

The authors of [4] present how a face detection and recognition algorithm will perform better when trained and tested with similar quality images instead of a full mixed dataset so it is a very useful consideration that the focus should be on training and testing algorithms on images with similar quality. On the other side [5] shows how face recognition in video cameras needs a good lightning condition.

Another of the challenges of trying facial verification of the missing person is the passage of time, so it is important to consider the way aging can affect algorithm's performance, in [6] the authors exposes as a result of a longitudinal study using about 230234 images that "commercial off the shelf" face recognition systems are capable of recognizing a person with a success rate of 99% using 8 images per subject as long as the time span difference of the images is not greater than 8.5 years where the performance of the algorithms decrease dramatically. In [7] it is shown how adding age information when using a deep neural network in the cross-age verification process can boost the performance in current state-of-art algorithms.

In terms of the confidence that public missing person data can provide, the author of [8] made a study comparing the information provided by the NameUs organization with the police records. While still having into account that the number of records that could be compared in the research is very limited, it was found that after running statistical analysis the variables age, eye colour, hair colour, and height reported in the NameUs data is very similar to the contained in the arrest reports, and the delay in the NameUs information upload compared to police's date of the missing did not show a significant effect on the accuracy of the reported information.

In [9] the authors analyse missing person data in the UK and then focused on the 3352 children cases, given that their dataset had a unique identifier for each person they intended to find temporal patterns and found that the number of children reported missing 10 times is greater that what would be expected just by chance and that the number of kids reported missing 10 or more times under authorities care is bigger than those reported missing 9 or less. Although there was not enough information to determine an explicit relation between drug and alcohol dependency in the population of children reported missing 10 or more times, they showed a greater reported level of dependencies to these substances. Finally, the study shows that children reported missing are more likely to be reported missing again in a time lapse of four weeks after the first report and this likelihood decreases with the passage of time.

For the analysis of the missing person records Data Mining could be a valuable tool as "Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes."[10]. In [11] the authors tried to find patterns in the data of missing persons in New South Wales between 1980 and 2000 just focusing on runaways and intents or committed suicides reported in this dataset. They used WEKA to generate decision rules that can contribute to police officers to determine the cause of the disappearance of a person, although the algorithm they used PART resulted worse in predicting the outcome of the missing case when they compared it with an Artificial Neural Network, they emphasize that the conducted study not only generated rules that can be useful, but also "provided insight into variables that have potential to accurately predict outcomes for missing persons cases and highlighted issues pertaining to data capture, pre-processing and rule determination".

3 Materials and methods

For the present study from the common data mining process (Figure 1) we divided the stages of the analysis of data selection, data exploration, pre-processing and patterns extraction.

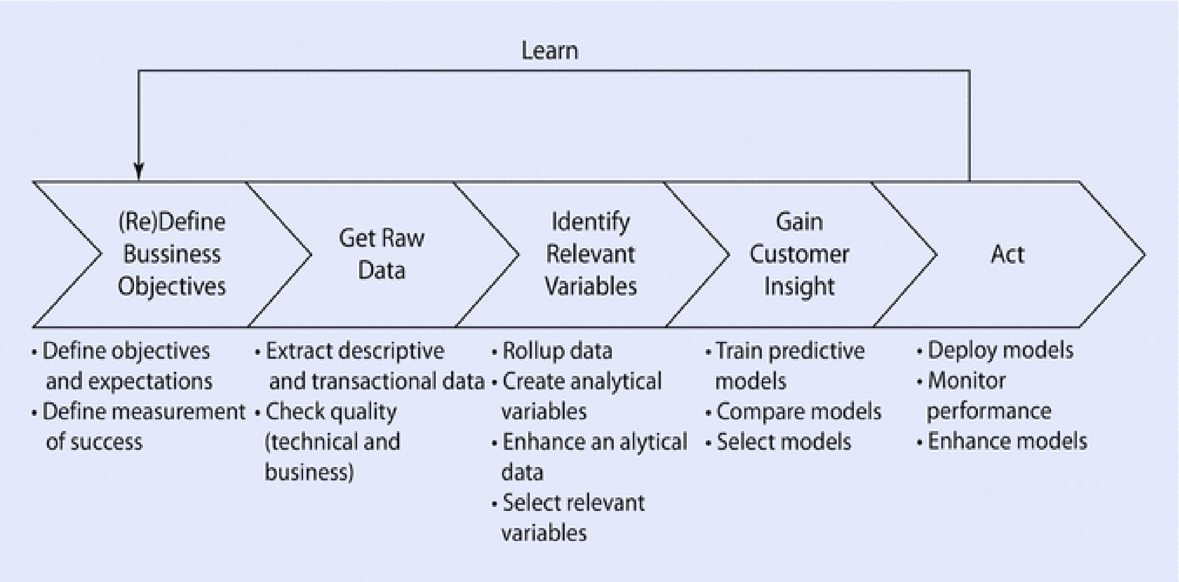
Learn
(Re)Define Bussiness Objectives
Get Raw Data
Identify Relevant Variables
Gain Customer Insight
Act
• Define objectives and expectations
• Define measurement of success
• Extract descriptive and transactional data
• Check quality (technical and business)
• Rollup data
• Create analytical variables
• Enhance an alytical data
• Select relevant variables
• Train predictive models
• Compare models
• Select models
• Deploy models
• Monitor performance
• Enhance models

Figure 1. Data mining process. Source:[12].

3.1 Data Selection

Several governmental websites were consulted, but not much useful information was found, in the public consults site of El Instituto Nacional de Medicina Legal y Ciencias Forenses where some data is made public in a paginated table under the title "Cadáveres CNI identificados y registrados como desaparecidos. Convenio Interadministrativo de Cooperación 01 de 2010 MinInterior - Medicina Legal - Registraduria" just by going through the pages it is easy to see how pretty much all records show missing or incomplete data with the name of "update pending", and this table cannot be downloaded. Later on a dataset in [13] with information of missing person in Colombia in 2017 was located, this data was consulted in the SIRDEC(Sistema de Información red de Desaparecidos y Cadáveres) in december of 2017, there are 6203 records with the following 20 columns:

- Year (same for all)
- Moth of the report
- Day of the report
- Age group
- Under/over age
- Definitive life cycle (age group and a nominal category)
- Gender
- Marital status
- Education level
- Racial ancestor
- Appeared live/dead or still missing.
- Locality
- Occurrence zone
- Presumed cause.
- Government institution that reported the missing
- Vulnerability factor
- Missing cause
- Department
- Municipality
- Municipality code

3.2 Data exploration

Once we started to analyse the dataset using WEKA for visualizing the variable's distribution it was found that more males than females were reported missing (Figure 2) which is concordant with the tendency that remains from 2016 to 2020 as shown by Medicina Legal statistics (Figure 3).

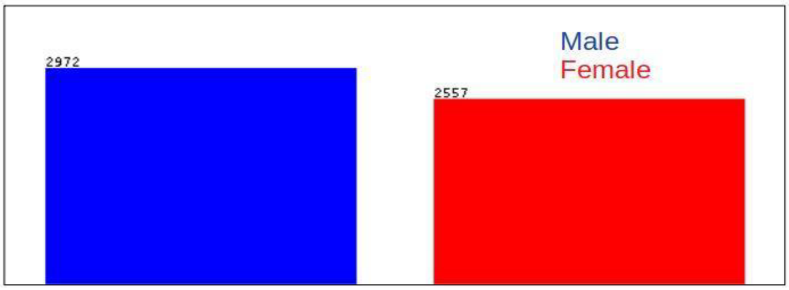

Figure 2. Male vs Female distribution in the 2017 missing person dataset in Colombia. Source: own work.

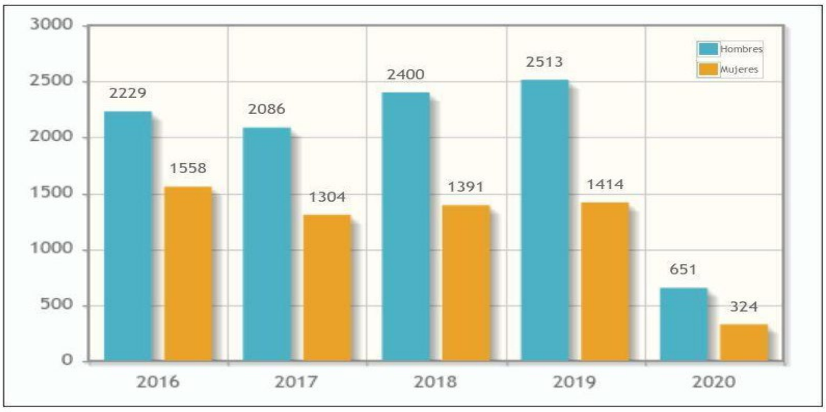

Figure 3. Male vs Female distribution of missing person in Colombia from 2016. Source: El Instituto Nacional de Medicina Legal y Ciencias Forenses.

Also, we visualized that from the 6202 reports 2087(33.5%) belongs to persons between 12 and 17 years old and the majority of these records are from females (Figure 3).

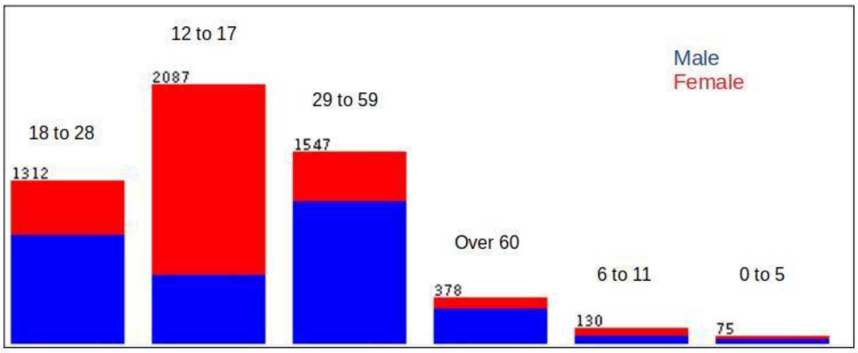

Figure 4. Age range distribution in Colombia in 2017 missing person. Source: own work.

3.3 Pre-processing

Before trying to use any data mining algorithm in the data, it is important to make sure that the information is as clean as possible so accurate and realistic results are obtained, in our case by visualizing the values of the records it was found that many of them contained the “no information” value which may lead to decision rules like “if category A and B have no information then C will have no information”, this essentially tells us nothing useful about the missing person, also inspecting the categories of the data from now on called nominal variables it is clear that some of them are redundant, i.e. one can be calculated from another or shows the same information than another, but just with a different aggregation level. So we decided to remove some of the redundant or irrelevant variables (e.g. the year as is the same for all records), the day and month of the report as we are not making a temporary study with just one year data, the government institution that reported the missing and the repeated location variables in order to get shorter decision rules and decrease computation time, then we removed some of the records that contained most “no information” values so the final dataset to be processed ended up containing 3358 records and 10 variables.

Among the remaining variables we chose to use as target the missing cause which contains the values voluntary absence, death, enforced disappearance, involuntary(mental disorder) and involuntary, from this variable we found that the distribution of the values in unbalance, 63% are voluntary absence, 7% death, 0.41% enforced disappearance, 2.59% involuntary(mental disorder), 6.55% involuntary and a 20% of remaining missing values.

3.4 Patterns extraction

For generating a set of rules that can be used to try to explain what conditions can be decisive to the missing cause, we decide to use PART [14] algorithm of WEKA which makes a set of rules from partial J48 threes, given the previously mentioned unbalance of the data a Cost Matrix was used to perform a cost-sensitive learning in order to penalize bad classification of the missing cause[15] with the class “voluntary absence” because it is easy to see that if any classification algorithm decided to assign the value of “voluntary absence” to all records, just by the distribution of the data it would get a 63% of correctly classified instances.

As evaluation method in the training process a ten-fold cross validation was chosen which is a method were “We train a model using some instances of the dataset and leave some instances out of it to test the model after it has been trained.”[16], in our case with the ten folds, the data is divided into ten parts, nine of them are used for training and one for testing and the process is repeated until all ten parts are used for testing. After that WEKA creates the final model that generates the set of decision rules that better fits the whole dataset.

To obtain consistent results of the performance of the chosen combination (i.e. cost-sensitive learning, cross validation and PART decision rule algorithm) we used the experimenter of WEKA[3] to run ten tests where the software changes the random numbers seeds so that every run will not perform exactly the same, so from those ten runs with ten-fold cross validation the accuracy of the processing gets evaluated one hundred times.

4 Results

From the experimenter of WEKA, we obtained that from 100 evaluations, the algorithm got an average accuracy of 72.93% with a standard deviation of 2.39 so the employed data mining process performs better than just assigning voluntary absence as missing cause to all records which would produce a 63% of accuracy.

From the model that produced the final set of rules we got that even though the ROC Area for every value is greater than 0.5 which indicates that our algorithm classifies better than chance, but apart from voluntary abuse and death, it is evident that the model ability to classify the other values is very poor as they ended up with an F-Measure below 0.3. (Table 1) all of these points to the need of getting a more balanced dataset so the algorithm would not be affected by most values having the same class.

Table 1**.** Model missing cause identification performance. Source: own work

	Precision	Recall	F-Measure	ROC Area	Class
	0.871	0.876	0.874	0.717	Voluntary Absence
	0.763	0.504	0.607	0.822	Death
	0	0	0	0.711	Enforced Disappearance
	0.152	0.08	0.105	0.772	Involuntary (mental disorder)
	0.171	0.264	0.208	0.653	Involuntary
Weighted Avg.	0.777	0.763	0.766	0.723	

WEKA also produces a confusion matrix which is another way of visualizing the algorithm performance and it shows how many of the records in the dataset were correctly classified in the principal diagonal and in the rest of the columns appears the number of misclassifications into that category, for example for Voluntary Absence 1864 record were associated correctly with its true value of missing cause.

Table 2. Model confusion matrix. Source: own work.

a	b	c	d	e	classified as
1864	10	2	28	224	a = Voluntary Absence
87	119	1	3	26	b = Death
13	0	0	0	1	c = Enforced Disappearance
49	0	1	7	30	d = Involuntary (mental disorder)
126	27	1	8	58	e = Involuntary

Despite the bad performance of the model when trying to classify most classes, still 38 decision rules were produced, although none of them comes from a pure node, which means neither of these rules can determine uniquely a result and applying it results in some records misclassified. Among them the ones that are not redundant or trivial are:

- Municipality = Mocoa: Death
- Definitive life cycle = (12 to 17) Teenager: Voluntary Absence
- Municipality = Barranquilla: Involuntary(mental disorder)
- Municipality = Soacha: Involuntary
- Municipality = Reported abroad: Voluntary Absence
- Municipality = Bogotá D.C. AND Gender = Female AND Marital status = Single: Voluntary Absence
- Municipality = Bogotá D.C. AND Vulnerability factor = user of psychoactive substances (drugs, alcohol, etc.): Voluntary Absence
- Municipality = Bogotá D.C. AND Vulnerability factor = None AND Locality = Ciudad Bolivar: Involuntary
- Municipality = Bogotá D.C. AND Vulnerability factor = None AND Locality = Bosa: Involuntary
- Municipality = Bogota D.C. AND Vulnerability factor = None AND Locality = Engativa: Voluntary Absence
- Municipality = Bogotá D.C. AND Vulnerability factor = None AND Locality = Suba: Involuntary
- Municipality = Bogotá D.C. AND Vulnerability factor = None AND Locality = Usme: Involuntary

- Municipality = Bogotá D.C. AND Vulnerability factor = None AND Locality = Chapinero: Death
- Municipality = Bogotá D.C. AND Vulnerability factor = None AND Locality = Tunjuelito: Involuntary
- Municipality = Bogotá D.C. AND Vulnerability factor = None AND Locality = Santa fe: Involuntary
- Municipality = Bogotá D.C. AND Locality = Fontibon: Involuntary
- Municipality = Bogotá D.C. AND Locality = Los Martires: Death
- Municipality = Bogotá D.C. AND Locality = Kennedy: Involuntary
- Municipality = Santiago de Cali AND Marital status = Free Union: Involuntary
- Municipality = Ibagué AND Vulnerability factor = None: Involuntary
- Municipality = Bucaramanga: Involuntary
- Municipality = Ibagué AND Education level = No information: Enforced disappearance

From the generated rules we can see that some interesting patterns emerge, the rules that conclude as missing cause “involuntary" in Bogotá suggest an association with some of the localities where most of its socioeconomic stratification[17] is conformed mainly by levels three or below as stated in the local sheets of each locality[18].

Some of the other rules suggest that reported missing teenagers, single females in the municipality of Bogotá and users of psychoactive substances go missing voluntarily while in the municipality of Ibagué is where enforced disappearances are occurring.

5 Discussion and Conclusions

This study conducted a research on what public information about missing persons is available, and what can be done in terms of image analysis and data processing, using a data mining process.

The utilized dataset contained several values with missing information so most of it could not be used in the study which shows the need to improve the way this data is registered and points to the fact that if more of this information was available publicly better datasets could be formed to analyse them getting better results and even temporal patterns could be found.

The final dataset resulted in an unbalanced set of records to be used in the actual pattern extraction limiting the algorithm performance and reliability of the resulting rules, however the data mining process showed that useful patterns and interesting links can be extracted between socio-economic stratification, age, gender and specific locations of Colombia and missing person cases.

Some of the obtained rules suggest profiles that are consistent with studies that have been done in Colombia, like the rule that states Ibague as the municipality where enforced disappearance is occurring which agrees with [19] where they state Ibague as the second city with more forced displacements in the country. For the case of the rule that shows as missing cause death for the people reported in Mocoa, it can be deeply related with the fact that in such city in April 2017 about 300 people died because of a flash flood as reported in [20].

In future work, the results of this research should be presented to governmental organizations, so they can explore the utility of the obtained results and possibly get motivated to make public more data so further studies can be conducted. Also, more information should be gathered from the websites that publicly expose the data and also social media where several posts about missing persons are published by their family and shared by lots of people, in order to form datasets containing more records with less missing values. Patterns in the images of the missing person could be studied and comparisons among the characteristics of a reported missing person from different countries would be made.

References

[1] R. S. Ferreira, C. G. de Oliveira, y A. A. B. Lima, «Myosotis: An Information System Applied to Missing People Problem», en Proceedings of the XIV Brazilian Symposium on Information Systems - SBSI'18, Caxias do Sul, Brazil, 2018, pp. 1-7, doi: 10.1145/3229345.3229379.

[2] «Missing Children and Adults - A cross government strategy», p. 28.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009, doi: 10.1145/1656274.1656278.

[4] F. Yang, Q. Zhang, M. Wang, and G. Qiu, “Quality Classified Image Analysis with Application to Face Detection and Recognition,” *ArXiv180106445 Cs*, Jan. 2018, doi: 10.1109/ICPR.2018.8545476.

[5] F. P. Mahdi, M. M. Habib, M. A. R. Ahad, S. Mckeever, A. S. M. Moslehuddin, and P. Vasant, “Face recognition-based real-time system for surveillance,” *Intell. Decis. Technol.*, vol. 11, no. 1, pp. 79–92, Apr. 2017, doi: 10.3233/IDT-160279.

[6] D. Deb, L. Best-Rowden, and A. K. Jain, “Face Recognition Performance under Aging,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 548–556, doi: 10.1109/CVPRW.2017.82.

[7] X. Wang, Y. Zhou, D. Kong, J. Currey, D. Li, and J. Zhou, “Unleash the Black Magic in Age: A Multi-Task Deep Neural Network Approach for Cross-Age Face Verification,” in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, DC, USA, May 2017, pp. 596–603, doi: 10.1109/FG.2017.75.

[8] S. Duncan, “Unsolvable? Assessing the Accuracy of Missing Person Case Data,” p. 117.

[9] A. Babuta and A. Sidebottom, “Missing Children: On the Extent, Patterns, and Correlates of Repeat Disappearances by Young People,” *Polic. J. Policy Pract.*, Sep. 2018, doi: 10.1093/police/pay066.

[10] P. Prasdika and B. Sugiantoro, “A Review Paper on Big Data and Data Mining Concepts and Techniques,” *IJID Int. J. Inform. Dev.*, vol. 7, no. 1, p. 33, Dec. 2018, doi: 10.14421/ijid.2018.07107.

[11] K. Blackmore, T. Bossomaier, S. Foy, and D. Thomson, “Data Mining of Missing Persons Data,” in *Classification and Clustering for Knowledge Discovery*, vol. 4, S. K. Halgamuge and L. Wang, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 305–314.

[12] V. Kumar and W. Reinartz, “Data Mining,” in *Customer Relationship Management: Concept, Strategy, and Tools*, V. Kumar and W. Reinartz, Eds. Berlin, Heidelberg: Springer, 2018, pp. 135–155.

[13] Instituto Nacional de Medicina Legal y Ciencias Forenses, “Base de datos preliminar de personas reportadas como Desaparecidas Enero-Noviembre 2017.” https://www.datos.gov.co/, Dec. 19, 2017, Accessed: Dec. 10, 2019. [Online]. Available: https://www.datos.gov.co/Estad-sticas-Nacionales/Base-de-datos-preliminar-de-personas-reportadas-co/85g8-qemt.

[14] E. Frank and I. Witten, “Generating Accurate Rule Sets Without Global Optimization,” *Mach. Learn. Proc. Fifteenth Int. Conf.*, 1998.

[15] A. Kim, K. Oh, J.-Y. Jung, and B. Kim, “Imbalanced classification of manufacturing quality conditions using cost-sensitive decision tree ensembles,” *Int. J. Comput. Integr. Manuf.*, vol. 31, no. 8, pp. 701–717, Aug. 2018, doi: 10.1080/0951192X.2017.1407447.

[16] S. Yadav and S. Shukla, “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, Bhimavaram, India, Feb. 2016, pp. 78–83, doi: 10.1109/IACC.2016.25.

[17] J. G. Yunda, “Densificación y estratificación social en Bogotá: distribución sesgada de la inversión privada,” *EURE Santiago*, vol. 45, pp. 237–257, 2019.

[18] “Fichas locales 2019 | Veeduría Distrital.” https://www.veeduriadistrital.gov.co/content/Fichas-locales-2019 (accessed Mar. 15, 2020).

[19] L. J. B. Caicedo and A. J. Q. Genneco, “DELITO DESPLAZAMIENTO FORZADO POR LA VIOLENCIA,” p. 17.

[20] J. E. Vásquez Santamaría, M. I. Gómez Vélez, and H. D. Martínez Hincapié, “The Mocoa tragedy: Example of a retrospective without an end point in the management of the risk of disasters detonated by natural events?,” *Rev. Derecho Uninorte*, no. 50, pp. 145–186, Jul. 2018, doi: 10.14482/dere.50.0007.