Rom J Leg Med [29] 74-80 [2021]
DOI: 10.4323/rjlm.2021.74
© 2021 Romanian Society of Legal Medicine

A PREDICTIVE MODEL FOR THE MISSING PEOPLE PROBLEM

Enrique Delahoz-Domínguez*, Silvana Mendoza-Brand

Facultad de Ingeniería, Universidad Tecnológica de Bolívar, Cartagena, Colombia

Abstract: The disappearance of people is a multidimensional phenomenon, in which several aspects must be considered. It affects people's security perception and consumes police resources in its treatment. Therefore, does exists an emotional circumstance for the relatives of the missing person. At the same, the police departments must develop a search task, in most cases with much uncertainty. In this research, a predictive model to predict missing people's status is presented. The information used to create the model come from the Colombian legal Medicine Institute, in a public dataset composed of 6202 cases and 11 variables. The output variable was the final disappearance status, with the categories Appears Dead, Appears Alive, and Still Disappeared. Three supervised machine-learning algorithms were trained and tested for the model creation, K-Nearest Neighbours, Decision Trees, and Random Forest. The study was divided into three phases, first considering all the output categories. In the second phase, generating a binary classification for the Appeared and Not appeared instance. Thirdly, models were built to predict the status of appeared persons, Appears Alive or Appears Dead. The K-NN algorithm outperforms the other models with an Area under the curve value of 94.8%.

Keywords: missing people, supervised learning, knowledge discovery, predictive modeling.

INTRODUCTION

Each year, 10 million people worldwide go missing (i.e., people who get lost), and this problem has been a popular subject of study as seen in several articles [1]. The main motivation to characterize and model the process of missing persons is two-dimensional. For one, police agencies often have limited resources and need to use them efficiently; for example as presented in [2], the police spend 14% of their time on missing people incidents. Therefore, they require information that allows them to prioritize search tasks based on objective information. On the other hand, there are family members and other persons affected by the disappearance, who undoubtedly experience suffering and despair initiated by the disappearance of a loved one [3]. In these cases, the possibility of generating relevant information about the future status of the person represents an opportunity to cover this space of uncertainty and misinformation.

These models are generators of knowledge that allow for the development of new theories and methodologies that optimize the decision-making process and promote a change in citizens' philosophical approach to day-to-day problems [4]. The present research focuses on the prediction of the status of a disappeared person, with particular emphasis on the identification of the variables that determine their possible status. These include whether the missing individual is dead, alive, or still disappeared.

Implementing machine learning models in the field of crime analysis is a line of dynamic research and current validity. For example, through a supervised learning model, the dangerousness of individuals convicted of domestic violence was determined to establish their future punitive penalties [5]. From a general point of view, [6] presents a model to assess community perception of criminality based on a Geographic Information System (GIS).

Given the magnitude and importance of the phenomenon of disappeared persons in police administration and citizens' perception of security, this research answers the following questions: is it possible to create a model that allows for the characterization and prediction of missing persons' status? What is the ideal methodology for the creation of a model that

*Correspondence to: Enrique de la Hoz Dominguez, Universidad Tecnológica de Bolívar, Facultad de Ingeniería, Km 1. Vía a Turbaco. Cartagena - Colombia, E-mail: edelahoz@utb.edu.co

predicts the status of the disappeared? Which machine learning technique is best suited to the phenomenon of missing persons? Following the proposal, this research presents an alternative model for predicting missing persons' future status by identifying key variables in the final classification of missing persons in the three defined statuses.

METHODOLOGY

This research aims to characterize the phenomenon of disappearances, to predict the situation or the status of new disappeared persons. This problem affects Colombia in general; therefore, through machine learning techniques, a supervised learning model will be created to obtain relevant information to support and manage missing persons. Based on the above, it is introduced a new approach to the treatment of the problem. It is proposed to articulate data mining tools and machine learning to generate a holistic approach to the problem, aiming to generate information to modeling the uncertainty. Taking into account the contextualization of the problem in all its dimensions. In Figure 1, it is introduced the research methodology of the model proposed.

Data

The database used for this research contains information about 6202 disappearing cases in 2017 at Colombia: 59% male and 41% female with average age 38.5 years (range = 0-80 years) at time of disappearing. The full database and variable's description is found in (Datos Abiertos Colombia, Datos.gov.co). The information was collected by the Colombian general prosecutor's office in its annual inform of criminality and security.

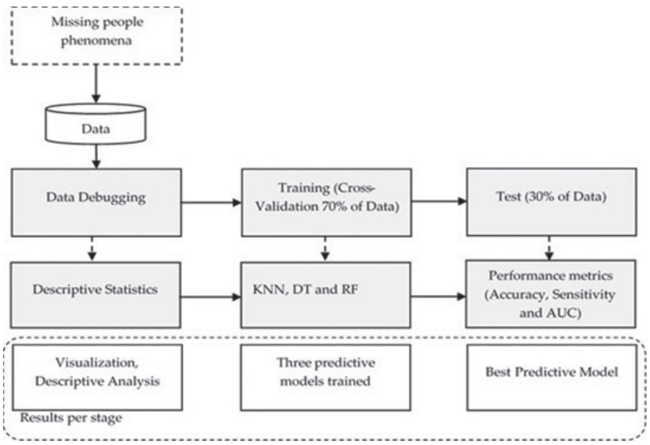
Missing people phenomena
Data
Data Debugging
Training (Cross-Validation 70% of Data)
Test (30% of Data)
Descriptive Statistics
KNN, DT and RF
Performance metrics (Accuracy, Sensitivity and AUC)
Visualization, Descriptive Analysis
Three predictive models trained
Best Predictive Model
Results per stage

Figure 1. Research framework.

Independent variables

The following 11 baseline variables after debugging comprised the input characteristics analyzed in the model.

- Seven variables corresponding to social and personal information of the person: vital cycle, the range of age, adult, gender, marital status, scholarship level and racial ancestrality.

- Two Variables corresponding to the date of the event: Month and day.

- Two variables corresponding to the location of the event: Occurrence zone and risk factor.

Dependent variable

The output (classification) variable is the final status of the disappearing event, a categorical variable composes of the levels Dead, Alive, Still disappeared. Data corresponding to 80% of total cases, sampled randomly, were used as training data and the remaining 20% as test data. In data, 8% of cases contained partial information or missing fields. Therefore, in consideration of the research, the missing data were included in the study to ensure that the model can deliver predictions using incomplete data, which is very common in these circumstances.

Machine Learning modelling

The development of this project is based on the use of machine learning applied to a social problem. A machine learning algorithm is a process in which a model is fitted to a set of data, this stage of the process is called training, after that stage, the model is used with a set of independent data to determine how much you can generalize the model, this other stage is called a test [8]. The models used in this project are all supervised algorithms (Table 1).

k - Nearest Neighbours algorithm (KNN)

It consists of creating a neighborhood for all categories and subsequently comparing a new value or data with a number k of neighbors based on distances (Kataria y Singh, 2013). It is essential to choose an appropriate value for k since this depends on the success of the classification. For this, it is possible to execute several times the algorithm with different values of k until obtaining the one with a better performance.

Decision Trees (DT)

It is a machine learning model that bases its classification by analyzing and examining each of the data variables using statements of the type yes-then. That is to say; it asks if an observation meets a criterion, then the observation probably belongs to the category.

Its purpose is to divide the data into two branches based on some value; this value is called the point of division [10].

Random Forest (RF)

It creates multiple decision trees and combines them randomly to obtain higher prediction accuracy. This model can be used for classification and regression. The objective of this technique in this work will be to classify it. Unlike the decision trees, Random Forest adds additional randomness to the model while the trees grow. Instead of looking for the most essential feature when dividing a node, look for the best feature among a random subset of features, this results in a wide diversity that generally results in a better model [11].

Performance metrics

The predictive process's key occurs by minimizing the difference between the predicted value and the actual value. This relationship is described by the Positive True (VP), True Negative (VN), False Positive (FP), and False Negative (FN) metrics. The metrics used to assess performance will be the correct classification rate or Accuracy (A), Positive predictive value (PPV), Negative predictive value (PPN), Sensitivity (S) and Specificity (E), and the area under the curve (AUC). The area under the curve represents the rate of TP and FP at various discrimination thresholds. A model with a perfect classification will have an AUC = 1. On the other hand, a very random model would yield an AUC value = 0.5.

RESULTS

As explained in the methodology, four prediction models applied to the database were used, adapting the categories of the response variables

Table 1. Variables description

Variable	Type	Values
Month of disappeareance	Categorical	January to December.
Day of disappeareance	Categorical	Monday to Sunday.
Range of age	Categorical	(0 - 4 , 5 - 9, 10 - 14, 15 to 17, 18 to 19, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49, 50 to 54, 55 to 59, 60 to 64, 65 to 69, 70 to 74, 75 to 79, > 80) years.
Adult or	Binary	Younger (under 18 years) or Adult (over 18 years).
Definite Vital cycle	Categorical	Early childhood (0 a 5 years), Childhood (6 a 11 years), Adolescence (12 a 17 years), Youth (18 a 28 years), Adulthood (29 a 59 years), Elderly (mas de 60 years).
Sex	Categorical	Male - Female.
Marital status	Categorical	Married, divorced, single, free union, widowhood.
Adjusted schooling	Categorical	Kindergarden, primary basic education, secondary basic education, university, technical education or professional, Postgraduate education, No schooling, without information.
Ultimate Racial Ancestor	Categorical	Asian, white, indigenous, half blood, mulatto, black.
Occurrence Zone	Categorical	Municipal head, populated center, sparse rural town, no information.
Vulnerability factor	Categorical	Consumers of psychoactive substances (drugs, alcohol, etc.), farmers, displaced by violence, homeless, belonging to ethnic groups, civic leaders, exercise of judicial activities, exercise of political activities,exercise of trade union or union activities, exercise of sex work, former convicts, injured and / or sick under health or medical protection, homeless, teacher or educator, people in custody, people with diverse sexual orientation (LGBTI), demobilized or reinserted people, belonging to gangs, recyclers, health workers or humanitarian mission, religious, none, others, without information.
Department	Categorical	This variable can take a value among the 32 departments of Colombia in addition to 3 additional values that are the following: Bogota, missing nationals abroad, San Andres, Providencia and Santa Catalina archipelago.

according to the disappearance status. First, the results for the three models used in the original database will be presented, which classify the disappeared into 3 categories: Appears Alive, Appears Dead, and Still Disappeared. These results will be called Phase 1. The results of the four models used in the second division of the data will be exposed, which will classify the disappeared into two categories: Appears and Still Disappeared, this will be Phase 2. Finally, using only the data Appears category, the results of the machine learning models used in this database will be presented. So, classifying the reports into two categories: Appears Alive and Appears Dead, this will be Phase 3 of the results (Fig. 2).

Results of Phase 1

Results of Phase 1 for KNN model

The KNN model presented overfitting in the test data, this occurs when the model adapts to the characteristics of the data that is available, and this worsens its predictive capacity when faced with new data, for this reason, a noticeable difference in the performance with the test data is perceived at Table 2. Following this, the performance of the model was evaluated using different values for the parameter K. The best value of k was seven, a relatively small value compared to the amount of data; this means that the model can make decisions with little data, which makes it efficient.

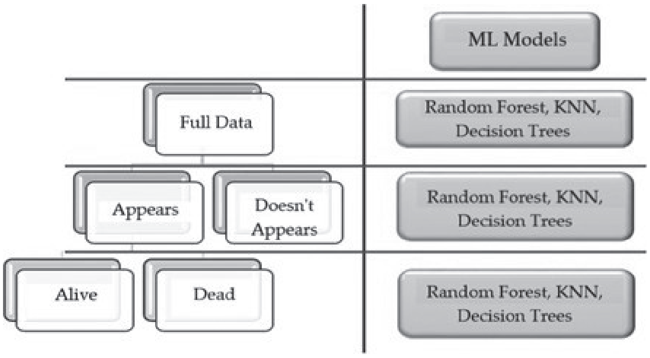
ML Models
Full Data
Random Forest, KNN, Decision Trees
Appears
Doesn't Appears
Random Forest, KNN, Decision Trees
Alive
Dead
Random Forest, KNN, Decision Trees

Figure 2. Research framework.

Analyzing the confusion matrix for the test data (Table 2), we can see that thirty predictions of the Appears dead category were correctly identified while thirteen did not; this gives a percentage of successes for this category of 69.77%, and an error of 30.24%. Likewise, 309 predictions for the category Appears alive were correctly identified, while 233 were wrong. Finally, 572 predictions for the category' Still disappeared' were correctly made, and 394 were equivocal. The performance of this model with the test data was 61.18%. Next, Table 4 shows all the performance indicators of this model for the test data. Among these is the sensitivity of the model, taking into account this parameter. The model could predict the category Still disappeared' more effectively than the remaining categories for which this parameter reflected a lower value. We also found the specificity that, in this case, was 0.99, the highest value found; this is the percentage of people belonging to the categories' Appears alive' and' Still disappeared' correctly identified as not belonging to the category appears dead'.

Results of phase 1 for the Decision tree model

This model bases the decision-making on the analysis of each of the variables. So, the confusion matrix is presented for the results of this model's predictions with the training data and the test data.

Results of phase 1 for Random Forest

The Random Forest model has a parameter called (n) that refers to the number of decision trees that will be created and mtry that indicates the number of variables that will be taken to create the trees, in this

Table 2. Confusion matrix for knn on Phase 1

Data	Categories	Appears Dead	Appears Alive	Still Disappeared	AUC
Train	Appears Dead	242	2	3	97.14%
	Appears Alive	2	1893	68	
	Still disappeared	3	55	2383	
Test	Appears Dead	30	0	13	61.18%
	Appears Alive	12	309	221	
	Still disappeared	44	350	572	

Table 3. Performance metrics for each output category

	KNN		
Performance metrics	Appears Dead	Appears Alive	Still disappeared
Sensitivity	0.3	0.452	0.705
Specifity	0.992	0.729	0.455
Pos Pred Value	0.71	0.537	0.595
Neg Pred Value	0.958	0.657	0.576

case it was 9. The higher the number of trees, the better to learn from the data, however, many trees make the training process considerably slow. Therefore, it was tested with different values of n from 100 to 2000, always obtaining the same performance. Finally, n = 500 was chosen. The confusion matrix for the test and training data of this model are shown in Table 5.

Results of Phase 2

Table 6 shows the results for the models mentioned with Phase 2 of the investigation. There are three boxes for the KNN model, considering preliminary tunning for the k parameter. It was tested with values from 1 to 25 to observe within that range whose value favored the model's performance to a greater extent, finding the best value k equal to 19. The model that presented the best performance for the test data within the three compared in the previous table was Decision Tree, with a performance of 64.53%.

Results of Phase 3

For Phase 3, a total of 2942 cases match the criteria for Appears alive and Appears dead, that is 47.44% of the total database reports, divided as follows: 333 cases for the category' Appears dead' (11.32%) and 2609 cases for the category' Appears alive' (88.68%). The highest precision model was KNN with k equal to five, with an AUC value of 94.84%.

DISCUSSION

The disappearance issue has been studied from different approaches and scenarios, such as news, research, and consultations. Other studies based on the analysis of DNA profiles to determine the individual's identity are found in Colombia [12], by developing a scientific tool called MESP (Predictive Spatial and Statistical Modeling), whose purpose was the prediction of the burial sites of missing persons, which could be a complement to the scheme proposed in the

Table 4. Confusion matrix for Decision Tree model at Phase 1

Data	Categories	Appears Dead	Appears Alive	Still disappeared	AUC
	Appears Dead	112	2	59	
Train	Appears Alive	45	1057	622	63.14%
	Still disappeared	96	890	1768	
	Appears Dead	33	1	15	
Test	Appears Alive	18	385	213	64.53%
	Still disappeared	29	274	583	

Table 5. Confusion matrix for the Random Forest model at Phase 1

Data	Categories	Appears Dead	Appears Alive	Still Disappeared	AUC
	Appears Dead	240	1	1	
Train	Appears Alive	3	1893	65	97.20%
	Still Disappeared	4	56	2388	
	Appears Dead	34	3	13	
Test	Appears Alive	16	361	241	61.05%
	Still Disappeared	36	295	552	

Table 6. Summary of Confusion matrix for the Phase 2

Data	Model	Category	Still Disappeared	Appears	AUC
Train	KNN	Still Disappeared	2380	61	0.974
		Appears	57	2153	
Test	KNN	Still Disappeared	483	336	0.564
		Appears	340	392	
Test	KNN (K=19)	Still Disappeared	565	363	0.599
		Appears	258	365	
Train	Decision Tree	Still Disappeared	1743	953	0.646
		Appears	694	1261	
Test	Decision Tree	Still Disappeared	589	316	0.645
		Appears	234	412	
Train	Random Forest	Still Disappeared	2374	52	0.975
		Appears	63	2162	
Test	Random Forest	Still Disappeared	526	334	0.593
		Appears	297	394	

Table 7. Summary of confusion matrix for the Phase 3

Data	Model	Categories	Appears Dead	Appears Alive	AUC
Train	KNN	Appears Dead	261	0	100%
		Appears Alive	0	1945	
Test	KNN	Appears Dead	40	17	93.61%
		Appears Alive	30	647	
Test	KNN (k=5)	Appears Dead	35	1	94.84%
		Appears Alive	37	663	
Train	Decision Tree	Appears Dead	128	5	93.74%
		Appears Alive	133	1940	
Test	Decision Tree	Appears Dead	30	3	93.89%
		Appears Alive	42	661	
Train	Random Forest	Appears Dead	258	1	99.82%
		Appears Alive	3	1944	
Test	Random Forest	Appears Dead	31	7	93.48%
		Appears Alive	41	657	

present investigation; the database used was based on cartographic information provided by the communities and geo-referenced data of the exhumations carried out by the national prosecutor's office.

The Decision Tree and KNN models presented a better performance concerning the rest of the models in all study phases. For phase 1, the AUC values are between 61% and 64%. Therefore, it is essential to emphasize that I exceed 50%, demonstrating that the results are not due to chance, but to the model's ability to discriminate between the response categories. In this case, it is advisable to inquire about the use of more predictive variables to improve the model's performance. In the third phase, we found that the highest performance model was the KNN, with a value greater than 94%. Despite reflecting a very high precision value, KNN presents a Sensitivity output below 50%, his indicates our estimator's ability to identify as belonging to those elements that belong to the category Appears Dead. As explained in previous paragraphs, some models tend to overfit, predicting a specific category if it represents most of the data.

When developing a machine learning model, it is necessary to devote special attention to the generation and collection of the database; this is undoubtedly one of the most important aspects because the quality of the data largely determined the model's accuracy. As mentioned earlier, for this investigation, a database with 12 variables was used, which contained 6202 reports of missing persons for 2017 in Colombia, obtaining precision results up to 65%. The number of reports turns out to be more than 18 times the number of cases used in a relapse prediction study in childhood acute lymphoblastic leukemia (ALL), in which the number of cases used was 336, even so, the latter model reached one accuracy levels close to 80% [13]. Most machine learning applications occur in medical and economic fields; in the case of forensic issues, specifically in terms of missing persons, the application of technological tools is deficient.

In conclusion, this research presents to the scientific and academic community a model for the prediction of the status of missing persons, providing a quantitative vision for making objective decisions in a work context with a high level of uncertainty. Therefore, contextualizing the research findings could be essential for social organizations to identify variables affecting the disappearance process's outcome, fostering optimal use of resources, enabling the search task, re-conceptualizing the police function, and moving from a passive role to an active one based on objective arguments.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Burch RA. Presumed Dead: Why Arizona Should Shorten the Required Time for Beloved Missing Persons to Be Declared Legally Dead. Ariz. Summit Law Rev. 2017; 10: 59.
2. Yang Y, Hu X, Liu H, Jiawei Z, Li Z, Yu P.S. r-instance Learning for Missing People Tweets Identification. 2018.
3. Parr H, Stevenson O, Woolnough P. Search/ing for missing people: Families living with ambiguous absence. Emot. Space Soc. 2016; 19: 66-75.
4. Mannini A, Sabatini AM. Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers. Sensors. 2010; 10 (2).
5. Berk RA, Sorenson SB, Barnes G. Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions. J. Empir. Leg. Stud. 2016; 13(1): 94-115.
6. Cutajar J, Formosa S, Calafato T. Community Perceptions of Criminality: The Case of the Maltese Walled City of Bormla. Soc. Sci. 2013; 2(2).

7. «Base de datos preliminar de personas reportadas como Desaparecidas Enero-Noviembre 2017 | Datos Abiertos Colombia». https://www.datos.gov.co/Estad-sticas-Nacionales/Base-de-datos-preliminar-de-personas-reportadas-co/85g8-qemt (accedido oct. 07, 2020).

8. Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc. 2016; 3(1).

9. Kataria A, Singh MD. A Review of Data Classification Using K-Nearest Neighbour Algorithm. 2013.

10. Therneau T, Atkinson B, Ripley B. Recursive Partitioning and Regression Trees. 2019.

11. Breiman L. Random Forests. Mach. Learn. 2001; 45(1): 5-32.

12. Franco RMF, Giraldo GCV. Identifying missing people in the National Database of Genetic Profiles for Application in Judicial Investigation —CODIS—: Two case reports. Case Rep. 2015; 1(2).

13. Pan L, Liu G, Lin F, Zhong S, Xia H, Sun X, Liang H. Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Sci. Rep. 2017; 7(1).